Effect of auditory and visual stimuli on distance perception presented by head-mounted display and headphone

Takumi Asakura* and Atsuya Ishikawa

Tokyo University of Science, 2641, Yamazaki, Noda, 278–0022 Japan

(Received 7 December 2018, Accepted for publication 4 April 2019)

Abstract: This paper investigates the relationship between the effects of auditory and visual stimuli on the distance perception between the subjects themselves and the sound sources. As the auditory stimuli, the effect of the male speech voices and the binaural room impulse responses on the distance perception are comparatively investigated. The relative error of the estimated distances in various audiovisual conditions are discussed. It was found that the subjective impression of the distance to the sound source was effected by both the auditory and visual information, however the extent of the distance perception varied depending on the time variant characteristics of the reproduced sounds.

Keywords: Distance perception, Visual and auditory stimuli, Head-mounted display, Virtual reality, Binaural simulation, Reverberation

PACS number: 43.66.Lj [doi:10.1250/ast.40.265]

1. INTRODUCTION

The virtual-reality technologies have been greatly developed in these days, and various kinds of sound reproduction methods such as the binaural reproduction and multi-channel reproduction are available. On the other hand, from the viewpoint of visual reproduction, more realistic images can also be easily available by using such developing hardwares as the head-mounted display (hereafter called HMD). While further realistic and immersive reproduction of the fields are required in many industrial sectors, various kinds of technological development are performed to attain high precision and quality for the audiovisual reproduction. Fujisaki et al. has reported that the human recognizes various kinds of events occurring in the real world by integrating the auditory and visual information [1]. Among these two modalities, Zhou et al. has confirmed that the scaled 3-D sound significantly increases the accuracy of depth judgments inside virtual reality environments [2].

In order to experience realistic field in the virtual spaces, it may be important to model accurately the visual and acoustic conditions. In particular, the distance perception between the location of the participant and each virtual objects inside the virtual spaces may be influenced by the visually and auditorily perceived subjective information.

The contributions by the visual and auditory perception against the distance perception between the objects are influenced by physical situation of each field, and each factors are not clearly indicated in the previous researches. Mershon et al. [3] investigated the influence of the intensity and reverberation on the auditory perception of egocentric distance. Lin et al. [4] investigated the accuracy of distance perception in near field augmented reality visual targets viewed by stereoscopic glasses. Spagnol et al. [5] proposed a filter model of near-field effects based on the distance variation function. They also investigated [6] how temporal order affects distance discrimination of receding and approaching pairs of sound sources rendered binaurally in the anechoic near-field. Rungta et al. [7] investigated the effects of reverberant sounds generated using different propagation algorithms on acoustic distance perception. As can be seen in these investigations, the distance perception has been the subject of interest from the viewpoints of auditory or visual perception, however multiple effect of the auditory and visual perception on the distance perception has not been much investigated. Although subjective influences of the auditory and/or visual stimuli on the distance perception are investigated [8], only speech sounds were convolved with simulated impulse responses and used as auditory stimuli. The subjective effect of the visual and auditory stimuli including more impulsive sounds largely influenced by the reverberation of the spaces should be comparatively

*e-mail: t.asakura@rs.tus.ac.jp
investigated together with such quasi-stationary sounds like speeches.

The basic purpose of this paper is to investigate the relationship between the effects of auditory and visual stimuli on the distance perception between the location of subject and the sound source. As the auditory stimuli, the effect of the male speech voices and the auralized binaural room impulse responses (BRIR) are comparatively investigated, whereas actually recorded panoramic video and virtually constructed panoramic video of 3D space are comparatively investigated as the visual ones. The relative error of the estimated distances in various audiovisual conditions are discussed.

2. METHOD OF DISTANCE ESTIMATION EXPERIMENT

To investigate the effect of binaural sounds and panoramic videos, which includes the audiovisual information on the sound source, on the distance perception between the locations of the subjects and the sound sources, the following subjective evaluation experiment was carried out.

2.1. Acquisition of the Auditory and Visual Data

In this section, the detailed scheme to obtain the auditory and visual data used in the experiment are described. Firstly, the data acquisition method of the auditory signals is shown. The impulse responses between the source and receiving points were measured in the space of a corridor in the campus of Tokyo University of Science as shown in Fig. 1(a). As indicated in the figure, the dummy-head simulator (ACO, SAMURAI HATS Type 7828 B) was located at the receiving point R which varied at the distances of 2 m, 5 m, 10 m, 20 m, and 30 m from the source point S with the loudspeaker. In the measurement of the impulse responses, time-stretched pulse (TSP) signals with the data duration of $2^{21}$ in 48 kHz sampling were used to decrease the influence of the background noise. The impulse responses were obtained through convolution between the inverse filter of the adopted TSP signal and the measured TSP responses. Because the loudspeaker has non-flat frequency characteristics of acoustic radiation, the frequency components of the impulse responses were compensated to have flat frequency characteristics in the frequency domain from 100 Hz to 12.5 kHz octave bands by equalization. The obtained two-channel impulse responses in each measurement condition were used in two ways for the subjective evaluation experiment. The impulse response was converted to wave files, and were directly reproduced to the subjects in the first way whereas the convolved sound files with the impulse responses and the male voice data in Japanese words having time duration of 10 seconds were reproduced to the subjects as reverberant sounds in the second way.

Next, the data acquisition method of the visual data is described. In the subjective evaluation experiment mentioned later, two kinds of visual stimuli were used; one is the panoramic video recorded at the receiving point of Fig. 1(a), and the other is the panoramic video projecting the scenery inside the same space as the former one in the corridor with one loudspeaker set in front of the receiving point which was made by three-dimensional computer graphics software (The Unity, 3-D modelling software). The snapshots of one part of the above-mentioned real and virtual panoramic videos are shown in Figs. 2(a) and 2(b), respectively. The real videos were recorded by using the panoramic camera (Galaxy Gear 360) in each condition of the distance between the source and the receiving points.

2.2. Procedure of the Subjective Evaluation Experiment

The subjective evaluation experiment was performed to investigate the subjective effect of presenting the immersive panoramic video by HMD and the binaural sounds by headphones against the distance perception. To present the audiovisual data to the subjects, a headphone (AUDIO-TECHINICA, ATH-W1000Z) and an HMD (OCULUS, Oculus Rift cv1) shown in Fig. 2(c) were used. In the experiment, two kinds of sound data; the measured impulse responses and the male voice, and two kinds of visual data; the recorded panoramic video and the three-dimensional computer-graphic video were adopted. The auditory data of the impulse response and the male voice have time durations of 2 s and 10 s, respectively. The Schroeder’s decay curves and the energy decay curves of the impulse
responses in the five conditions of the source-receiver distances are shown in Figs. 3(a) and 3(b), respectively. In addition, their reverberation times are shown in Table 1 which was evaluated using the integrated impulse response method over a 10-dB ($T_{10}$, from 0 dB to $-10$ dB below the starting level) or 30-dB ($T_{30}$, from $-5$ to $-35$ dB below the starting level) decay ranges in the reverberation curve. These reverberations have a little different decay characteristics between the early- and late-time domains. The early-time reverberation times have relatively shorter values than the late-time reverberation times in all conditions. On the other hand, the sound source of the male voice was reproduced by assuming not such a situation of face-to-face conversation but a situation that one person speaks to relatively large number of people is assumed. For this reason, the male voice data in condition of the nearest position of sound source at a two-meter distance was presented to the subject at the reproduction level that satisfies $L_{Aeq}$ of 60 dB at the ear part. This reproduction level was set by following the Ref. [9] which indicated that the average voice level of 27 teachers was estimated to be 65.3 dB A at a one-meter distance, which was converted into 59.3 dB A at a two-meter distance by assuming the speaker’s voice as a point source.

The subjective evaluation experiment was performed by presenting the audiovisual data as shown in Table 2(a), (b) and (c) which are combination of the conditions of five kinds of source-receiver distances, two visual conditions of recorded and CG visual presentation, and two auditory conditions of impulse response and male voice. The experiment was composed of three parts. In the first part, the auditory or visual stimuli were independently presented in random order. The subjects answered their perceived distance between their position and the loudspeaker by
using integer numbers. It should be noted that the actual loudspeaker (Fig. 4) which were set at a one-meter distance from the subject were presented to the subjects before the experiment. It was because the presentation of the actual size of the loudspeaker shown as the visual stimuli gives a clue to judge the distance to the loudspeaker from the subjects’ location. There are another way of not indicating the actual size of the loudspeaker, but in that situation, it is considered to be difficult to judge the distance. For such a reason, the actual loudspeaker set at a one-meter distance from the subject (Fig. 4) was presented to them before the experiment. The experiment of the first part includes 20 conditions which are combinations of four kinds of audiovisual stimuli and five kinds of source-receiver distances (2 m, 5 m, 10 m, 20 m, and 30 m). In the second part, the auditory and visual stimuli were simultaneously presented. The experiment of the second part includes 20 conditions which are combinations of two kinds of auditory stimuli, two kinds of visual stimulus conditions, and five kinds of source-receiver distances. It should be noted that the source-receiver distance of respective auditory and visual conditions were made to be the same with each other in this second experiment. In contrast, the auditory and visual stimuli which have different (mismatching) source-receiver distances with each other were evaluated in the third part of the experiment. In addition to these experiments, in order to estimate the influence of the personal distance-sense ability of each subject on the experimental results, the distance sense from the subject to the same speaker as that in the main experiments located in a real situation were measured. In this additional measurement, the arranged distance between the subject and the target

Table 2 Investigated conditions of the (a) Part 1, (b) Part 2 and (c) Part 3 of the subjective experiment.

| Part 1 | Condition number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---------|------------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sound   | Male Voice       |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Visual  | —                |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Source-receiver distance [m] | Sound | 2 | 5 | 10 | 20 | 30 | 2 | 5 | 10 | 20 | 30 | 2 | 5 | 10 | 20 | 30 | 2 | 5 | 10 | 20 | 30 |
| Visual  | 2 | 5 | 10 | 20 | 30 | 2 | 5 | 10 | 20 | 30 | 2 | 5 | 10 | 20 | 30 | 2 | 5 | 10 | 20 | 30 |

| Part 2 | Condition number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---------|------------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sound   | Male Voice       |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Visual  | Male Voice       |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Source-receiver distance [m] | Sound | 2 | 5 | 10 | 20 | 30 | 2 | 5 | 10 | 20 | 30 | 2 | 5 | 10 | 20 | 30 | 2 | 5 | 10 | 20 | 30 |
| Visual  | 2 | 5 | 10 | 20 | 30 | 2 | 5 | 10 | 20 | 30 | 2 | 5 | 10 | 20 | 30 | 2 | 5 | 10 | 20 | 30 |

| Part 3 | Condition number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---------|------------------|---|---|---|---|---|---|---|---|---|---|
| Sound   | Imp. Resp.       |   |   |   |   |   |   |   |   |   |   |
| Visual  | Virtual          |   |   |   |   |   |   |   |   |   |   |
| Source-receiver distance [m] | Sound | 2 | 5 | 10 | 20 | 30 | 2 | 5 | 10 | 20 | 30 |
| Visual  | 5 | 10 | 10 | 20 | 5 | 20 | 10 | 30 | 10 | 20 | 20 |

Fig. 4 Adopted loudspeaker in the experiment.
speaker were set as 2.2 m, 5.5 m, 8.5 m, 12.5 m, 21.5 m, and 28.5 m, which were set as different intervals from the abovementioned setting of 2 m, 5 m, 10 m, 20 m, and 30 m to avoid easier estimation of the distance.

Ten male subjects in their twenties participated in the experiment. Throughout the experiment, the same ten subjects were used, and they participated in the first, second, and third part of the experiment in this order. If the subjects haven’t experienced all the audiovisual stimuli before the main experiment, the evaluated results may be effected by the presentation order of the stimuli in each part, which were randomly presented to the subject, because they gradually change the criteria of the judgment of the egocentric distance as the progress of the experiment. In order to avoid the effect of such a presentation order of each stimulus as possible, the subject was presented all the conditions from No. 1 to No. 20 used in the second part of the experiment before the experiment of the first part. In addition, it was sufficiently instructed to each subject to evaluate each condition independently, and not to evaluate each condition by comparing with each other.

3. RESULTS AND DISCUSSION

3.1. Effect of Separately Presented Auditory or Visual Stimuli

The results of the experiment of the first part is shown in Fig. 5 as the relationship between the relative error of the averaged value of the perceived source-receiver distance with their standard deviation and the correct distance. The relative error is calculated as

\[
\text{Relative error} = \frac{D_{\text{per}} - D_{\text{corr}}}{D_{\text{corr}}} \cdot 100 \% \quad (1)
\]

where \(D_{\text{per}}\) is the averaged value of the perceived source-receiver distance, and \(D_{\text{corr}}\) is the correct source-receiver distance.

In Fig. 5(a) with only the visual stimuli, the tendency of the real image indicates relatively correct perception of the distance in the conditions of 2 m, 5 m, and 10 m while those of the conditions over 20 m indicate a little larger error of misperception to nearer position. The tendencies of the error in the conditions of real and virtual images indicates almost the same characteristics, but in the conditions with further source-receiver distance, the conditions presented with the virtual ones indicate slightly larger error. It may be considered to be effected by the difference of the number of clues including some doors, ceiling lights and so on. On the other hand, the estimated results of the distance in Refs. [8] and [10] are additionally shown in the figure. In these studies, the egocentric distances between their position and the target objects (loudspeaker) were evaluated, which are located in a corridor space [8] or on the stage of a concert hall [10]. It should be noted that the visual stimuli were presented on the visual display of a stereoscopic screen [8] or a large flat screen HDTV [10]. The referred results indicate relatively underestimated values compared to the obtained results in this paper. One of the reasons for the difference of the error of the evaluated distance may be that the additional information on the actual size of the object (loudspeaker) instructed to them before the experiment was used as a clue to estimate the distance to the object. In addition, the target corridor of the distance evaluation in this paper was such a space that are familiar to all the subjects.

In contrast to these results of the egocentric distances evaluated by giving visual stimuli by the HMD, the measured results of the egocentric distances in the real corridor space are shown as a blue solid line in Fig. 5(a). It indicates that the subjects can also evaluate accurately the distance in comparison to the virtual situation. Based on this result, it can be considered that the variation of each subject’s ability of distance sense has little impacts on the experimental results.

On the other hand, in Fig. 5(b) with only the auditory stimuli, the tendency of the conditions with impulse response and male voice indicates quite difference charac-
teristics with each other. The tendency of the male voice indicates relatively correct distance in the two-meter condition while the farther conditions show relatively larger error of misperception to nearer position than the correct one. Herein, the obtained results in the Refs. [8] and [10] are also shown in the figure by black solid and broken lines, respectively. The Ref. [8] adopted a sound signal which were obtained by convolving a dry source of male voice and the BRIR, whereas the Ref. [10] adopted a sound signal which were obtained by convolving a dry source of Gaussian noise with a duration of 100 ms and the BRIR. Then, the results of the Ref. [8] show similar tendencies to those of the experiment in this paper, because they have similar time-transient characteristics of voice. In contrast, the results of the conditions with impulse response indicate large errors of misperception to farther position in the conditions of 2 m and 5 m while those of the conditions over 10 m conversely indicate those to nearer position. The results of the Ref. [10] show a similar tendency to the results of this paper from the viewpoint of overestimation in the near-distance conditions of under 3 m, whereas those show underestimated values in the conditions of over 5 m, which are conversely similar to the conditions of the male voice, not to the condition of the impulse response. The reasons for the abovementioned tendency of misperception for each sounds are considered as follows. Firstly, in comparison between the results presented by transient (impulse response) and non-transient (male voice) sounds, the egocentric source-receiver distances in the former case are more underestimated than the latter one in the conditions around 2 m. The reason for this tendency may be considered that the transient sounds like the impulse response including many reflected sound generated at the each part of the sound field may make the location of the sound image slightly farther than that of the original object location. Secondly, in comparison between the results of the impulse response obtained in this paper (red line) or in the Ref. [10] (broken black line), the relative error of the former case indicates larger value than that of the latter one, especially in the conditions with relatively shorter source-receiver distances. It may be due to the reflected sounds included in the impulse responses adopted in this paper. The energy decay curve of the impulse responses adopted in this paper is shown in Fig. 3(b). The surrounded parts by broken lines in the two-meter and five-meter conditions indicate prominent reflected sounds which can also be perceived separately from the direct sounds. The reflection of these sounds are estimated to be generated at the left-side end of the corridor shown in Fig. 1(a). Then, the increase of the relative error may be caused because the subjects misperceive the location of the sound source interfered by the incoming sounds reflected at farther points than the real location of the sound source.

3.2. Effect of Simultaneously Presented Auditory or Visual Stimuli

The results of the experiment of the second and third parts are shown in Figs. 6 and 7, respectively. The results of the first part is additionally shown again in the figure. Firstly in Fig. 6(a), the results simultaneously presented with both auditory and visual stimuli show similar tendency to the conditions among the auditory-only and visual-only results of the source-receiver distance of 10 m and 20 m, whereas those of 5 m and 30 m show intermediate values between the auditory-only and visual-only results. The result of the condition of 2 m shows slightly closer value to the visual-only one. In the situation of the virtual image presentation (Fig. 6(b)), the results simultaneously presented with both auditory and visual stimuli also show similar trend as those of Fig. 6(a), from the abovementioned viewpoints, but that of the condition of 2 m also show intermediate values between the auditory-only and visual-only results. On the other hand in Fig. 7(b), the evaluated results simultaneously presented with both auditory and visual stimuli show closer values to those presented with only the visual stimuli in all conditions of the source-receiver distance. Figure 7(a) shows that the evaluated results simultaneously presented with both auditory and visual stimuli also show closer values to those presented.
with only the visual stimuli, whereas the results under the condition of “30 m” show intermediate value between the auditory-only and visual-only results. Then, the results of Figs. 6 and 7 have shown different dependency of the egocentric distance perception on the impulse response or the male voice. From these results, it can be considered that the extent of the subjective effects of the auditory and visual information on the overall distance evaluation may differ depending on the characteristics of the sounds, especially the time variation characteristics of them.

The results of the experiment of the third part are shown in Fig. 8. In this result, the mismatching audiovisual stimuli were presented. For example, in the condition (i) with “impulse response” in the thirty-meter source-receiver distance and the virtual video of the ten-meter source-receiver distance, the estimated value resulted in 15.5 m which indicates an intermediate value between the 22 m (with only the visual stimuli) and 9.5 m (with only the auditory stimuli). In other conditions, the same tendencies are observed. From these results, it may be considered that the estimated results of each source-receiver distance indicate different characteristics depending on the types of sound, and such a transient sound like the impulse response gives more drastic influence on the distance perception.

4. CONCLUSION

The estimation of the distance between the sound sources and the evaluator perceived by presentation of the auditory and/or visual stimuli were investigated through subjective experiment. It was found that the subjective impression of the distance to the sound source was affected by both the auditory and visual information, and the extent of the influence by the auditory and visual stimuli varies depending on the time variant characteristics of the sound. As our future work, such an evaluation experiment of distance estimation in an anechoic space without any reflection will be conducted, and the obtained results by the experiment should be compared to those of this paper effected by reverberation.

REFERENCES

[1] W. Fujisaki, S. Shimojo, M. Kashino and S. Nishida, “Recalibration of audiovisual simultaneity,” Nat. Neurosci., 7, 773–778 (2004).
[2] Z. Zhou, A. D. Cheok, Y. Qiu and X. Yang, “The role of 3-D sound in human reaction and performance in augmented reality environments,” IEEE Trans. Syst. Man Cybern. A, 37, 262–272 (2007).
[3] D. H. Mershon and L. D. King, “Intensity and reverberation as factors in the auditory perception of egocentric distance,” Percept. Psychophys., 18, 409–415 (1975).
[4] C. J. Lin, B. H. Woldegiorgis, D. Caesaron and L.-Y. Cheng, “Distance estimation with mixed real and virtual targets in stereoscopic displays,” Displays, 36, 41–48 (2015).
[5] S. Spagnol, E. Tavazzi and F. Avanzini, “Distance rendering and perception of nearby virtual sound sources with a near-field filter model,” Appl. Acoust., 115, 61–73 (2017).
[6] S. Spagnol, R. Hoffmann, Á. Kristjánsson and F. Avanzini, “Effects of stimulus order on auditory distance discrimination of virtual nearby sound sources,” J. Acoust. Soc. Am., 141, EL375 (2017).
[7] A. Rungta, N. Rewkowski, R. Klatzky, M. Lin and D. Manocha, “Effects of virtual acoustics on dynamic auditory distance perception,” J. Acoust. Soc. Am., 141, EL427 (2017).
[8] M. Paquier, N. Côté, F. Devillers and V. Koehl, “Interaction between auditory and visual perceptions on distance estimations in a virtual environment,” Appl. Acoust., 105, 186–199 (2016).

[9] H. Sato and J. S. Bradley, “Evaluation of acoustical conditions for speech communication in working elementary school classrooms,” J. Acoust. Soc. Am., 123, 2064–2077 (2008).

[10] P. W. Anderson and P. Zahorik, “Auditory/visual distance estimation: Accuracy and variability,” Front. Psychol., 5(1097), pp. 1–11 (2014).