The 23rd IEEE International Symposium on Robot and
Human Interactive Communication
August 25-29, 2014. Edinburgh, Scotland, UK,

Temporal Processing on Audiovisual Simultaneity during Perception of Apparent Motion

Jinhwan Kwon, Ken-ichiro Ogawa, Taiki Ogata and Yoshihiro Miyake

Abstract— **The relation between visual motion information and temporal perception has a significant effect on the development of man-machine interface. However, the relation is still not fully understood. This study aims to investigate temporal processing of audiovisual simultaneity during perception of apparent motion, which is the fundamental unit of human motion perception. Participants performed an audiovisual temporal order judgment (TOJ) task under two conditions: apparent motion condition and non-apparent motion condition. Our result shows that visual motion information contributes to the acceleration of visual processing and the increase of temporal resolution in temporal processing of audiovisual simultaneity. Our findings will provide useful information to construct the frame of temporal processing in man-machine interface.**

I. Introduction

The relation between visual motion information and temporal perception is an important key for flexible human behavior in a dynamic environment. The relation further has a significant effect on the temporal interaction between humans and artifacts such as robots in the real world and avatars in virtual worlds. The artifacts therefore should be designed based on human perceptual system so that human naturally interact with artifacts. However, the relation is still not fully understood. In this study, we aim to investigate how visual motion information affects human temporal perception. In particular, we focus on apparent motion, which is the fundamental unit of human motion perception, as visual motion information.

Visual motion information is used for various display systems such as film and television. Although such display systems deliver a discrete temporal sequence of static views, we cannot avoid perceiving it as continuous moving images under a specific condition (e.g., 24 fps in films, 30 fps in television), so-called “window of visibility” [1]. This phenomenon is attributed to apparent motion. Apparent motion is a visual phenomenon that makes continuous motion appear with an appropriate spatiotemporal interval despite two discrete stimuli [2], [3] and well represents the specific characteristics of human motion perception. In particular, apparent motion is systematically influenced by the temporal interval between the stimuli. When the temporal interval is too short, both the stimuli are perceived as simultaneousness. Whereas, when the temporal interval is too long, both the stimuli are perceived as successiveness. Namely, when the temporal interval between the two stimuli is too short or long, continuous motion cannot be perceived. For example, when the stimulus onset asynchrony (SOA) of two visual stimuli is within a range of 50 to 150 ms the two visual stimuli are perceived as a continuous motion. Conversely, the visual stimuli are perceived as successive beyond an SOA of 300 ms [4], [5], [6].

In this study, to investigate the relation between visual motion information and temporal perception, we performed temporal order judgment (TOJ) tasks as a psychophysical experiment for examining temporal factors in multisensory processes [7], [8]. The TOJ task is known as a way to measure how human perceive temporal synchrony between two types of senses. In TOJ tasks, a point of subjective simultaneity (PSS) and a just noticeable difference (JND) are calculated as statistic quantities. The PSS represents a specific interval between the applications of two sensory stimuli at which both are perceived at the same time. The JND represents temporal resolution for identifying the simultaneity [8].

Two experiments were conducted to investigate whether visual motion information affects audiovisual TOJ tasks. In Experiment 1, participants conducted a TOJ task for audiovisual simultaneity in the apparent motion condition with two flashes and in the normal condition with a single flash. In Experiment 2, we eliminated the influence of prediction from the TOJ task by randomly presenting the intervals of two visual stimuli because there remained an influence of specific prediction as a higher-order brain function from the constant interval between two flashes.

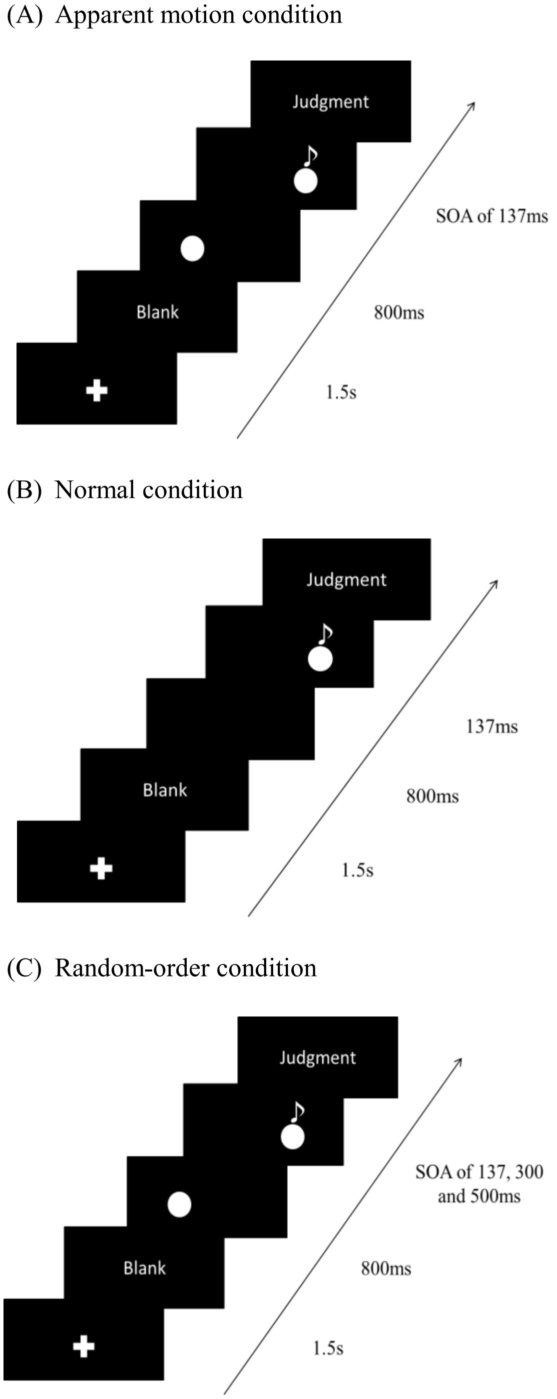
II. Methods

A. Participants

Sixteen participants (15 males and one female, with a mean age of 24.3 years) participated in Experiment 1. Twelve participants (10 males and two females, with a mean age of 23.9 years) took part in Experiment 2. All participants had normal hearing and normal or corrected-to-normal visual acuity and were naive as to the purpose of the experiment. Participants were paid for taking part in the experiment and written informed consent was obtained. This experiment was approved by the ethics committee of the Tokyo Institute of Technology.

J. Kwon, K. Ogawa, Y. Miyake are with Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, 4259 Nagatsuda, Midori, Yokohama 226-8502, Japan (e-mail: kwon@myk.dis.titech.ac.jp, ogawa@dis.titech.ac.jp, miyake@dis.titech.ac.jp)

T. Ogata is with Research into Artifacts, Center for Engineering(RACE), The University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa-shi, Chiba, 277-8568, Japan (e-mail: ogata@race.u-tokyo.ac.jp)



(A) Apparent motion condition

(B) Normal condition

(C) Random-order condition

Fig. 1 Schematic illustration of Experiment 1 and Experiment 2. The two conditions in Experiment 1: Apparent motion condition with a SOA of 137 ms (A) and Normal condition with single flash (B). The condition in Experiment 2: Random-order condition with a SOA of 137 ms (apparent motion condition) and SOAs of 300 and 500 ms (successive condition).

B. Apparatus and stimuli

All TOJ task experiments were conducted in a dark and soundproof room. Visual stimulation was provided by a 27-inch LCD display (Samsung S27A950D, Korea) with a screen resolution of 1920×1080 pixels, and a refresh rate of 120 Hz. The display was operated by a PC workstation (Mac pro, 3.2GHz Quad-Core Intel Xeon, ATI Radeon HD 5770 graphic card, 1GB GDDR5 memory, US) and placed in front of the subjects. Their head position was fixed by a chin rest at a viewing distance of 100 cm. A white cross of 2 cm in length was displayed as a fixation point in the center of the screen. Visual stimuli consisted of one or two white disks 3.2 cm in diameter on a black background. The visual angle was 2.8° for the single stimulus and 5.6° for the two stimuli. Sound stimuli were presented as mono sounds (65dB, 1,000Hz) delivered via two speakers (MM-SPWD3BK, Sanwa supply, Japan). The speakers were located on top of the screen. These visual and auditory stimuli were developed and operated by a computer program (Matlab and Psychtoolbox-3, US).

C. Procedure

In Experiment 1, the participants sat on a chair facing visual stimuli, and their head position was maintained by the chin rest. Then, audiovisual TOJ tasks were performed under two conditions with visual stimuli: the apparent motion condition and the normal condition. Figs. 1(A) and 1(B) illustrate the procedure for Experiment 1. In the apparent motion condition (Fig. 1(A)), each trial began with the presentation of a fixation cross for 1.5 seconds, followed by a dark blank screen for 800 ms. Next, one white circle as the first visual stimulus was displayed for 30 ms; then with a SOA of 137 ms, the second stimulus was presented for 30 ms [9]. To assess the temporal discrimination of the auditory and visual stimulus pairs, one brief sound (30ms) as an auditory stimulus was presented at different times relative to the second visual stimulus. The participants were instructed to conduct a TOJ task between the second visual stimulus and the audio stimulus. The onset time of the auditory stimulus paired with a visual stimulus was randomly selected from the following SOA values: –120, –90, –60, –30, 0, +30, +60, +90, and +120 ms. Here the negative values indicate that the auditory stimulus preceded the visual stimulus. Then the participants made a forced-choice judgment with respect to the order of the audiovisual stimuli by answering the question ‘which one came first?’ The answers consisted of ‘light first’ and ‘sound first’, which were chosen by pressing the Z key and the X key, respectively. The response ‘light first’ was selected when the flash was ahead of the sound, and it is the same with ‘sound first.’ As shown in Fig. 1(B), the procedure for the normal condition was the same as that for the TOJ task in the apparent motion condition. The only one difference was that the first visual frame was not presented. Then, the same procedure for evaluating the temporal discrimination between the auditory stimulus and the visual stimulus, and the same SOA values were used as in the apparent motion condition. Experiment 1 consisted of 270 trials (2 visual conditions × 9 audiovisual

SOAs × 15 repeats) in a counterbalanced order. The participants performed 27 trials (9 audiovisual SOAs × 3 repeats) as one block for each condition.

In Experiment 2, the apparatus, stimuli, and procedure were the same as shown in Experiment 1. In Experiment 2, however, only the random-order condition was studied. In the random-order condition, the participants conducted TOJ tasks with SOAs between the visual stimuli of 137 ms, 300 ms, and 500ms presented in a random order. That is, we set up two kinds of TOJ tasks in the apparent motion condition with an SOA of 137 ms between the two flashes, which is the same spatiotemporal interval as in Experiment 1, and in a successive condition with SOAs of 300 and 500 ms between the two flashes, which are perceived as successive stimuli without motion perception. The timing of the auditory stimulus relative to the second flash was the same as in Experiment 1. The participants were instructed to judge the order of the second visual stimulus and the auditory stimulus. Experiment 2 consisted of 432 trials (3 visual conditions × 9 audiovisual SOAs × 16 repeats) with a counterbalanced order. The participants performed 54 trials (3 visual conditions x 9 audiovisual SOAs x 2 repeats) as one block for each condition and only the data of apparent motion was calculated in Experiment 2. The practice of each experiment was conducted and the total performance took about one and a half hours in each experiment.

Before starting each experiment, we examined whether the participants perceived motion between two flashes, and we confirmed that the motion was perceived during the TOJ task after each experimental session was completed.

D. Data analysis

The ratio of the answers indicating the earlier presentation of the auditory stimulus was calculated for each SOA. We conducted logistic regressions using a generalized linear model with the ratio data of each experiment [10]. Fig. 2 shows a logistic regression curve for data analysis of an audio–visual temporal order judgment. The following equation was applied to the regression analysis:

$$y = \frac{1}{1 + e^{\frac{(\alpha - x)}{\beta}}} \qquad (1)$$

where α represents the estimated PSS, x denotes SOA, and β shows the temporal unit for temporal difference of numerator as $\alpha - x$. JND is defined as:

$$JND = \frac{X_{75} - X_{25}}{2} = \beta \log 3 \qquad (2)$$

where X_p represents the SOA with p percent of 'auditory first' responses. We determined the JND and PSS values for each participant using regression analysis (Eqs. (1) and (2)) and processed the data statistically to obtain mean and standard error values.

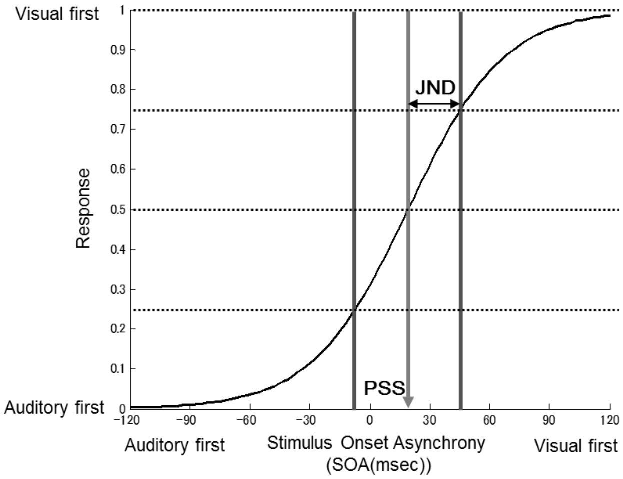


Fig. 2 Logistic regression curve for data analysis of an audio–visual TOJ task. The point of subjective simultaneity (PSS) represents a specific interval between the applications of two sensory stimuli at which both are perceived at the same time. The just noticeable difference (JND) represents temporal resolution for identifying the simultaneity.

III. Results

A. Experiment 1

Fig. 3 presents the results of Experiment 1. As shown in Fig. 3(B), the PSS in the normal condition had a positive value, 12.47 ms (SE = 6.45), but the PSS in the apparent motion condition shifted to a negative value, –4.90 ms (SE = 5.84). The PSS of negative values indicates that the audiovisual stimulus pairs were perceived as simultaneous when the auditory stimuli preceded the visual stimuli. A paired t-test of PSSs indicated a significant difference between the TOJ task in the apparent motion condition and that in the normal condition (t(15) = –2.33, $P < 0.05$). In addition, the JND in the apparent motion condition was smaller than that in the normal condition (see Fig. 3(C)), and the JND values were 35.72 ms (SE = 3.96) and 48.23 ms (SE = 5.17), respectively. A significant difference between the JNDs was observed in the paired t-test (t(15) = –3.57, $P < 0.01$).

B. Experiment 2

In Experiment 2, all participants performed the TOJ task with the intervals between the visual stimuli in a random order, and only the results of the apparent motion condition were extracted. The participants perceived continuous motion, and the PSSs and JNDs were computed as in Experiment 1. Fig. 4 shows the results of Experiment 2, Fig. 4(B) and 4(C) show the results for PSSs and JNDs in Experiment 2. The values of PSS and JND in the apparent motion condition of the random-order condition were almost the same as those in the apparent motion condition in Experiment 1. An unpaired t-test of PSSs and JNDs of the TOJ tasks in the apparent motion condition indicated no significant difference between Experiment 1 and Experiment 2 (t(26) = –0.11, P = 0.92, t(26) = –0.12, P = 0.91).

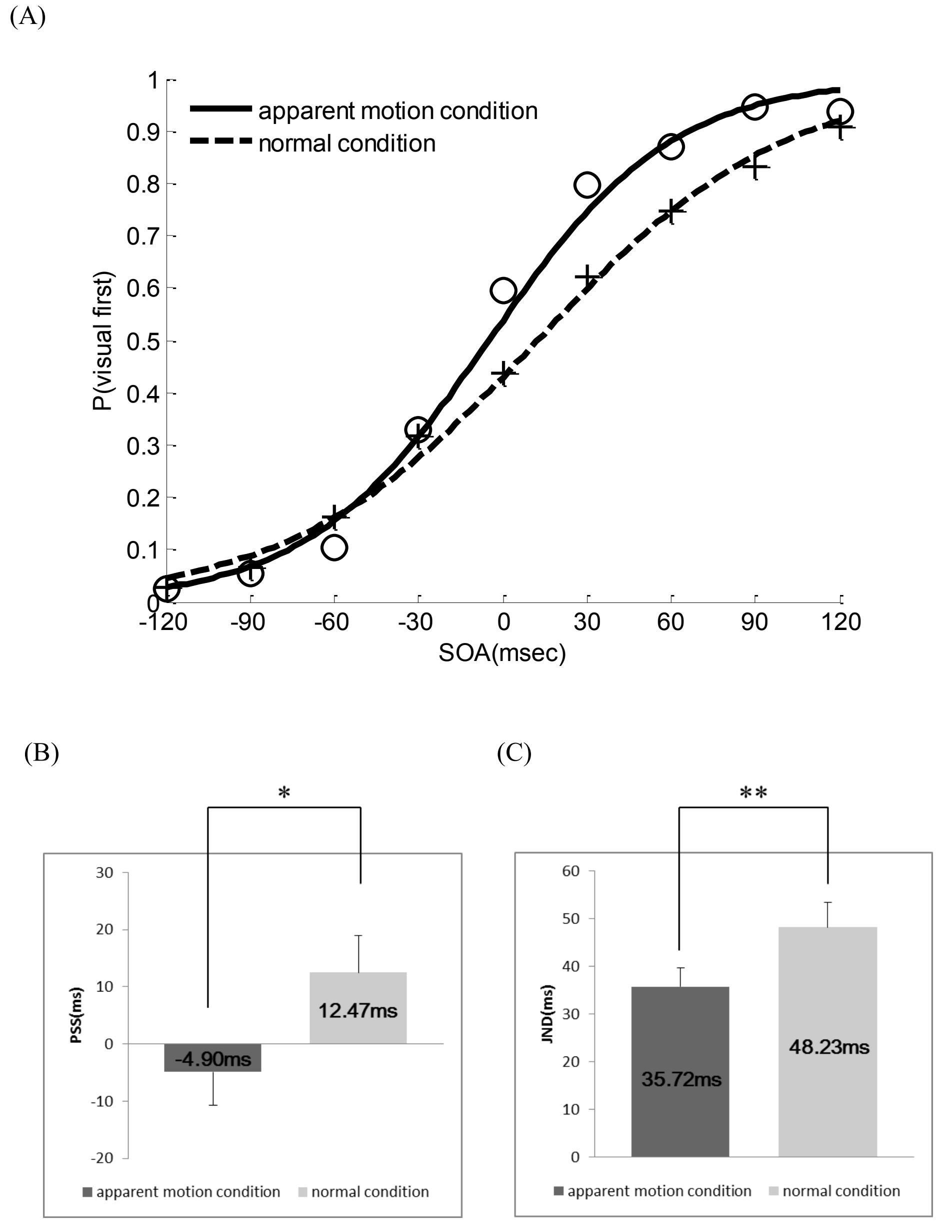


Fig. 3 The results of Experiment 1. (A) Psychometric curves fitted to the distribution of the mean TOJ data in Experiment 1. (B) Mean PSSs and JNDs in the apparent motion condition and the normal condition. The error bars represent the standard error of the means, * : P < .05, ** : P < .01, *** : P < .001.

IV. Discussion

In the apparent motion condition, the PSS was shifted to a sound-lead stimulus and the JND was smaller. Previous studies have reported that the PSS usually shifts toward a visual-lead stimulus within a range of 20-40ms, and therefore simultaneity is maximally perceived if light comes slightly before sound [11], [12], [13], [14]. However, the PSS in apparent motion condition was shifted to a sound-lead stimulus that is the opposite result in the normal condition in Experiment 1. The sound-lead value of PSS indicates a possibility that visual processing was faster. With respect to temporal resolution, it is known that the JND is within a range of 30-60 ms in audiovisual TOJ tasks [15], [16], [17]. We however found that visual apparent motion resulted in the smaller JND, which indicates higher temporal discrimination. These results show that motion information differs from non-motion information on temporal processing in multisensory integration.

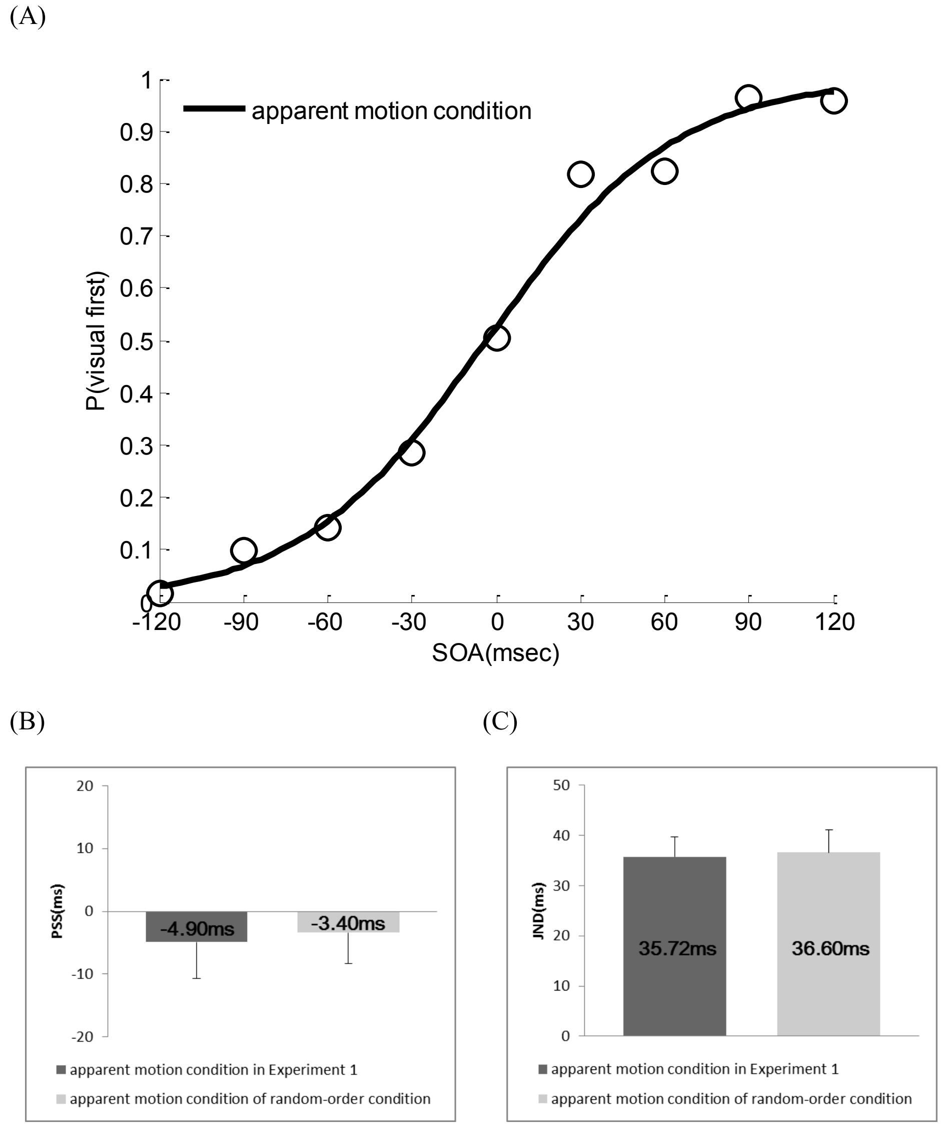


Fig. 4 The results of Experiment 2. (A) Psychometric curves fitted to the distribution of the mean TOJ data in Experiment 2. (B), (C) Mean PSSs and JNDs in the apparent motion condition in Experiment 1 and in the apparent motion condition of the random-order condition. The error bars represent the standard error of the means.

In Experiment 1, we found that motion perception influences temporal perception on audiovisual processing. However, there remained an influence not only of apparent motion but also of specific prediction as some top-down factor, because the interval of two flashes in the apparent motion condition was constant. That is, the prediction effect caused by the constant interval itself may influence the temporal perception on audiovisual processing. Therefore, it is necessary to conduct a supplementary experiment in which the interval is changed randomly. In Experiment 2, the participants could not predict the presence or absence of apparent motion. Nevertheless, the result showed that motion information was equivalently processed regardless of prediction as a top-down factor. It is known that predictable information improves the temporal resolution [18] and allocates a faster time course for motion processing [19]. However, this result of unpredictable apparent motion did not differ from that of predictable apparent motion. Thus, we could eliminate the effect of prediction as a top-down factor.

Alternatively, the pathway of motion and non-motion processing may differ in audiovisual temporal perception, and there may exist some activation for determining the pathway of motion and non-motion processing by bottom-up signals. Many researchers have been claimed that audio-visual stimulation was integrated at an early processing stage [20], [21]. Fendrich and Corballis [20] suggested that the sensory capture phenomenon may be connected with low-level inter-sensory linking processes and it seems that auditory driving or auditory dominance depends on such low level sensory linking processes. Moreover, Bruns and Getzmann [21] also reported that their findings are consistent with a low-level audiovisual integration between visual apparent motion and single sound. In this study, the same properties on temporal perception appear between predictable motion (in Experiment 1) and unpredictable motion (in Experiment 2) on audiovisual processing. Therefore, we can conclude that the pathway for motion and non-motion processing is determined by the bottom-up signals. With respect to brain function on audiovisual processing, it has been reported that the superior temporal sulcus (STS) have been suggested as an audiovisual association area [22], [23], [24]. However, in recent years, with the growing interest in multisensory properties of motion, some researchers have raised a possibility that the area MT playing an important role in visual motion processing is engaged in the audiovisual processing [24], [25], [26], [27]. This fact may lead us to the mechanism which decides the pathway of motion or non-motion processing on audiovisual temporal perception by bottom-up signals.

The synchronization of audio-visual signals is one of the important factors for multimedia content and artifacts. However, audio-visual desynchronization sometimes occurs in multimedia applications and artifacts. The PSS indicates the optimal simultaneity on cross-modal processing and JND shows a threshold to perceive asynchrony on cross-modality. Therefore, it is necessary to be controlled by the simultaneous range of 36 ms (from JNDs in the apparent motion condition in Experiment 1 and 2) on the basis of sound-lead stimulus, approximately 4 ms from PSSs in apparent motion condition in Experiment 1 and 2, for audiovisual simultaneity considering visual motion information. On the other hand, the visual-lead stimulus within a range of 20-40ms and the simultaneous range of 30-60 ms is required for the audiovisual simultaneity in the case of non-motion information [11]-[17].

Because errors in audio-visual synchronization cause poor interaction between human and artifacts, the cross-modal information is important for the designs of artifacts. Our findings can be applied to multisensory integration technology for robots. Current robots process the external information depending on each sensory stimulus such as spatial information via visual signals, dialog information through auditory signals. However, the researches for creating robots capable of multisensory integration are in progress. Our findings of PSS and JND will provide an important time scale for integration of multisensory information in robot design.

In the present study, it was found that motion perception correlates with temporal perception in audiovisual processing. This motion perception resulted in faster processing and higher temporal resolution in audiovisual temporal perception relative to non-motion processing. Moreover, the effect of motion perception on temporal perception has shown automatic processing mechanisms regardless of prediction. It may be caused by the activation processing separately motion and non-motion information. Such activation on the temporal perception of multisensory processing should be incorporated in artifacts such as robots and avatars for better communication with human.

Acknowledgment

We would like to thank Mr. JongHwan Kim for assistance in data programming.

References

[1] A.B. Watson, A.J. Ahumada and J.E. Farrell, “Window of visibility: a psychophysical theory of fidelity in time-sampled visual motion displays,” Optical Society of America Vol. 3, No. 3, pp.300-307, 1986.

[2] V.S. Ramachandran and S.M. Anstis, “The perception of apparent motion,” Scientific American 254(6), pp.102–109, 1986.

[3] A. Larsen, J.E. Farrell and C Bundesen, “Short- and Long-Range Processes in Visual Apparent Movement,” Psychol Res 45, pp.11–18, 1983.

[4] M.R.W. Dawson, “The how and why of what went where in apparent motion: modeling solutions to the motion correspondence problem,” Psychological Review 98: 569–603, 1991 .

[5] S. Getzmann, “The effect of brief auditory stimuli on visual apparent motion,” Perception 36:1089–1103, 2007.

[6] T.Z. Strybel, C.L. Manligas, O. Chan and D.R. Perrott “A comparison of the effects of spatial separation on apparent motion in the auditory and visual modalities,” Perception & Psychophysics 47: 439–448, 1990.

[7] S. Grondin, “Timing and time perception: a review of recent behavioral and neuroscience findings and theoretical directions,” Atten Percept Psychophys 72, pp.561–582, 2010.

[8] J. Vroomen and M. Keetels, “Perception of intersensory synchrony: A tutorial review,” Attention, Perception, & Psychophysics 72 (4), pp.871–884, 2010.

[9] V. Harrar, R. Winter and L.R. Harris, “Visuotactile apparent motion,” Perception & Psychophysics 70(5), pp.807–817, 2008.

[10] D.J. Finney, “Probit analysis: A statistical treatment of the sigmoid response curve,” Cambridge Univ. Press, 1952.

[11] C. Spence and C. Parise, “Prior-entry: A review,” Consciousness and Cognition 19, pp.364–379, 2010.

[12] M. Zampini, S. Guest, D.I. Shore and C. Spence, “Audio-visual simultaneity judgments,” Perception & Psychophysics 67(3), pp.531–544, 2005.

[13] M. Kanabus, E. Szelg, E. Rojek and E. Pöppel, “Temporal order judgement for auditory and visual stimuli,” Acta Neurobiol. Exp 62, pp.263–270, 2002.

[14] P. Jakowski, F. Jaroszyk and D. Hojan-Jezierska, “Temporal-order judgments and reaction time for stimuli of different modalities,” Psychol Res 52, pp.35–38, 1990.

[15] S. Morein-Zamir, S. Soto-Faraco and A. Kingstone, “Auditory capture of vision: Examining temporal ventriloquism,” Cognitive Brain Research 17, pp.154–163, 2003.

[16] M. Keetels and J. Vroomen, “The role of spatial disparity and hemifields in audio–visual temporal order judgments,” Experimental Brain Research 167, pp.635–640, 2005.

[17] M. Zampini, D.I. Shore and C. Spence, “Audiovisual temporal order judgments,” Experimental Brain Research 152, pp.198–210, 2003.

[18] K. Petrini, M. Russell and F. Pollick, “When knowing can replace seeing in audiovisual integration of actions,” Cognition 110, pp.432–439, 2009.

[19] L. Busse, S. Katzner and S. Treue, “Temporal dynamics of neuronal modulation during exogenous and endogenous shifts of visual attention in macaque area MT,” Proceedings of the National Academy of Sciences of the United States of America, 105(42), pp.16380–16385, 2008.

[20] R. Fendrich and P.M. Corballis, “The temporal cross-capture of audition and vision,” Perception & Psychophysics 63(4), pp.719-725, 2001.

[21] P. Bruns and S. Getzmann, “Audiovisual influences on the perception of visual apparent motion: Exploring the effect of a single sound,” Acta Psychologica 129, pp.273–283, 2008.

[22] B.E. Stein and T.R. Stanford, “Multisensory integration: current issues from the perspective of the single neuron,” Nature Reviews Neuroscience vol. 9, pp.255–266, 2008.

[23] M.S. Beauchamp, “See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex,” Curr Opin Neurobiol, 15, pp.1–9, 2005.

[24] I.R. Olson, J.C. Gatenby and J.C. Gore, “A comparison of bound and unbound audio–visual information processing in the human cerebral cortex,” Cognitive Brain Research 14, pp.129–138, 2002.

[25] H. Kafaligonul and G.R. Stoner, “Auditory modulation of visual apparent motion with short spatial and temporal intervals,” Journal of Vision 10(12), 31, pp.1–13, 2010.

[26] G.A. Calvert, et al, “Response amplification in sensory-specific cortices during crossmodal binding,” Neuroreport, 10, pp.2619–2623, 1999.

[27] R.T. Born and D.C. Bradley, “Structure and function of visual area MT,” Annual Review of Neuroscience, 28, pp.157–189, 2005.