Recalibration of Audiovisual Simultaneity in Speech

Akihiro Tanaka[1,2], Kaori Asakawa[3,4], Hisato Imai[4]

[1] Department of Psychology, University of Tokyo
[2] Cognitive and Affective Neuroscience Laboratory, Tilburg University
[3] Graduate School of Information Science, Tohoku University
[4] Department of Psychology, Tokyo Woman's Christian University
a.tanaka@uvt.nl

Abstract

Recent studies have shown that the audio-visual synchrony is recalibrated after adaptation to a constant timing difference between auditory and visual signals (i.e. temporal recalibration). Here we investigated whether the temporal recalibration occurs for audio-visual speech using an off-line adaptation method. After 3 minutes of lag observation, the audio-visual synchrony is recalibrated toward the adapted lag. The point of subjective simultaneity shifted after 10 seconds of lag observation, whereas the just noticeable difference did not change during this short observation period. The width of the temporal window extended only to the direction of audio delay. These findings extend the findings in previous studies and suggest different properties of temporal recalibration in speech.

Index Terms: audio-visual integration; speech perception; temporal recalibration; temporal order judgment

1. Introduction

Multisensory integration requires temporal coordination of signals from multiple sensory modalities. These signals, however, do not need to be precisely synchronous to be perceived as a single event. Audio-visual asynchrony is tolerated to some extent. The sensitivity to audio-visual asynchrony was first reported by Hirsh and Sherrick [1].

Speech perception is one of the examples of multisensory perception. Listeners use the visual information from the speaker's mouth for speech perception as well as auditory speech. Audio-visual asynchrony in speech signals is often observed in live televised satellite broadcasts. Sensitivity to audio-visual asynchrony in speech has been measured by indirect methods such as McGurk task (e.g., [2–4]) and visual enhancement of intelligibility (e.g., [5–8]) as well as direct methods such as temporal order judgment (TOJ) task (e.g., [9,10]) and simultaneity judgement task (e.g., [4,11–13]). Both lines of evidence have shown that sensitivity to audio-visual asynchrony is lower in speech than in simple nonspeech. These results suggest that human can handle with a relatively large amont of lag between auditory and visual speech signals.

Another important findings on how humans handle with intersensory lags are reported from the studies of "temporal recalibration." Recent studies have shown that the audio-visual synchrony is recalibrated after adaptation to a constant timing difference between auditory and visual signals (e.g., [14,15]). In these studies, observers were exposed to a series of visual (e.g., flash) and auditory (e.g., tone pip) stimuli with a constant lag and then judged the simultaneity and temporal order of visual and auditory stimuli. The results suggest that subjective simultaneity changes after adaptation to a time lag. Vatakis et al. [16] investigated temporal recalibration using speech materials. In their study, participants were exposed to two streams of audiovisual stimuli. One is a foreground stream, in which visual and auditory signals were presented with various timing. The other is a background stream, in which visual and auditory signals were either synchronous or asynchronous. Observers were engaged in a 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 background speech stream). Results revealed a significant shift in the point of subjective simultaneity (PSS) toward the direction of the asynchrony in the background stream in the dual-task condition. No shift was observed in the single-task condition.

So far, it is not clear whether temporal recalibration *following* exposure to asynchrony (i.e., off-line adaptation method), not the *concurrent* exposure to asynchronous speech stream in a dual-task situation, occurs for an audio-visual speech signal. Therefore, we investigated temporal recalibration for audio-visual speech using an off-line adaptation method. This methodology enables us to compare between the results obtained from nonspeech stimuli [14,15] and speech stimuli more directly.

2. Experiment 1

2.1. Methods

2.1.1. Participants

Thirteen graduate and undergraduate students participated in Experiment 1. All of them reported normal hearing and normal or corrected-to-normal visual acuity. All were native Japanese speakers.

2.1.2. Materials

The audiovisual stimuli were created from digital audio and video recordings of male and female Japanese speakers. Adaptation and re-adaptation stimuli were a series of monosyllable (/pa/, /ta/, or /ka/) spoken by three male and three female speakers. In half of the sessions, we used congruent audiovisual speech syllables and the other half contained incongruent ones (visual /ka/ and auditory /pa/). Test stimuli were monosyllables (/pa/, /ta/, or /ka/) spoken by either of the female speakers. The video clip (640 * 480 pixels, 29.97 frames/s) and the auditory speech (digitized at 48 kHz, with a 16-bit quantization resolution) were combined and desynchronized using Adobe Premiere Pro 2.0.

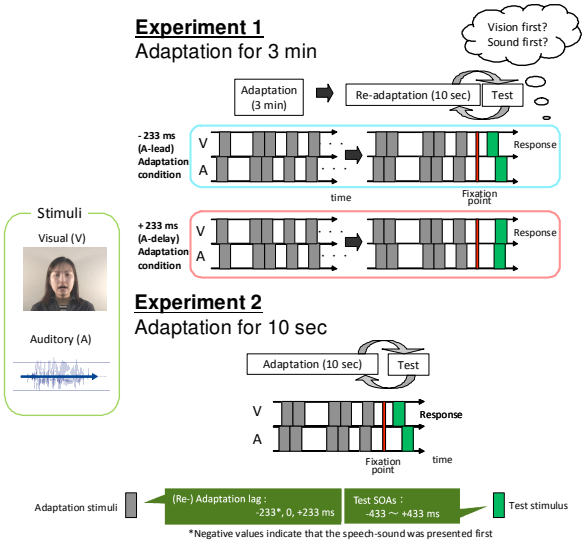

Figure 1: Schematic diagrams of the two experiments.

2.1.3. Procedure

Participants were 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 approximately 70 dB SPL. Pink noise was added to the speech sounds, resulting in a +10 dB signal-to-noise ratio.

Each session started with an adaptation phase of 3 min with a constant time lag between the visual and auditory speech (–233, 0, or +233 ms: positive value indicates audio delay). In each adaptation phase, a single syllable was presented repeatedly. The adaptation phase was followed by test trials, each preceded by a 10-s re-adaptation (see Figure 1). The audiovisual lag of the re-adaptation stimuli was identical with that of the adaptation stimuli in each session. The test stimulus was presented with various stimulus onset asynchronies (SOAs) between visual and auditory speech (13 SOAs ranging between –433 and +433 ms). The participants' task was to judge whether the auditory or visual stimulus was presented first. Participants were instructed to respond accurately rather than quickly. The experimental session, which lasted approximately 25 minutes, consisted of 78 test trials (6 repetitions of the 13 SOAs). Four experimental sessions were run for each adaptation condition. Participants were engaged in one adaptation condition per day.

2.2. Results and Discussion

For each participant, the proportion of 'vision-first' responses was calculated for each combination of adapted lag and SOA. For each adapted lag, an individually determined psychometric function was calculated by fitting a cumulative normal distribution using maximum likelihood estimation. The interpolated 50% crossover point represents the PSS. The steepness was expressed in terms of the just noticeable difference (JND), which represents half the difference in SOA between the 25% and 75% point, corresponding to the smallest interval each observer can notice. The average psychometric functions for each of the adapted lag are shown in Figure 2. The average PSSs and the JNDs are shown in Figure 3 and Figure 4, respectively.

A one-way analysis of variance (ANOVA) on the PSS revealed a significant main effect of adapted lag [$F(2,24) = 5.69$, $p < .01$]. Multiple comparison (Ryan's method) revealed that the PSS at –233 ms lag was significantly smaller than at 0 and +233 ms. In a one-way ANOVA on the JND, there was a significant main effect of adapted lag [$F(2,24) = 5.49$, $p < .05$]. Multiple comparison revealed that the JND at +233 ms lag was significantly larger than at 0 and –233 ms.

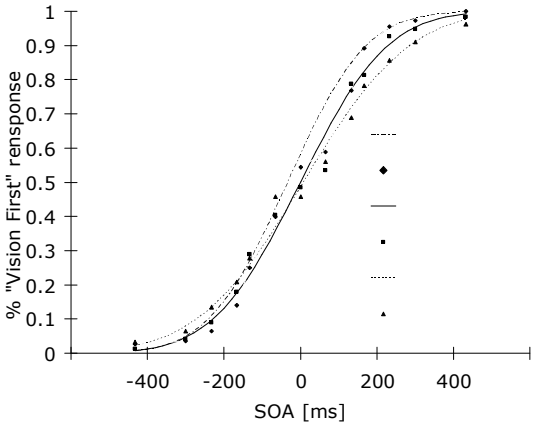

Figure 2: Mean proportion of 'vision first' responses in Experiment 1. Lines show the average psychometric functions for each of the adapted lag.

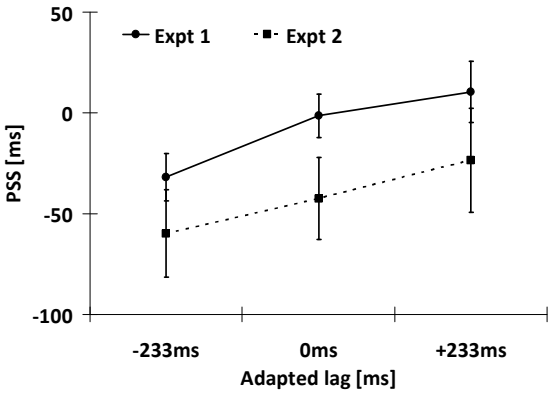

Figure 3: Mean PSS values for the two experiments as a function of the adapted lag. The error bars represent the standard errors of the means.

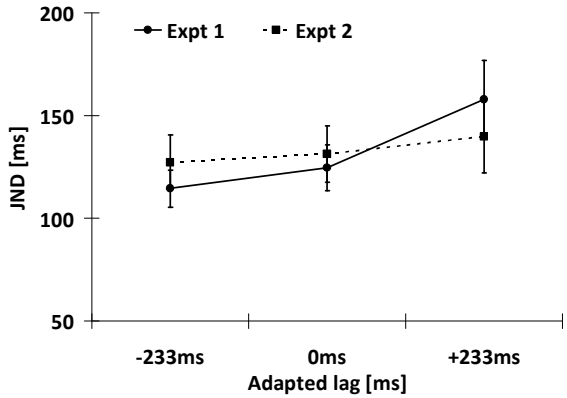

Figure 4: Mean JND values for the two experiments as a function of the adapted lag. The error bars represent the standard errors of the means.

Although we used congruent and incongruent (McGurk like) audiovisual speech syllables as adaptation and re-adaptation stimuli, there were no interactions between congruency and adapted lag (PSS: $F(2,24) = 0.40$, $p = .67$; JND: $F(2,24) = 1.17$, $p = .33$). Also, there were no interactions between the gender of adaptation stimulus (male and female) and adapted lag (PSS: $F(2,24) = 0.34$, $p = .72$; JND: $F(2,24) = 0.15$, $p = .86$). These results suggest that temporal recalibration occurs irrespective of whether the adaptation and the test stimuli are identical or not.

The results of Experiment 1 clearly showed that adaptation to asynchronous speech affects the synchrony perception. In Experiment 2, we examined whether this temporal recalibration occurs only by a brief presentation to a time lag before each trial.

3. Experiment 2

3.1. Methods

Participants were eleven graduate and undergraduate students. There was no adaptation phase of 3 min in Experiment 2. Test trials followed a 10-s adaptation to audiovisual lag (–233, 0, or +233 ms), which was equivalent to the re-adaptation in Experiment 1. The order of the audiovisual lag was randomized among trials. Except for the above points, experimental methods were the same as in Experiment 1.

3.2. Results and Discussion

The average psychometric functions for each of the adapted lag are shown in Figure 5. The average PSSs and the JNDs are shown in Figure 3 and 4, respectively.

In a one-way ANOVA on the PSS, there was a significant main effect of adapted lag [$F(2,20) = 6.32$, $p < .01$]. Multiple comparison revealed that the PSS at –233 ms lag was significantly smaller than at +233 ms. This result shows that participants' response changes after 10 seconds of observation of asynchronous speech. In contrast, a one-way ANOVA on the JND revealed that the main effect of adapted lag was not significant [$F(2,20) = 1.29$, $p = .30$].

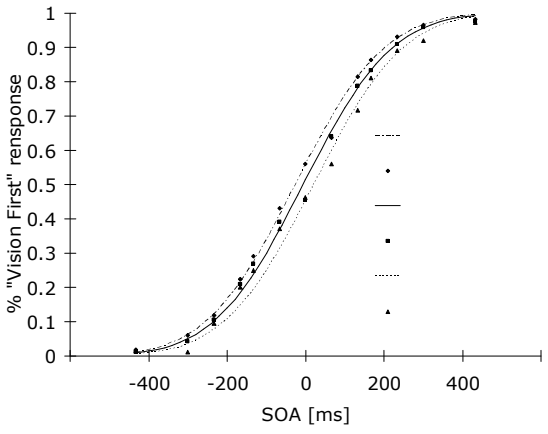

Figure 5: Mean proportion of 'vision first' responses in Experiment 2. Lines show the average psychometric functions for each of the adapted lag.

4. General Discussion

In two experiments, we demonstrated temporal recalibration after exposure to a constant time lag between visual and auditory speech. This result extends the findings using simple nonspeech stimuli in an off-line adaptation method [14,15] and the finding using speech stimuli in a dual-task method [16]. Taken together with these studies, temporal recalibration seems a robust phenomenon which occurs for both speech and nonspeech stimuli and is observed through either off-line and on-line adaptation methods.

The width of the temporal window (i.e., JND) extended only to the direction of audio delay. While previous studies have shown bidirectional extension of the temporal window (e.g., [14]), our results demonstrated asymmetric patterns between video-delay and audio-delay directions. Given the main difference between these studies is the stimuli (flash and tone-pip in a previous study [14] and audiovisual speech in our study), this asymmetry seems to be related to some properties of speech (e.g., complexity, familiarity, causality, etc.). Among them, causality can account for the asymmetric effect of lag adaptation. Speech sound is generated through the movement of articulatory organs. This constrains the order of visual and auditory changes in audio-visual speech signals. Sound never comes first because it is generated as a result of (visible) mouth movement. A recent study revealed that perceived causality in audio-visual stimuli influences synchrony perception [17]. The same might apply to the temporal recalibration. That is, when observers were presented with visual-leading speech (+233 ms condition in this study), they might have perceived a causality between visual and auditory speech signals, leading to the temporal recalibration. On the other hand, when observers were presented with audio-leading speech (–233 ms condition), they might not have perceived a causality, leading to the absense of temporal recalibration. It is noteworthy, however, that the previous study [14] using stream-bounce illusion [18] showed bidirectional shift. To speculate, only the human-origin causality (speech, hand clap, etc.) might modulate temporal recalibration.

In previous studies using off-line adaptation [14,15], exposure time was 3 min. In the current study, the results of Experiment 2 showed a shift in PSS after observation of a fixed time lag only for 10 seconds. This result suggests that temporal recalibration can occur after a short exposure to audio-visual lag. This interpretation is consistent with adaptation studies in other domains. In the visual domain, aftereffect in face stimuli is observed with both long and short adaptation periods (e.g., [19–21]), suggesting that face adaptation occurs over a range of exposure times. Also in audio-visual speech domain, some studies report aftereffects of visual speech following exposure for a couple of trials (e.g., [22]), although the relevant attribute of adaptation is different from our study.

While the PSS shifted after 10 seconds of lag observation, the JND did not change during this short observation period. A shift in the JND was observed after 3 minutes of lag adaptation, which is consistent with previous studies [14,15]. This discrepancy between the results of the PSS and the JND is noteworthy for the discussion on the relationship of these measures. The result can be interpreted as showing that temporal recalibration begins with a shift in the criterion of simultaneity and then is extended to the width of the temporal window, although other studies propose the opposite build-up process (e.g., [23]). Future studies should focus on the build-up process in temporal recalibration.

One might say that the observed effect of lag adaptation comes from the response bias of the participants; their responses might have been biased in the opposite direction of the adapted lag and there might have been neither perceptual change nor integration of auditory and visual inputs. However, a previous study [14] has shown that temporal profile of stream bounce illusion is changed after adaptation to a constant audio-visual lag. Since the responses required in their task do not involve any explicit simultaneity judgement, their result supports the notion that temporal recalbration is a perceptual phenomenon rather than a shift in cognitive criterion. In speech domain, temporal recalibration was shown using indirect method (i.e., the McGurk effect) [24]. Taken together, it seems difficult to account for the observed effects only in terms of the response bias. Rather, our results are consistent with the view that auditory and visual speech signals are integrated and there is a perceptual change.

In the experiments, we presented isolated monosyllabic speech. This paradigm was adopted to compare the temporal recalibration effect more directly between nonspeech [14,15] and speech materials. One might say that monosyllabic speech used in our experiments is not realistic in naturalistic settings. However, a previous study [16] used continuous speech and obtained a temporal recalibration effect. Although there are some methodological differences between these studies, this implies that our findings would also be observed from more realistic continuous speech.

5. Conclusions

In this study, we demonstrated temporal recalibration after exposure to a constant time lag between visual and auditory speech. The PSS shifted after 10 seconds of lag observation, whereas the JND did not change during this short observation period. The width of the temporal window extended only to the direction of audio delay. These findings extend the findings in previous studies and suggest different properties of temporal recalibration in speech.

6. Acknowledgements

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. The first author (A.T.) was supported by the Postdoctoral Fellowships for Research Abroad from the Japan Society for the Promotion of Science. The authors are grateful to Martijn Baart, Shuichi Sakamoto, and Yo-iti Suzuki for their helpful comments on earlier versions of this paper.

7. References

[1] Hirsh, I. J., & Sherrick, C. E. (1961). Perceived order in different sense modalities. *Journal of Experimental Psychology*, 62, 423–432.

[2] Massaro, D. W., Cohen, M. M., & Smeele, P. M. (1996). Perception of asynchronous and conflicting visual and auditory speech. *Journal of the Acoustical Society of America*, 100, 1777–1786.

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

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

[5] Grant, K.W., van Wassenhove, V., and Poeppel, D. (2004). Detection of auditory (cross-spectral) and auditory-visual (cross-modal) synchrony. *Speech Communication*, 44, 43–53.

[6] McGrath, M., & Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *Journal of the Acoustical Society of America*, 77, 678–684.

[7] Pandey, P. C., Kunov, H., & Abel, S. M. (1986). Disruptive effects of auditory signal delay on speech perception with lipreading. *The Journal of Auditory Research*, 26, 27–41.

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

[9] Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience*, 19, 1964–1973.

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

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

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

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

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

[15] 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.

[16] 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.

[17] Eijk, R.L.J. van, Kohlrausch, A.G., Juola, J.F., Par, S.L.J.D.E. van de (in press). Perceived causality in audio-visual stimuli influences asynchrony detection thresholds. *Journal of Experimental Psychology: Human Perception and Performance*.

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

[19] Leopold, D. A., O'Toole, A. J., Vetter, T. & Blanz, V. (2001). Prototype-referenced shape encoding revealed by highlevel aftereffects. *Nature Neuroscience*, 4, 89–94.

[20] Seyama, J. and Nagayama, R. S. (2006). Eye direction aftereffect. *Psychological Research*, 70, 59–67.

[21] Webster, M. A. & MacLin, L. H. (1999). Figural aftereffects in the perception of faces. *Psychonomic Bulletin & Review*, 6, 647–653.

[22] Vroomen, J., Van Linden, S., de Gelder, B. & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory–visual speech perception: Contrasting build-up courses. *Neuropsychologia*, 45, 572–577.

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

[24] Asakawa, K., Tanaka, A., & Imai, H. (2009). Temporal recalibration in audio-visual speech integration using a simultaneity judgment task and the McGurk identification task. *Proceedings of the 31st Annual Meeting of the Cognitive Science Society,* 1669-1673.