Evaluating Effects of Listening to Content with Lip-sync Animation on Head Mounted Displays

Naoya Isoyama
Kobe University
1-1 Rokkodaicho, Kobe, Hyogo,
6578501 JPN
isoyama(at)eedept.kobe-u.ac.jp

Tsutomu Terada
Kobe University / PRESTO, JST
1-1 Rokkodaicho, Kobe, Hyogo,
6578501 JPN
tsutomu(at)eedept.kobe-u.ac.jp

Masahiko Tsukamoto
Kobe University
1-1 Rokkodaicho, Kobe, Hyogo,
6578501 JPN
tuka(at)kobe-u.ac.jp

UbiComp/ISWC'17 Adjunct , September 11–15, 2017, Maui, HI, USA
ACM 978-1-4503-5190-4/17/09.
https://doi.org/10.1145/3123024.3129265

Abstract
Users should always be able to receive information when using a head-mounted display (HMD) anytime, anywhere. Users can watch content shown on an HMD hands-free even when moving or working. It seems that presenting specific information affects humans. In this paper, we investigate the effects on listening to speech information that are caused by presenting animation on an HMD. It is difficult to listen to information that is presented in noisy surroundings. If the solution were only to turn up the volume, we would feel uncomfortable because this is very inconvenient. Therefore, by additionally presenting animation, we aim to make it easy for users to listen to speech information. With our method, we use lip-sync animation that matches specific speech information. We performed two experiments to determine whether it is easier to get speech information with animation.

Author Keywords
Head Mounted Display, Wearable Computing, Speech Information, Cognitive Psychology

ACM Classification Keywords
H.5.1 [Information interfaces and presentation (e.g., HCI)]: Multimedia Information Systems.

INTRODUCTION

A user equipped with a monocular head-mounted display (HMD) can always browse information hands-free in various situations such as when moving or doing other work in everyday life. The user can check e-mail or train transfer information even while walking without a mobile terminal. For these situations, the existing methods for presenting information in desktop computing environments are not always sufficient for users to fully grasp information, so several novel methods that take into consideration user needs and situations are proposed. HMDs have appeared one after another for daily use, such as Google Glass[1], M100[2], and Telepathy Walker[3]. These HMDs are loaded with Android OS and other OSs, and users can use them in the same way as smartphones.

Currently, when we look at a display in a desktop environment, we face the display with the intention of working. Smartphones, which are widely common now, have a larger display than conventional mobile phones. Although information can be browsed casually, this is active information browsing. In comparison, HMD users are always presented with some information. This is a passive browsing. Our physical and mental behavior is affected by what we see. Unlike real world scenarios in which what we see changes naturally, the content presented on an HMD can be intentionally changed by others. It is important to investigate the effect of this because it is believed that the influence on our behavior increases if we always look at the display.

In this paper, we focus on auditory perception and investigate its influence. Auditory perception does not depend on only the volume of the sound (that a person wants to hear) and other surrounding sounds. It is also affected by senses other than the auditory sense. There is much research on the relationship between the auditory sense and other senses. As research on auditory and visual perception, the McGurk effect [1] is a compelling demonstration of how we all use visual speech information. In this research, a person who is in a video is mouthing the syllables /ga-ga/, but the video has been dubbed with a sound track of him saying /ba-ba/. Trying to reconcile the conflicting information from our eyes and ears, the brain will decide that the syllables are those that are acoustically closest to /ba-ba/, which is articulated with the lips open, and we will “hear” /da-da/ or /tha-tha/. The ventriloquist effect[2] refers to perceiving speech sounds as coming from a different direction than their true direction. We regularly experience the effect when watching television and movies, where the voices seem to emanate from the actors’ lips rather than from the actual sound source. Moreover, various studies have revealed that not only auditory information derived from voice but also the visual information of the movement of a mouths useful for listening when talking face to face[3, 4]. On the basis of these findings, we assume that using the mouth as a visual stimulus makes it easier to acquire related speech information. In this paper, we investigate the influence of listening to a voice while watching a lip-sync animation on an HMD in which a displayed mouth moves synchronously with a specific sound and seems to be talking. If it becomes easier to listen to specific speech information with this animation, we can easily listen to information even in noisy environments.

We examined whether a person can easily listen to a certain voice by watching a lip-sync animation that is synchronized with the voice in a noisy environment. In experiments, we investigated whether the animation makes it easier to understand what a certain person said and whether there

[1] https://www.google.com/glass/start/
[2] https://www.vuzix.com/Products/m100-smart-glasses
[3] http://www.telepathywalker.com

is a change in the subjectively perceived volume of a specific voice. We evaluated the number of correct answers and the change in volume under the presence or absence of lip-sync animation.

The remainder of this paper is organized as follows. In Section 2, we explain related work. We then describe the experiment in Section 3. Finally, we present our conclusion and future work in Section 4.

RELATED WORK

There are many studies that examined that perceptions made with the auditory sense are changed by other senses (such as vision), knowledge, and preconception [1, 2, 3, 4]. The cocktail party effect is the phenomenon of being able to focus one's auditory attention on a particular stimulus, much the same way that a person can focus on a single conversation in a noisy room[5]. As a system that utilizes the relationship between visual and auditory senses, SmartVoice [6] makes it possible to show a speaker speaking by him/herself directly by outputting voice data synchronized to the movement of the speaker's mouth.

There are also many attempts to convey a voice properly or change the subjective impression of a voice. Yataka investigated the relationship between the recognition of audio information and a user's activity and ambient sound [7]. On the basis of an evaluation, he designed and implemented a system that changes the way audio information is presented on the basis of the user's activity, the volume of ambient sound, and the recognition accuracy that the user desires. Okazaki investigated the possibility of tactile-audio crossmodal interaction in frequency perception [8].

EXPERIMENT

As far as we know, there is no study that examined the influence of presenting lip-sync animation on an HMD. Therefore, we investigated whether the animation makes it easy to understand what somebody said and whether there is a change in the subjectively perceived volume of a specific voice. We used a PC as well as an HMD for the experiments and examined whether there is a difference between an HMD and PC.

Experiment on understanding content

We investigated whether lip-sync animation helps a user understand what somebody said in a noisy environment. We hypothesize that the animation makes understanding information easy.

Experimental method

We created five audio tracks that simultaneously reproduced three pieces of audio data (weather forecast, presentation, and English learning). The tracks (a, b, c, d, e) each had different content. Each track was approximately one and a half minutes long and normalized to eliminate the difference in volume. All speech was spoken by a female in Japanese. Participants listened to the tracks and tried to understand only the content of the weather forecast. We created five lip-sync animations that synchronized with the voices of the weather forecast speech. We utilized CubismAnimator by Live2D Inc.[4] to create the animations. As the animated character, we used epsilon data, which is a standard model. The experiment system showed only the mouth part of the character. Figure 1 shows the experiment. The participants listened to the audio track from a speaker connected to the PC in three conditions: without the animation and with the animation on the PC and on the HMD. After listening, they answered six questions with three

[4] http://www.live2d.com/

[I] No animation [II] PC [III] HMD

Figure 1: Experiment in action

choices regarding the content of the weather forecast. They were only allowed to answer the questions that they were able to listen to (without their intuition). If they were not confident enough to answer, they answered the choice written "I could not understand." We used an M100 by the Vuzix Corporation as the HMD. When using the HMD, the participants placed it over their dominant eye and looked at a blank wall.

Before the experiment, we told the participants "We will have you listen to three kinds of audio speech simultaneously. After listening, we will ask six questions regarding the content of a weather forecast. Thus, please listen to the weather forecast carefully." First, we explained how the experiment would work to them by using audio track e. We reproduced the track and showed the appearance of the animation on the PC and HMD. After the explanation, the participants listened to track e and answered the six questions as an exercise. Next, they listened other four tracks for all conditions. They answered the questions whenever they listened to one track. On another day, we conducted the experiment in a different order. In short, they listened to

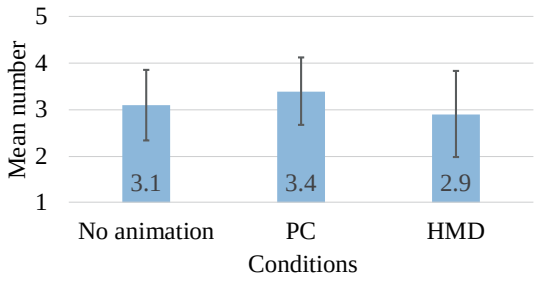

Figure 2: Result of experiment on understanding content

four audio tracks three conditions over separate three days. We decided the order on the basis of a Latin square design, and each participant underwent the experiment in respective order. There were 16 participants (male: 13, female: 3) in their 20's.

Results

Figure 2 shows the mean number of participants answering correctly for each condition (error bars are S.D.). The

values were 3.1 (no animation), 3.4 (PC), and 3.0 (HMD). The S.D. values were 0.76 (no animation), 0.73 (PC), and 0.93 (HMD). A comparison between "no animation" and "PC" showed that 11 participants had better results for "PC browsing" than "no animation" and 5 participants had worse results for "PC browsing." A comparison between "no animation" and "HMD" showed that 5 participants had better results for "HMD browsing" than "no animation", and 10 participants had worse results for "HMD browsing." A comparison result "PC" and "HMD" showed that 12 participants had better results for "PC browsing" than "no animation" and 3 participants had worse results for "PC browsing." Four participants had better results for both "PC browsing" and "HMD browsing" than "no animation." Four participants had worse result for both "PC browsing" and "HMD browsing" than "no animation." We conducted an ANOVA (within-subject factor) by using the number of correct answers. There was a significant difference ($F_{(2,30)} = 3.58$, $p < .05$). Then, we conducted multiple comparisons by using the Bonferroni method. There was a significant difference; the result for "PC" was better than that for "HMD" ($MSe = 0.220$, $p < .05$).

Experiment on adjusting volume

We investigated whether a lip-sync animation changed the subjectively perceived volume. We hypothesized that the animation made the volume subjectively louder.

Experimental method

We used the same audio tracks (a–e) without the English learning data and the same lip-sync animation as the experiment in Section 3.1. The participants listened to the audio tracks from the speaker connected to the PC under three conditions: without the animation and with the animation on the PC or the HMD. We used M100 as the HMD the same as in Section 3.1. The participants were allowed to adjust only the volume of the weather forecast data while listening. The volume was increased one point per right click and decreased one point per left click by using a wireless mouse. One point is a value obtained by dividing the volume at the start of the experiment into 50. For the evaluation, the participants adjusted the volume and made it the same to that of another piece of audio data (presentation audio data). The participants informed the experimenter when they finished adjusting the volume to the value that they wanted. We performed evaluation on the basis of the difference in the volume of the weather forecast data at the end of the experiment. The volumes of the two pieces of audio data were made the same by normalizing them at the start of the experiment. We conducted the experiment without telling this to the participants. Each participant listened to four audio tracks (a–d) in one day. On the first day, they listened to track e as practice. The order of the tracks and browsing states was set on the basis of a Latin square design, and each participant conducted the experiment in respective order. In short, the participants listened to four audio tracks under three conditions over separate three days. There were 16 participants, the same from Section 3.1.

Results

Figure 3 shows the mean value at which the participants adjusted volume for each condition (error bars are S.D.). First, for the evaluation, we standardized (mean: 0, S.D.: 1) the points (that were at the end of each experiment) for each participant. The numerical characters in the graph are the mean values of the standardized values. The values were defined to be zero at the start of each experiment and change by $\pm$1 point each time the mouse was clicked once. The values increased because the participants perceived the volume of the weather forecast data to be low. We assumed that the value would be smaller when listening with the animation.

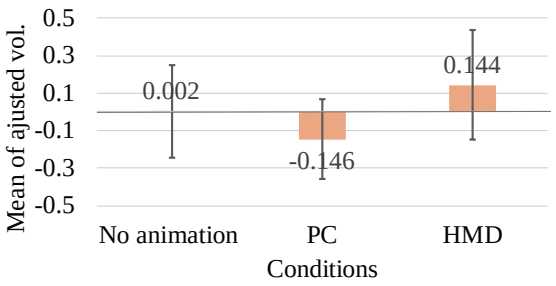

Figure 3: Result of volume adjusting experiment

A comparison between “no animation” and “PC” showed that nine participants had smaller values for “PC browsing” than “no animation” and six participants had larger values for “PC browsing.” A comparison between “no animation” and “HMD” showed that 5 participants had smaller values for “HMD browsing” than “no animation” and 11 participants had larger values for “HMD browsing.” A comparison between “PC” and “HMD” showed that 11 participants had smaller values for “PC browsing” than “no animation” and 5 participants had larger values for “PC browsing.” Four participants had smaller values for both “PC browsing” and “HMD browsing” than “no animation.” Five participants had larger values for both “PC browsing” and “HMD browsing” than “no animation.” We conducted an ANOVA (within-subject factor) by using the values of adjusted volume. There was a significant tendency ($F_{(2,30)} = 3.27, p < .10$). Then, we conducted multiple comparisons by using the Bonferroni method. There was a significant difference in that the value for “PC” was smaller than that for “HMD” ($MSe = 0.103, p < .05$).

Consideration

Regarding the experiment on understanding content, the mean number for which the participants answered correctly for PC browsing was the highest, and the difference between the HMD browsing and the no animation condition was small. The result for HMD browsing was the worst. In advance, we conducted a pre-experiment. There were differences from this paper’s experiment. The participants listened to the audio tracks in order from data a and gazed at the HMD. At that time, the number of people was 22 (male: 18, female: 4) in their 20’s, and they were different people from those of this paper’s experiment. The mean numbers for which they answered correctly were 2.5 (no animation), 2.8 (PC), and 2.7 (HMD). The S.D. values were 1.06 (no animation), 1.05 (PC), and 0.94 (HMD). There was no significant difference ($F_{(2,42)} = 1.65, p > .05$). Unlike the results in this paper, the result for “no animation” was the worst, and there was little difference between “PC browsing” and “HMD browsing.” There is the possibility that there is no value or that there is a negative influence on listening if a user does not gaze at the lip sync animation, although it is useful to gaze at the animation.

Regarding the experiment on adjusting volume, the result for HMD browsing was the biggest. This shows that the users listened to the audio information at a smaller volume if there was an animation in a corner of the field of vision. This is not useful for our purpose.

We need to increase the number of participants and proceed with each experiment.

As a feature when using the HMD, there was the opinion that “the sound is heard as if it is flowing from the HMD” although it was output from the speaker. From this opinion, even if the speaker from which the voice is heard is behind the user, there is the possibility that the perceptual direction

of the sound source could be changed and that this would make it easier to listen to the sound. It is also possible to change the perceptual distance to the speaker by changing the size of the mouth displayed on the HMD. To take these possibilities into consideration, we also need to perform experiments by changing the size of the mouth or making the speaker not face the user.

The influence of the animation also could be negative. For example, it is conceivable that the attention of the user could be led to a specific advertisement voice even though he/she does not want this to happen. Due to long-term use, there is the possibility that the ability of the user to localize a sound source would be hindered. While considering these negative aspects, it is important to investigate the influence of images on an HMD.

CONCLUSION

We considered the importance of investigating the influence of animation presented on an HMD. In this paper, we investigated the influence of listening to speech information by presenting a lip-sync animation. We aimed to make it easier to listen to specific speech information by showing the animation on an HMD and examined the effectiveness of doing so in two experiments. In the first experiment, we investigated whether it is easier to understand the content of specific speech with the lip-sync animation. Although the participants could easily understand the speech with the animation on a PC, there was no difference between the conditions of no animation and HMD browsing. In the second experiment, we examined whether the subjective volume changed for specific speech. Although the participants perceived the audio volume to be louder with the animation on the PC, they perceived it to be lower with the animation on the HMD. We will continue to perform further evaluation experiments in the future.

Acknowledgements

This research was supported in part by a Grant in aid for Precursory Research for Embryonic Science and Technology (PRESTO) and CREST from the Japan Science and Technology Agency.

REFERENCES

1. M. Harry and M. John: Hearing Lips and Seeing Voices, *Nature*, Vol. 264, Issue. 5588, pp. 746–748 (1976).
2. R. B. Welch and D. H. Warren: Immediate Perceptual Response to Intersensory Discrepancy, *Psychological Bull*, Vol. 88, No. 3, pp. 638–667 (1980).
3. R. Campbell and B. Dodd: Hearing by Eye II, *Psychology Press* (1998).
4. D. W. Massaro: Speech Perception by Ear and Eye, *A Paradigm for Psychological Inquiry* (1987).
5. E. C. Cherry and W. K. Taylor: Some Further Experiments upon the Recognition of Speech, with One and with Two Ears, *Journal of Acoustical Society of America*, Vol. 26, pp. 554–559 (1954).
6. X. Li and J. Rekimoto: SmartVoice: A Presentation Support System for Overcoming the Language Barriers, *Proc. of CHI2014*, pp. 1563–1570 (2014).
7. S. Yataka, et al. : A Context-aware Audio Presentation Method in Wearable Computing, *Proc. of SAC2011*, pp. 405–412 (2011).
8. R. Okazaki, et al. : Judged Consonance of Tactile and Auditory Frequencies, *Proc. of the IEEE World Haptics Conference*, pp. 663–666 (2013).