CalmResponses: Displaying Collective Audience Reactions in Remote Communication

Kiyosu Maeda*
University of Tokyo
Tokyo, Japan
kiyosu775@g.ecc.u-tokyo.ac.jp

Riku Arakawa
Carnegie Mellon University
Pittsburgh, USA
rarakawa@cs.cmu.edu

Jun Rekimoto
University of Tokyo
Tokyo, Japan Sony CSL Kyoto
Kyoto, Japan
rekimoto@acm.org

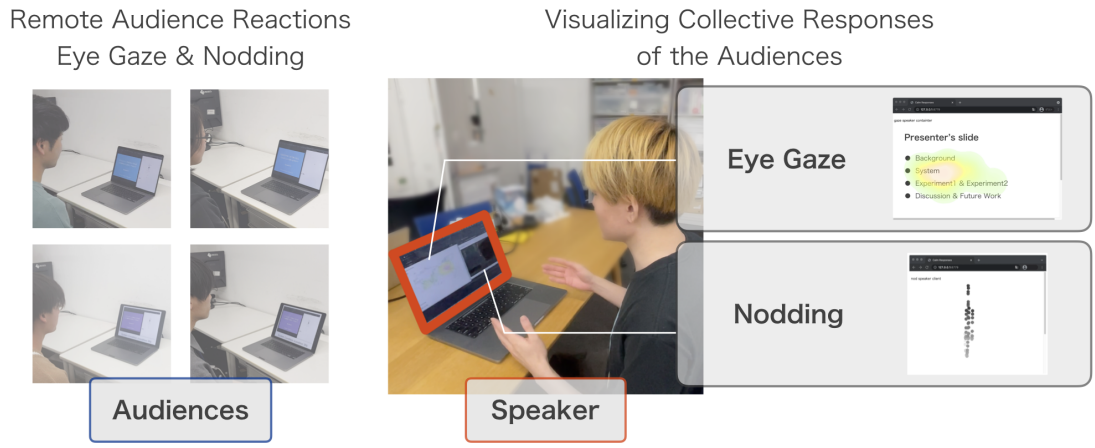

Figure 1: The overview of the proposed system. The system obtains eye gaze and nod reactions of audiences with webcams (left) and collectively presents them to speakers in real time (right) during remote communication.

ABSTRACT

We propose a system displaying audience eye gaze and nod reactions for enhancing synchronous remote communication. Recently, we have had increasing opportunities to speak to others remotely. In contrast to offline situations, however, speakers often have difficulty observing audience reactions at once in remote communication, which makes them feel more anxious and less confident in their speeches. Recent studies have proposed methods of presenting various audience reactions to speakers. Since these methods require additional devices to measure audience reactions, they are not appropriate for practical situations. Moreover, these methods do not present overall audience reactions. In contrast, we design and develop CalmResponses, a browser-based system which measures audience eye gaze and nod reactions only with a built-in webcam and collectively presents them to speakers. The results of our two user studies indicated that the number of fillers in speaker's speech decreases when audiences' eye gaze is presented, and their self-rating score increases when audiences' nodding is presented. Moreover, comments from audiences suggested benefits of CalmResponses for them in terms of co-presence and privacy concerns.

IMX '22, June 22–24, 2022, Aveiro, JB, Portugal

ACM ISBN 978-1-4503-9212-9/22/06…$15.00
https://doi.org/10.1145/3505284.3529959

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**; **Interactive systems and tools**.

KEYWORDS

remote communication, audience sensing, eye gaze, nodding, feedback design

ACM Reference Format:
Kiyosu Maeda, Riku Arakawa, and Jun Rekimoto. 2022. CalmResponses: Displaying Collective Audience Reactions in Remote Communication. In *ACM International Conference on Interactive Media Experiences (IMX '22), June 22–24, 2022, Aveiro, JB, Portugal.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3505284.3529959

1 INTRODUCTION

We frequently speak to others in distant locations using online communication tools. Some people give lectures to students remotely, and others present to groups of people in online meetings. One of the advantages of such online communication is that we are not constrained to physical space. We can access this communication style from anywhere as long as we connect to the Internet with laptops or smartphones. Due to the COVID-19 pandemic, there has been a soaring demand for online communication [65], especially

in education [5, 49] and work [70] situation. As a result, many institutions have undergone a rapid transition to remote environments.

Despite its benefits and demand, audio- and video-based online communication has a crucial problem: it is difficult for speakers to see others' reactions compared to face-to-face communication. Existing teleconferencing tools, such as Skype, cannot effectively convey non-verbal cues [44]. This disadvantage is particularly remarkable when there are multiple audiences. For example, most students do not want to turn on their videos due to various reasons such as privacy concerns (e.g. appearance, background) as well as problems in their internet connection [10]. Even if audiences turn on these features, speakers cannot observe the reactions of the entire audience at a glance unlike offline situations due to the limited screen size or screen sharing [39]. When speakers cannot receive these non-verbal signals from audiences, they feel anxious and stressed during their speech [6, 46].

In response, several techniques have been proposed in the HCI literature to support speakers during online communication by transmitting audience reactions to them. One way to present audience reactions to speakers is for audience members to send texts or emoticons [52, 64]. Although speakers can observe real-time reactions through this method, it forces audience members to manipulate laptops explicitly, making them feel bothersome. Hence, speakers' ability to obtain enough information depends on the active participation of the audiences. Moreover, texts or emoticons cannot necessarily reflect the real thoughts of their senders [27, 43]. Another way to communicate reactions is for audience members to share unconscious behaviors represented by non-verbal cues or physiological signals of audiences with speakers. This method can obtain audience reactions without burden. Existing research has exploited these signals as audience reactions [20, 40, 50, 62].

However, most of these systems require extra sensors in addition to laptops to obtain these reactions. Thus, these methods can only be used in offline controlled laboratory settings [22, 28, 61, 67]. Also, some systems acquire reactions obtrusively (e.g. attaching a sensor), which can negatively affect the communication experience for the audience. To calmly obtain entire audience reactions online, we require scalable systems that only use commonly available devices, such as laptops, and do not depend on any additional device. Furthermore, it is desirable for speakers to see overall audience reactions simultaneously as they can in offline situations. In other words, we need to design feedback that aggregates their reactions so that speakers can experience the communication atmosphere on a single screen.

In this study, we propose CalmResponses, a browser-based system that obtains the eye gaze and nod reactions of remote audiences using a webcam and collectively presents them to speakers. Through a long time of research [9, 19], it is known that eye gaze and nodding are some of the most meaningful non-verbal cues in face-to-face communication. The same is true for online communication, and recent studies have shown that presenting eye gaze or nod reactions in real-time during conversation positively affects speakers [13, 14, 36]. CalmResponses presents these signals collectively, which enables presenters to see overall audience reactions simultaneously. In addition, since the proposed system does not require an additional device other than a webcam equipped with a laptop, it is easy to use in actual online settings.

We conducted two experiments and studied how collective audience reactions affect speakers through the objective and subjective evaluations of presentations, number of fillers, and qualitative comments about the presentation experiences. As a result, we found that CalmResponses positively affected not only the speakers but also the audiences. Specifically, the collective feedback of eye gaze reduced the number of fillers in speakers' speech and the collective feedback of nodding enhanced their self-rating scores. Moreover, the audiences found the experience positively, mentioning that it contributed to their feeling of co-presence as well as their low privacy invasiveness.

To summarize, our contributions are as follow:

- We designed and developed a browser-based system that collectively presents real-time eye gaze and nod reactions of entire audience in remote communication settings to speakers using the built-in webcam.
- We conducted online experiments to study how the system affected the speakers' presentations and the result indicated that displaying eye gaze reactions reduced speakers' number of fillers, and displaying nodding increased self-rating scores.
- Based on the findings of the experiments, we discussed the potential benefits of collective visualization in terms of co-presence and privacy concerns for future audience sensing and feedback technologies.

2 RELATED WORK

2.1 Audience Sensing and Feedback

As mentioned in the introduction, one often-used method of presenting real-time audience reactions to speakers in online communication is to use texts or emoticons [42, 45, 52, 64, 78]. For example, Teevan et al. [64] proposed a system that presents audience reactions, such as "like" or "dislike", in a slide during a presentation. Zhou et al. [78] investigated the relationship in live streams between paid gifting [1] and stimuli extracted from *danmaku*, moving comments on videos like bullets, and found that some variables such as number of comments positively affected paid gifting. Although these methods enable speakers to observe real-time audience reactions, they also force audiences to manipulate some devices explicitly such as pushing buttons. In other words, the success of these methods depends on active participation of the audience members.

In contrast, other studies have leveraged unconscious behaviors and states of audiences, such as non-verbal cues or physiological signals, and presented them to speakers [20, 28, 61, 63, 67, 75]. For example, EngageMeter [28] allows presenters to observe audiences' engagement and workload through electroencephalography (EEG) signal. Yao et al. [75] developed a system that presents students' gaze movements to a teacher in online exercise lectures. They showed that this system can help the teacher to grasp students' attention and the progress of their tasks. Despite their efficacy, however, most of these methods require additional devices for sensing the reactions of the audience, such as Tobii [2] to estimate eye gaze positions. Moreover, these systems often use obtrusive data collection methods, negatively affecting the audience experiences.

[1] paid gifting is a (virtual) gift or a donation that viewers can send to streamers.
[2] https://www.tobii.com

It is important to use simple interaction technologies that do not interfere with audiences experiences [4].

To obtain real-time reactions of multiple audiences in online situations unobtrusively, we require a scalable system that only uses commonly available devices, such as laptops. In this sense, AffectiveSpotlight [50] is a suitable system to share audience reactions with speakers because it estimates facial responses and head gestures of audiences with a webcam. However, it presents only one audience with the highest engagement simultaneously and it is impossible for speakers to grasp how the others feel because it is hard to deny that while one audience is engaging, the others are not during presentations. Given the assumed scenarios such as education as we mentioned in the introduction, it would be better for speakers to observe the collective audience reactions simultaneously. Thus, in this paper, we develop a system that can unobtrusively sense audience reactions and visually present them to speakers in a collected manner.

2.2 Non-verbal Behavior Analysis in Communication

Eye gaze and nodding has been subjects of research in human communication for a long time [19, 48, 69]. On one hand, eye gaze has been utilized as a cue to estimate various internal states since it has been shown to be correlated with human cognitive processes closely [9]. For example, we can infer one's interest or attention [1, 12, 32], one's linguistic abilities [21], the type of books one is reading [38], and even what picture one is recalling [71] based on eye gaze movements. On the other hand, nodding is also one of the main methods of conveying information in communication, as we can see from a finding that over 80% of listeners' head movements are nodding [74]. In face-to-face communication, nodding has various meanings and functions [26, 56]. For example, listeners' nods indicate confirmation, agreement, and even disagreement [56].

These days, researchers study the use of these modalities in online communication. For eye gaze, prior research has explored the effect of sharing eye gaze movements with remote users in online communication [16, 17, 59, 63, 68, 75]. In these cases, users share their gaze positions on their screens, which helps to disambiguate what the speaker is referring to [16]. GAZE groupware system [68] is a system that conveys gaze directions in multiparty mediated communication. In this system, users can easily grasp who is talking to whom and about what. Other research [59, 75] visualized eye gaze positions from multiple students in remote classes or exercises. Similarly, nodding has also been leveraged in computer-mediated communication [14, 25, 36, 47, 72]. For example, Kubota et al. [36] proposed a system that presents nodding images related to spoken words in online communication. Chollet et al. also developed a system leveraging virtual audiences for presentation training [14]. This virtual audience responds with non-verbal cues, including nodding during presentations. Likewise, Maloney et al. [47] investigated non-verbal communication in social virtual reality and found that nodding is a naturally used interaction.

While previous research has proposed systems collectively visualizing engagement of multiple audiences through facial expressions [62] and heart rate [61], little is known about how we design the collective visualization from multiple audiences' eye gaze and nodding reactions. In the present study, we propose a system that presents audience eye gaze and nod reactions collectively to speakers. The system estimates and displays the users' eye gaze and head movements (specifically nod reactions). In the next section, we introduce how we estimate and display these reactions without relying on external devices other than conventional laptops.

3 PROPOSED SYSTEM: CALMRESPONSES

3.1 System Overview

We propose CalmResponses, which displays audience eye gaze positions and head movements collectively to speakers for supporting them to grasp audience reactions in real time. Based on the existing research and issues as we discussed in Section 2, we need to develop the system appropriate for real-time, remote, and multiple-audience situations. Specifically, we aim to meet the following two requirements.

Do not need any additional device to unobtrusively estimate reactions. Although we can obtain richer information with additional devices [28], it increases the cost to utilize in actual online settings. To estimate and display entire audience reactions while keeping availability, we implemented the system which functions only with a built-in webcam and a web browser.

Display eye gaze positions and head movements of multiple audiences collectively. To present reactions of multiple audiences, we need to aggregate reactions on a single screen. The aggregation of reactions from multiple users has also been proposed in [62] leveraging facial expressions. We designed and developed collective feedback method for eye gaze and nodding.

CalmResponses satisfies the above requirements, implemented as a browser-based application using HTML, CSS, and JavaScript. The system consists of three sub-systems: audience client, speaker client, and server. The audience client obtains and sends audiences' eye gaze or head movements through a built-in webcam, and the speaker client receives and displays feedback in real time. We adopted WebSocket to connect the server with both the speaker and audience clients. We deployed the system on Heroku. We provide the source code [3] and the link [4] to try the system. Users can take this application together with other video conferencing tools such as Zoom as long as they open the browser interface. In the next section, we elaborate on the audience client and speaker client respectively, centering on how they estimate and display eye gaze positions or head movements of audiences.

3.2 Collective Eye Gaze Reactions

Sensing

To estimate one's eye gaze position on the screen, we used WebGazer.js [53], a javascript library to infer eye gaze positions based on webcams equipped with laptops in real time. It has shown as a reliable tool for eye tracking due to its low errors. Since this library requires only a built-in webcam and the browser, it is suitable for our purpose, as discussed in Section 3.1.

[3] https://github.com/kiyosumaeda/CalmResponses
[4] https://calmresponses.herokuapp.com/

Users need to conduct calibration before using the system. We introduced the calibration process at the beginning of the usage. We followed the original research [53] to implement the calibration process in the audience client.

Since raw gaze data are not stable even in fixation due to the involuntary movements of eyes [76], we need to smooth the data. While there are various methods to smooth them [37], we concisely smooth them by calculating the mean values of gaze positions for several frames. Thus, the audience client sends relative eye gaze positions smoothed on the screen to the speaker client through a server. Next, we show how we display gaze positions in the speaker client.

Feedback Design

In this section, we describe how we present eye gaze movements of entire audience collectively. Figure 2 shows the three ways to visualize eye gaze movements. Specifically, we considered three ways for presenting gaze collectively: Dots, Heat Map, and Heat Map only for the dense area.

Dots. Dots are the most common approach to visualize gaze data [18]. As described in Figure 2A, red dots on eye gaze positions of each user were presented. In detail, upon receiving data of relative gaze position from the audience client, the speaker client calculates an absolute gaze position. Although this approach is a primitive way to visualize gaze positions, one of the problems is that this dot visualization can be distracting for speakers [17].

Heat Map. Figure 2B shows the heat map visualization of eye gaze data. Previous work has adopted this heat map visualization [18, 51]. The color indicates the density of the audience gaze (e.g. red areas are where the audience is mostly looking at). The heat map is translucent so that users can see original contents under the visualization without being distracted. This visualization can help speakers to understand audience interests and attitudes intuitively. For example, from red areas, speakers can see which part of the content the audiences are interested in. Likewise, when the red area is exactly where the speakers are referring, they can see that audiences pay attention to their speeches. Conversely, when there is no dense area, they can infer that audiences are likely to be distracted. In this way, heat map visualization helps speakers understand the audience reactions based on their gaze movements. Heat map is also robust to noise that derives from accuracy because heat map represent ranges (not the exact position) [18], which is suitable for actual remote situations.

Heat Map Only for Dense Area. Another way to display gaze information is to draw only dense areas of heat map as in Figure 2C. This corresponds to Shared Area [18] where multiple users look at simultaneously. This visualization allows users to focus more on where many audiences are looking. On the other hand, this visualization is hard to reflect the minority's interests. Moreover, when audience eye gaze positions are distributed within the screen, it does not represent overall audience reactions.

Considering the advantages of Heat Map as real-time feedback for speakers, we adopted Heat Map visualization for CalmResponses.

3.3 Collective Nod Reactions

Sensing

We estimate head movements based on facial landmarks. We use clmtrackr[5], a javascript library to detect facial landmarks in videos or images. Since this library requires only a built-in webcam and a browser, it is suitable for online settings as we discussed in Section 3.1.

While there are various methods to estimate head movements [7], we estimate them only from one landmark corresponding to the nose tip position. The rationale of this method is as follows: on one hand, when we nod, we rotate our heads around the interaural axis [69]. In this condition, the nose tip position also moves up and down. On the other hand, when we shake our heads, we move our heads laterally. In this condition, the nose tip position also moves left and right. Therefore, we can estimate most head movements only from the nose tip position. Although this method cannot distinguish still head from other head movements, such as head rotation around the naso-occipital axis which expresses disbelief, we did not consider those because we mainly focused on nodding as we discussed in Section 2.2.

Feedback Design

We use trails of cursors to visualize head movement. Compared to the gaze visualization, there is little research regarding how to visualize head movements. We adopted cursors to visualize them more intuitively than other methods such as numbers (e.g. how many people are nodding). Figure 3 shows an example of how the system estimates and visualizes a user's head movement. The trails of cursors move based on the nose tip position's velocity vector. For example, when the user shakes their head down (Figure 3B), the nose tip position's velocity vector is downward direction. In this case, the trails of cursors extend from center to bottom (Figure 3E).

Next, we describe the way we present the head movements of multiple users. We superimpose all audiences' trails with some horizontal offsets. Figure 4 shows an example visualization of trails from three audiences. Since we overlay the multiple users' trails together, similar head movements are emphasized by overlapping many cursors. For example, when speakers observe the system and notice that cursors are extending vertically, they can infer that many audiences are nodding. This is the benefit of collectively displaying audience reactions as we can observe clearer vertical feedback in Figure 4 than in Figure 3. While speakers can see the gaze feedback that is visualized on their slide directly (See Figure 2), the nod feedback was presented alongside their slide (See Figure 8).

We calculate the velocity of the nose tip landmark about every 30 ms and send that velocity to the speaker client through the server. When the speaker client receives new data, it updates the position of each cursor. We can change the amplitude of horizontal and vertical movements so that the system can emphasize nodding.

[5] https://www.auduno.com/clmtrackr/

Figure 2: Various candidate ways to display audience eye gaze movements: Dots (A), Heat Map (B), and Heat Map for only dense area (C)

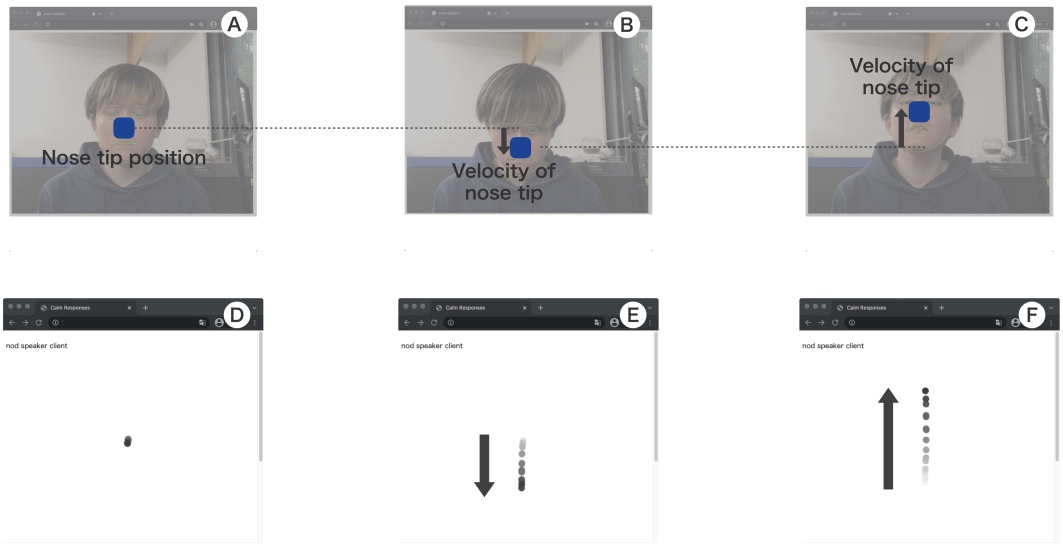

Figure 3: Illustration of how the system estimates and visualizes a user's head movement. When the user faces the front (A), the cursors' trails are still in the browser's center (D). Next, when the user shakes their head down (B), the nose tip position's velocity vector is downward direction. Then, the trails of cursors extend from center to bottom (E). Finally, when the user shakes their head up (C), the nose tip position's velocity vector is in the upward direction. Thus, the trails extend from bottom to top (F).

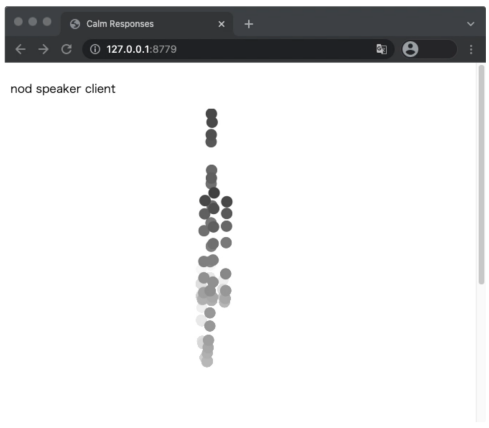

Figure 4: Collective visualization of head movements. We superimpose all audience's trails of cursors with some offsets.

4 EXPERIMENT OVERVIEW

Up to this point, we introduced our proposed system estimating and displaying collective reactions of multiple audiences. We conducted two experiments to evaluate the efficacy of the system displaying collective eye gaze and nodding reactions in online communication.

4.1 Design

We conducted two experiments that replicated online communication situations. Specifically, we asked participants to have presentations to audiences about given topics for a few minutes on Zoom, a common video conferencing platform. We employed a within-participants design comparing a treatment condition using CalmResponses with a baseline condition. Namely, there are three conditions: *condition B*, *condition CR-E*, and *condition CR-N*.

- *Condition B* : Baseline condition. In this condition, the audience turned off their video and audio. In other words, speakers could not see any reaction from the audience. This condition is based on the finding from existing research that most students tend to turn off their video and audio during online classes [10].
- *Condition CR-E*: System (CalmResponses) condition with eye gaze reactions. Although speakers could not see any video and audio from the audience, same as in *condition B*, they could see the collective eye gaze positions of the audience through the speaker client of the system (see Section 3.1).
- *Condition CR-N* : System (CalmResponses) condition with nod reactions. In this condition, although speakers could not see any video and audio from the audience, same as in *condition B*, they could see the collective head movements of the audience through the speaker client of the system.

In the first experiment, we compared *condition B* and *condition CR-E*; and in the second experiment, we compared *condition B* and *condition CR-N*. In each experiment, we performed three evaluations of the system: objective and subjective evaluation of the presentations, number of fillers, and qualitative comments. These are widely used measures in existing works to evaluate speech quality and indirectly estimate speakers' states such as anxiety [50, 66].

4.2 Measure

Objective and Subjective Evaluation of the Presentations. We used a questionnaire to evaluate the speaker's speech quality based on RoboCOP [66] after presentation. Speakers self-evaluated their speech quality, and the audience evaluated the speakers' speech quality. The questionnaire contained six items, and each item was rated from one to seven (1 = Not At All, 7 = Very Much). We calculated the sum of these scores for every speaker (note that we added reversed scores for the third question). The questionnaire for the speakers contained the following questions:

- How engaging was your presentation?
- How understandable was your presentation?
- How nervous were you during your presentation?
- How exciting was your presentation?
- How entertaining was your presentation?
- How competent were you during your presentation?

Note that the above questionnaire was used for the speakers. When asking the audience, we modified each item to be suitable for them. For example, "How engaging was your presentation?" was changed into "How engaging was the presenter's presentation?" We hypothesized that the system would help participants increase their evaluation scores.

Number of Fillers. In addition to the questionnaire evaluation, we evaluated the presentations using automatic speech analysis. We examined the number of verbal fillers used from the recorded presentation data to estimate speech performance. Since it is known that fillers are negatively correlated with speakers' performance or anxiety [11, 14, 23, 58], we anticipated that the number of fillers per minute would decrease in *condition CR-E* and *condition CR-N* compared to *condition B*.

Qualitative Comments. While the above two measures were effective, we also evaluated the system qualitatively. After the experiments, we asked the speakers (1) how and why they changed their self-rating scores compared to *condition B*, (2) how they perceived and interpreted the presented feedback, and (3) what the pros and cons of the system were. We also asked the audiences about their overall experience of the experiments. Specifically, we asked them how they perceived the system and how they reacted to the speakers.

4.3 Participants

We recruited 38 participants (nine females, 29 males, ages 19-44, mean 24.7). All participants were Japanese or Chinese, and all could fluently speak and understand Japanese. We randomly split them into six groups of five to seven participants as described in Table 1. They did not know each other at the time of the experiments. In each group, we randomly assigned one to three participants to be speakers, while the other members (four to five) served as an audience. The experiment was conducted online using Zoom, an online conferencing tool. In total, five speakers gave presentations in Experiment 1 (comparing *condition B* and *condition CR-E*), and six speakers in Experiment 2 (comparing *condition B* and *condition CR-N*). We use SE and SN to denote speaker participants in experiment 1 and experiment 2 respectively, while we use AE and AN to denote audience participants in experiment 1 and experiment 2 respectively.

4.4 Procedure

The overview of the two experiments is presented in Figure 5. The experiments contained four types of sessions: a pre-speech session, speech sessions (the first and the second), evaluation sessions (the first and the second), and a post-speech session. In the pre-speech session, all participants answered a questionnaire about their demographics. After finishing the questionnaire, the audiences accessed CalmResponses using Google Chrome, and an experimenter briefly explained how to use the system. Audiences were asked to put a webcam and a display in front of them.

In the speech-session, the speakers were asked to give presentations in Japanese twice to the audiences on given topics. They talked about different topics for the two presentations, and the topics were informed in advance. Although they could not read

Table 1: The groups of participants in the experiments. There are six groups of five to seven participants. We use SE1-SE5 / AE1-AE19 to denote speakers / audiences in experiment 1 and SN1-SN6 / AN1-AN8 to denote speakers / audiences in experiment 2.

group	conditions	number of speakers	number of audiences
group1	B & CR-E (experiment 1)	1 (SE1)	4 (AE1 - AE4)
group2	B & CR-E (experiment 1)	2 (SE2, SE3)	5 (AE5 - AE9)
group3	B & CR-E (experiment 1)	1 (SE4)	5 (AE10 - AE14)
group4	B & CR-E (experiment 1)	1 (SE5)	5 (AE15 - AE19)
group5	B & CR-N (experiment 2)	3 (SN1 - SN3)	4 (AN1 - AN4)
group6	B & CR-N (experiment 2)	3 (SN4 - SN6)	4 (AN5 - AN8)

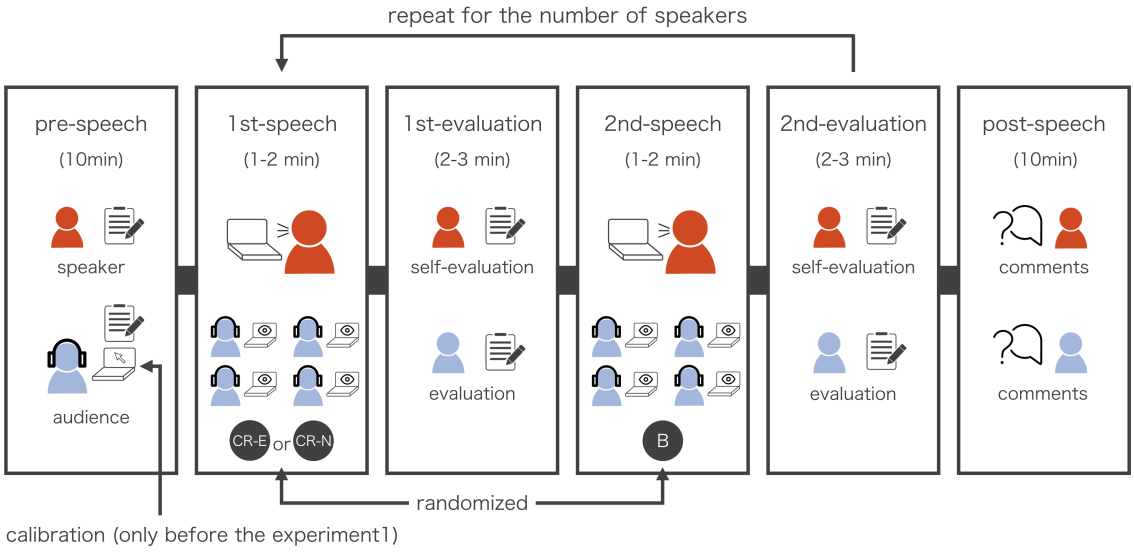

Figure 5: The overview of the experiment 1 and experiment 2.

manuscripts, they were allowed to prepare their thoughts before their presentations. Each presentation lasted for approximately two minutes. While this was a short time for a presentation, speakers did not need to make an effort to prepare, which could led to the natural speech. Before each presentation, speakers filled out a questionnaire about the State Anxiety [60] to check their anxiety level for the presentation. One presentation was given under *condition B*, while the other was given under *condition CR-E* (experiment 1) or *condition CR-N* (experiment 2). The order of conditions was randomized. The total time of the speech and evaluation sessions was about 10-30 minutes ((2-min speeches + 3-min evaluation periods) * (1 to 3)-presenter * 2-condition). After a speaker finished their presentation, the speaker and the audience answered a questionnaire on the presentation quality for about a few minutes (evaluation session). The speakers could not observe evaluations by the audiences and vice versa. Each speaker did not listen to the other speakers' presentations. A series of speech and evaluation sessions were repeated for each speaker. Finally, in the post-speech session, we collected their comments regarding the usability of the system, as we mentioned in Section 4.2. The audio of the presentations was recorded for analyzing the number of fillers.

5 EXPERIMENT 1: DISPLAYING EYE GAZE REACTIONS

In this section, we explain the first experiment, which explored the use of collective eye gaze reactions of multiple audiences in online communication.

5.1 Detailed Procedure

After audiences filled out the questionnaire in the pre-speech session, they started to calibrate their gaze following an experimenter's instruction as we described in Section 3.2. The topic in both *condition CR-E and B* was: "Please explain experiences in places where you've been before." The speakers could choose choose places either in Japan or in the world. They gave their presentations while seeing a map of Japan or the world. They talked about different places in the two presentations. The audiences also saw the same map as speakers' during the presentation through the audience client. Figure 6 presents a speaker's browser screen. A heat map was visualized in *condition CR-E*. On the other hand, speakers could see only a map and the remaining time in *condition B*. The speakers were allowed to place blue markers on a map in advance where to explain in their speech. These markers were also visible on the audience clients.

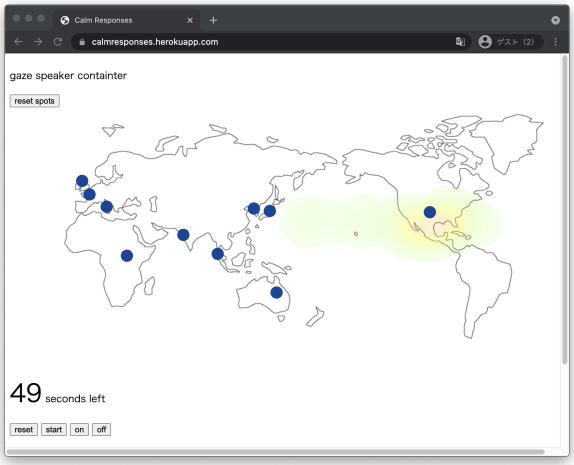

Figure 6: View of the speakers in *condition CR-E*, in which the audience eye gaze movements were visualized as a heat map. In both *CR-E and B conditions*, the speakers were asked to give presentations while seeing a map (Japan or the world map). They could place blue markers in advance where they would explain in their speeches. These markers were also visible on the audience clients.

5.2 Result

How Did the Speakers Change Subjectively?

The result of the paired t-test showed no significant difference between the baseline and system conditions on the State Anxiety (p = .36) which was measured right before the presentation (See Section 4.4). Although the average self-rating scores were higher in *condition CR-E*, there was no significant difference between the two conditions (p = .085) as denoted in Figure 7A. We further analyzed the effect of CalmResponses on the speakers via collected comments. Three out of five speakers raised their self-evaluation scores. We found three strengths of the system through feedback from these speakers. First, they mentioned that they felt a sense of relief in using the system: *"I felt more relieved because I could figure out whether the audiences understood my speech"* [SE1]. *"Heat map helped me feel at ease"* [SE2]. Second, the speakers mentioned that the system helped them feel more aware of the audience: *"I thought audiences were listening to my speech"* [SE1]. *"I became more aware of the presence of the audience"* [SE2]. *"Since I found that eye gaze positions moved every time I changed the location, I could confirm that the audience did listen to my presentation"* [SE4]. Finally, one speaker reported feeling more confident about his speech: *"I spoke with confidence, especially because I could tell if the audience understood me or not when I mentioned the names of places"* [SE4]. While all speakers interpreted eye gaze positions as audience interest or attention, it is interesting to note that some speakers also interpreted eye gaze as the audience understandings. These findings (feeling relieved, awareness of the audience, and confidence) are aligned with previous works [50, 54, 66]. Thus, it is implied that the system could positively affect speakers.

There were two speakers whose ratings did not increase due to the introduction of CalmResponses. One speaker felt that the audience did not necessarily see where the speakers were talking about: *"Although I felt relieved to know that audiences paid attention to me, at the same time, I felt a bit anxious because they sometimes saw where I was not explaining"* [SE3]. Although this was partly because of the inaccuracy of eye tracking, it is true that the audiences were not always paying attention to the same point that the speaker was talking about. The other participant pointed out that he did not see the visualization much: *"Since I was concentrating on remembering my experience, honestly, I didn't see the display so often"* [SE5]. This participant also argued that the system would be more beneficial when using more materials for the presentations. In fact, some audience members mentioned the same point: *"I thought that my eye gaze movements would changed more dynamically in slide-based lectures or presentations"* [AE5]. Therefore, we need to consider and compare various situations and explore more suitable ones for the visualization of eye gaze movements.

How Did the Speakers Change Objectively?

Audience rating scores were not significantly different between the two conditions (p = .06) as described in Figure 7B. We also found some comments from the audiences that supported this result: *"I couldn't notice the difference between the two speeches"* [AE9]. In contrast, in terms of the number of fillers, we did find a significant difference (p = .04) as described in Figure 7C. Since the number of fillers in speech is negatively correlated to with speakers' performance or anxiety [23], this result was incompatible with the result of audience rating scores. To understand more objective speakers' change, it will be a promising approach to use other measures such as speech evaluation by experts [14].

How Did CalmResponses Affect the Audience?

We asked the 19 audiences how they perceived the system. First, many of the audiences mentioned that it was easy to use the system (N=10): *"I didn't need any complex procedure"* [AE7]. *"Since all I needed was a laptop with a webcam, it was easy to introduce"* [AE8]. In contrast, two audiences had trouble with their laptops freezing in the middle of the *condition CR-E* (AE3, AE14), and a few audience members pointed out that it was inconvenient to use the system in addition to Zoom: *"One of the disadvantages was that I had to open the system through Google Chrome other than Zoom"* [AE17]. We expect that CalmResponses could be integrated into existing teleconferencing tools to address this usability issue.

We also asked audience members where they were looking during presentations. All audience members reported they looked at the locations mentioned by the speakers. However, the system made a few audiences feel nervous (N=3): *"I think it was weird for speakers both when I was gazing at a single point and when I changed points so frequently"* [AE4]. *"I felt cramped a little because speakers easily found out whether we were listening or not"* [AE11]. These comments imply that we need to consider their privacy. Still, we believe our collective visualization using a heat map could mitigate the issue since it does not identify individual gaze movements.

Interestingly, on the other hand, some audiences claimed that they were encouraged by the system to actively participate in the

Figure 7: Self-rating score (A), Audience rating score (B), Number of fillers per minute (C). p < .05 is marked as *.

presentations (N=5): *"I felt that I needed to be a good listener"* [AE2]. *"By being aware that my gaze positions were seen by speakers, I could focus more on listening to the speech"* [AE17]. In terms of online learning, it is difficult for learners to maintain their concentration, and several works have sought to address this problem [2, 32]. Our system can provide an alternative approach to the issue, although we need to balance users' feelings between active participation and nervousness as we mentioned in the previous paragraph.

Finally, we asked the audience members whether they wanted to see other audience members' eye gaze positions. Most of them answered positively about this question (N=15): *"I'm curious about what other audiences are interested in while listening to the presentations"* [AE3]. *"When I knew where others were looking at, I could notice the difference and similarity of the reactions between others and me"* [AE16]. In addition, they mentioned that they wanted to watch their own eye gaze positions as well. This point is consistent with a previous research that has emphasized the importance of providing audiences with visual feedback on their own actions [4]. In response, we modified the experiment to share the audience reactions not only with the speakers but also with the audience participants when we conducted the second experiment.

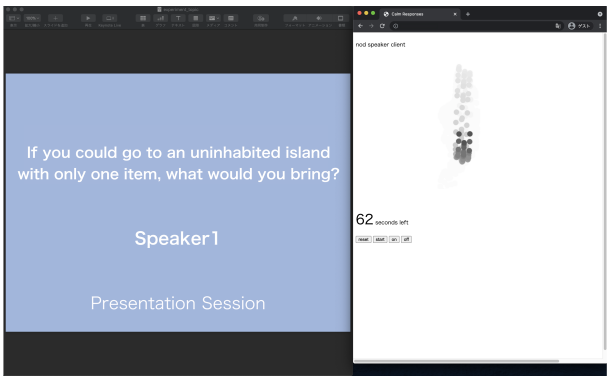

Figure 8: The example shared display on Zoom during presentation in *condition CR-N*. While only the slide about a topic and the remaining time were presented in *condition B*, the head movements of audiences were also presented in *condition CR-N*. All participants including audiences could see them.

6 EXPERIMENT 2: DISPLAYING NOD REACTIONS

In this section, we explain the second experiment, which explored the use of collective nod reactions of multiple audiences in online communication.

6.1 Detailed Procedure

Based on the feedback received in Experiment 1, we shared head movement visualizations with both speakers and audiences on Zoom in Experiment 2. After the audiences accessed CalmResponses, they switched to Zoom, and a speaker and audiences viewed a shared screen during several presentations. Figure 8 shows the shared screen on Zoom. All participants (speakers + audiences) could see the head movements' visualization, the remaining time, and a topic of the presentation in *condition CR-N*. In *condition B*, on the other hand, they could not see the visualization of the reactions. The presentation topics in both *condition CR-N* and *condition B* were: "If you could go to an uninhabited island with only one item, what would you bring?", "If you could use 100 thousand dollars in only one day, what would you use it for?", and "If you had a time machine, what period do you want to visit?". The speakers talked about different topics in the two presentations.

6.2 Result

How Did the Speakers Change Subjectively?

The result of the paired t-test showed no significant difference between the baseline and system conditions on the State Anxiety (p = .41). Figure 9A shows the self-evaluation scores in both *condition B* and *condition CR-N*. This score showed a significant difference between the two conditions (p = .024). This suggests that the use of the system increased speakers' self-evaluation ratings. Four out of six speakers exhibited higher self-evaluation scores in *condition CR-N*. In the collected comments, we found that they felt relieved when they were using CalmResponses, which led to high self-evaluation scores: *"I had a calm feeling when I used the system because I was sure that the audience was listening"* [SN4]. *"I was mentally relieved*

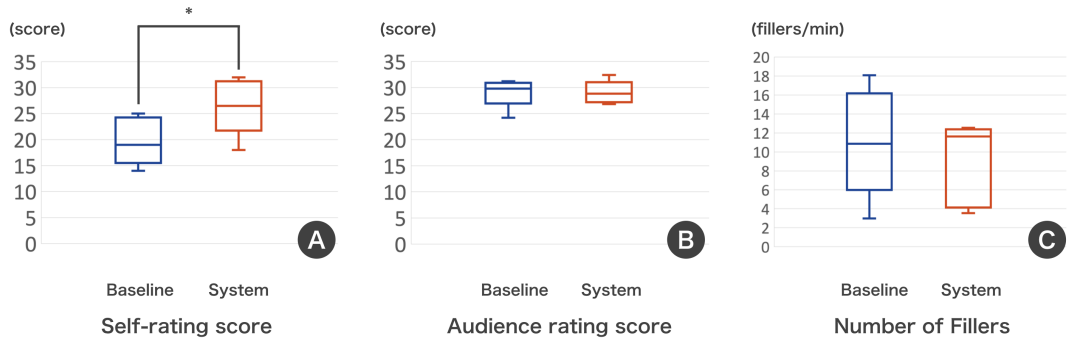

Figure 9: Self-rating score (A), Audience rating score (B), Number of fillers per minute (C). p < .05 is marked as *.

when I saw reactions because they expressed the agreement" [SN6]. This feedback was consistent with the finding of Experiment 1. Another speaker implied that the higher rating was related to the expansion of the content she talked about: *"I could expand on the content smoothly and speak more without thinking too much due to the feedback from the system"* [SN5]. This is consistent with the finding from prior research that audience feedback helped speakers speak more [50]. Another speaker associated the self-rating with the fact that they could look at themselves objectively: *"I was able to consider the audience reactions with the system, so my self-rating score got higher"* [SN1].

However, two speakers mentioned that the system did not affect their self-evaluations much. One of them pointed out that nodding was not enough reaction to estimate audience engagement: *"I could not tell whether the audience was feeling enjoyment or bored only from the nodding feedback"* [SN3]. The other participant did not observe many reactions: *"Even though the head movements were visualized, there were not many nodding reactions"* [SN2]. Despite this, SN2 was aware of his own presentation: *"I tried to improve my presentation so that audiences would react more"* [SN2]. This aspect could be considered one of the system's potential benefits.

How Did the Speakers Change Objectively?

Figure 9B shows the audience rating scores. There was no significant difference between the two conditions (p = .42). This result indicates that the system had a limited effect on the objective speech quality. This was also consistent with the small difference in the number of fillers per minute, as described in Figure 9C (p = .21).

While the number of fillers did not change much, one speaker's fillers were remarkably reduced. Figure 10 shows the timeline of the middle of the 20-second speeches given by SN5 who did the first speech under *condition B* and the second under *condition CR-N*, illustrating how the speaker's number of fillers per minute was much lower in *condition CR-N* (4.04 fillers / min) than in *condition B* (8.25 fillers / min). As can be seen in this figure, in *condition B*, SN5 often uttered fillers, which indicated that SN5 could not put the idea into words in those periods. Moreover, SN5 did not pause much during the speech in *condition B*, as there was no reaction for SN5 to observe. In comparison, in *condition CR-N*, the speaker rarely uttered fillers.

To further explore how the speaker perceived the system, we correlated the speech with head movements visualization. The bottom of the Figure 10 shows how audience head movements were displayed in *condition CR-N*. In this condition, SN5 gave a presentation about "If you could go to an uninhabited island with only one item, what would you bring?". Before (a), SN5 explained the premise of the topic, and at this time, the head movements were relatively still. When SN5 claimed that fire and water are important to life on an uninhabited island at (a), many audience members nodded. After that, SN5 paused the speech a while, presumably because SN5 wanted to observe the audience reactions. Cursors continued to move vertically, while the speakers decided to bring a tool to obtain fire or water at (b). At this point, the speaker sometimes paused the speech. From (c), SN5 was thinking about a concrete tool to bring in the island, and head movements were relatively still again. This finding implied that the audience members nodded when they heard the speaker's opinion and in that moment, the speaker used pauses to observe reactions. Although further verification is demanded, this analysis implied the efficacy of collectively presenting nodding reactions to the speakers in terms of improving their fluency of the presentation.

7 DISCUSSION

7.1 The Effect of Collective Reactions

In the present study, we examined how using a browser-based system to monitor audience reactions affected speakers' behaviors. We found a trend in which speakers' self-rating scores were higher in *condition CR-N*, and the number of fillers was smaller in *condition CR-E*. In this section, we discuss the implications of eye gaze and nodding as audience reactions and collective visualizations based on the experimental results.

7.1.1 **Advantage of Each Reaction**. It would be essential to take advantage of each reaction rather than comparison. In their interviews, participants suggested the suitable situations for eye gaze sharing as also described in the result of the first experiment: *"I thought slide-based presentations were good for the system because visual stimuli frequently changed"* [SE5]. *"The system would be more effective when speakers were using several slides"* [AE16]. In comparison, nodding might be helpful in bidirectional communication due

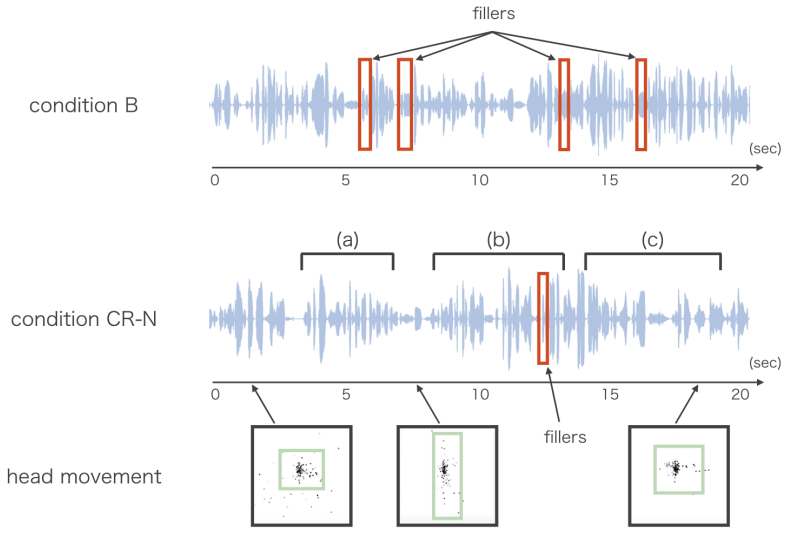

Figure 10: Analysis of a speaker's 20 second speech in both conditions. We annotated filler positions manually and they are marked as orange boxes. Moreover in *condition CR-N*, we show how nod reactions were presented during the speech.

to turn-taking and coordination functions. With regard to other reactions, speakers telling a funny story to audiences at a comedy show might want to observe smiles from facial expressions rather than eye gaze or head movements. Since we observed different results in terms of self-rating scores and the number of fillers between eye gaze and nodding, it is implied that the result will change depending on modalities and situations. It is desirable to further sophisticate the feedback for various situations given these different characteristics of different reactions.

7.1.2 **Privacy Concern**. Although we only sent eye gaze positions and nose tip velocities through the server, privacy concerns are an inevitable issue when using webcams [57]. Interviews with audience members reflected this: *"I was worried a little about being watched my video"* [AE10]. *"I felt I was being monitored"* [AE13]. *"I was a little embarrassed when using the system"* [AN1]. These statements imply that the system invaded some audiences' privacy. However, other audiences expressed the opposite: *"I was slightly embarrassed, but that was almost no problem because I thought that the system could keep anonymity"* [AE3]. *"I thought that my impression would change depending on whether I could identify individuals from head movements or not. In this experiment, I was relieved that others couldn't identify individuals from the visualization"* [AN4]. These comments suggest that collectively presenting information retained anonymity to some extent because individuals could not identify who were nodding based on the cursor visualization and where others were looking based on heatmap visualization. The comments in the second experiment had a stronger tendency than the first one, which suggests that the audiences in the second experiment reported fewer privacy concerns than those in the first experiment. Anonymity is one of the important advantages of remote communication, as mentioned in [30]. Since privacy concerns are the main issue in the psychological and physical aspects of online sensing [15], collective visualization will pave the way for privacy-preserving audience sensing and feedback technologies. However, in the present experiment, there was still a variation in audience perception regarding the point of privacy. This relationship between collective visualization and privacy concerns must be explored further.

7.1.3 **Co-Presence**. Regarding the shared visualization of collective head movements in the second experiment, some audiences reported another implication. *"I felt that I was involved in the presentations"* [AN2], *"I was not sure the speakers' speeches were improved, but I felt a sense of unity from others' head movements visualization"* [AN6]. These comments indicate that the proposed system formed co-presence, a sense of being together psychologically [8]. This can also be related to emotional contagion [29], through which audience members are affected by one another. This is consistent with prior studies that have proposed systems of sharing visual cues with remote users to improve co-presence [35]. Co-presence has been one of the main topics in distributed communication [73], and many works have proposed various methods of forming co-presence [3, 33, 35, 55]. These existing methods use additional devices (e.g., HMD) or require users' intentional actions. Our proposed approach's strengths, in contrast with these existing approaches, are that it (1) requires only a commonly available device and (2) influences multiple audiences at the same time. Our future research will investigate whether co-presence generated by the system can improve audience performance in terms of their concentration.

7.2 Further Applications

Regarding nod reactions, users can send reactions through the system in asynchronous situations, such as social networking services. Figure 11A shows the examples of how users send nod reactions with CalmResponses and how they are visualized on Twitter. Although existing reactions on Twitter (e.g, *liking* or *retweeting a post*) are discrete, two-bit information, we can send continuous and

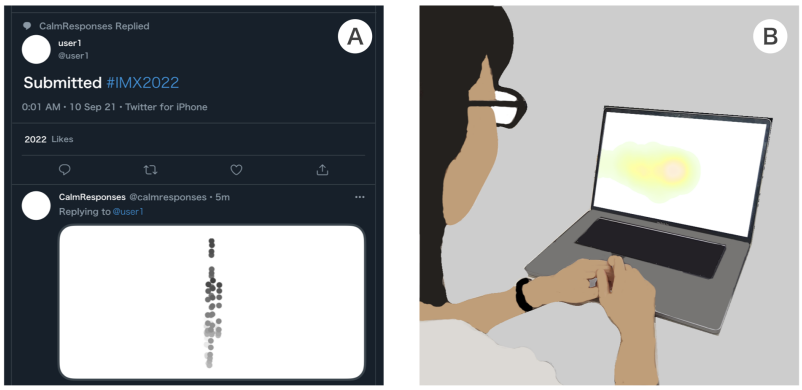

Figure 11: Application scenarios in asynchronous situations. (A) Nodding Reactions on Twitter. Instead of sending two-bit information such as like or not, users can send nod reactions changing the strength of head movements to show the degree of empathy. (B) Eye gaze positions in lecture videos. Accumulated gaze positions of past learners can help new learners to observe important points of the videos quickly.

complex reactions, such as one's degree of empathy, by changing the strength of one's head movements.

We can also leverage eye gaze reactions in asynchronous situations. Since we can infer one's interest from eye gaze movements, collective eye gaze reactions can be used to estimate the importance of the contents in videos. Figure 11B shows an example of visualization of past learners' gaze data on a lecture video. One of the problems with lecture videos is that learners seldom interact with their instructors and often lose their attention while watching them [24, 31]. To overcome this problem, past learners' collective gaze visualization can help new learners to understand the important topic of the lecture videos quickly. Since these applications do not require active participation of audiences, it will be easier to collect reactions than other systems visualizing explicit reactions and interactions of past audiences or learners [34, 41]. Note that in these contexts, we need to consider some aspects such as note-taking and additional displays, which could invalidate the sensing method of the system.

7.3 Limitations

Although the present study produced some positive results, it is not without limitations. We first refer to the limitations of the system. The accuracy of the current eye-tracking module was not so high. This can be attributed to several aspects, such as the lighting conditions. Some speakers and audience members pointed out this issue: *"I thought that it would be better if the accuracy was higher"* [SE3]. *"There was room for improvement in accuracy"* [AE10]. We will address this limitation to use other eye-tracking modules with higher accuracy such as [77].

In terms of the limitations of the experiments, we did not compare eye gaze and nod reactions with explicit reactions such as texts or emoticons. There is little research on the usage of these reactions in synchronous distributed communication compared to asynchronous communication. Moreover, we had limited sample sizes, especially the speaker sample size. We need to investigate further the difference between explicit reactions and eye gaze or nodding in the context of a larger-scale experiment. Another limitation is a relatively short time (approximately two minutes) and small number of topics (five topics in total) for presentation used in our experiments, compared to existing research that conducted similar experiments [50, 66]. This choice has both advantages and disadvantages. On one hand, since speakers do not need much time to prepare for short speech, it could induce spontaneous and natural presentations. On the other hand, it can not be denied that approximately two minute presentations are not enough for audiences to properly assess them. Furthermore, there are various topics for actual communications, and we still do not know whether the system can be applied to other topics. In future work, we will introduce the system to in-the-wild situations such as lectures or webinars and evaluate its efficacy with longer presentations with various topics.

8 CONCLUSION

In this study, we proposed a browser-based system that displays audience eye gaze positions and head movements collectively to speakers in synchronous remote communication using a built-in webcam. We conducted two experiments to evaluate the effectiveness of the proposed system. We found that the system helped speakers reduce the number of fillers during their presentation and raised their self-rating scores. Based on our results and findings, we discussed the potential benefits of the system in terms of privacy concerns, co-presence of the audience. We also discussed the application scenarios of the system beyond the synchronous online communication. We believe that our approach and findings will enhance the experience of future remote communication by leveraging its benefits for both speakers and audiences.

ACKNOWLEDGMENTS

This work was supported by the commissioned research by National Institute of Information and Communications Technology (NICT) Japan, JST CREST Grant Number JPMJCR17A3, and JST Moonshot R&D Grant Number JPMJMS2012.

REFERENCES

[1] Riku Arakawa and Hiromu Yakura. 2019. REsCUE: A framework for REal-time feedback on behavioral CUEs using multimodal anomaly detection. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*. ACM, New York, NY, 572. https://doi.org/10.1145/3290605.3300802

[2] Riku Arakawa and Hiromu Yakura. 2021. Mindless Attractor: A False-Positive Resistant Intervention for Drawing Attention Using Auditory Perturbation. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*. ACM, New York, NY, 99:1–99:15. https://doi.org/10.1145/3411764.3445339

[3] Riku Arakawa and Hiromu Yakura. 2021. Reaction or Speculation: Building Computational Support for Users in Catching-Up Series Based on an Emerging Media Consumption Phenomenon. *Proc. ACM Hum. Comput. Interact.* 5, CSCW1 (2021), 1–28. https://doi.org/10.1145/3449225

[4] Louise Barkhuus and Tobias Jørgensen. 2008. Engaging the crowd: studies of audience-performer interaction. In *Extended Abstracts Proceedings of the 2008 Conference on Human Factors in Computing Systems, CHI 2008, Florence, Italy, April 5-10, 2008*. ACM, New York, NY, 2925–2930. https://doi.org/10.1145/1358628.1358785

[5] Giorgi Basilaia and David Kvavadze. 2020. Transition to online education in schools during a SARS-CoV-2 coronavirus (COVID-19) pandemic in Georgia. *Pedagogical Research* 5, 4 (2020).

[6] Ronald Bassett, Ralph R Behnke, Larry W Carlile, and Jimmie Rogers. 1973. The effects of positive and negative audience responses on the autonomic arousal of student speakers. *Southern Journal of Communication* 38, 3 (1973), 255–261.

[7] Konstantinos Bousmalis, Marc Mehu, and Maja Pantic. 2013. Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools. *Image and Vision Computing* 31, 2 (2013), 203–221. https://doi.org/10.1016/j.imavis.2012.07.003

[8] Saniye Tugba Bulu. 2012. Place presence, social presence, co-presence, and satisfaction in virtual worlds. *Computers & Education* 58, 1 (2012), 154–161.

[9] Benjamin T. Carter and Steven G. Luke. 2020. Best practices in eye tracking research. *International Journal of Psychophysiology* 155 (Sept. 2020), 49–62. https://doi.org/10.1016/j.ijpsycho.2020.05.010

[10] Frank R. Castelli and Mark A. Sarvary. 2021. Why students do not turn on their video cameras during online classes and an equitable and inclusive plan to encourage them to do so. *Ecology and Evolution* 11, 8 (Jan. 2021), 3565–3576. https://doi.org/10.1002/ece3.7123

[11] Lei Chen, Gary Feng, Jilliam Joe, Chee Wee Leong, Christopher Kitchen, and Chong Min Lee. 2014. Towards Automated Assessment of Public Speaking Skills Using Multimodal Cues. In *Proceedings of the 16th International Conference on Multimodal Interaction, ICMI 2014, Istanbul, Turkey, November 12-16, 2014*. ACM, New York, NY, 200–203. https://doi.org/10.1145/2663204.2663265

[12] Mauro Cherubini, Marc-Antoine Nüssli, and Pierre Dillenbourg. 2008. Deixis and gaze in collaborative work at a distance (over a shared map): a computational model to detect misunderstandings. In *Proceedings of the Eye Tracking Research & Application Symposium, ETRA 2008, Savannah, Georgia, USA, March 26-28, 2008*. ACM, New York, NY, 173–180. https://doi.org/10.1145/1344471.1344515

[13] Mathieu Chollet and Stefan Scherer. 2017. Perception of Virtual Audiences. *IEEE Computer Graphics and Applications* 37, 4 (2017), 50–59. https://doi.org/10.1109/MCG.2017.3271465

[14] Mathieu Chollet, Torsten Wörtwein, Louis-Philippe Morency, Ari Shapiro, and Stefan Scherer. 2015. Exploring feedback strategies to improve public speaking: an interactive virtual audience framework. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp 2015, Osaka, Japan, September 7-11, 2015*. ACM, New York, NY, 1143–1154. https://doi.org/10.1145/2750858.2806060

[15] Martin Cooney, Sepideh Pashami, Anita Pinheiro Sant'Anna, Yuantao Fan, and Slawomir Nowaczyk. 2018. Pitfalls of Affective Computing: How can the automatic visual communication of emotions lead to harm, and what can be done to mitigate such risks. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*. ACM, New York, NY, 1563–1566. https://doi.org/10.1145/3184558.3191611

[16] Sarah D'Angelo and Andrew Begel. 2017. Improving Communication Between Pair Programmers Using Shared Gaze Awareness. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017*. ACM, New York, NY, 6245–6290. https://doi.org/10.1145/3025453.3025573

[17] Sarah D'Angelo and Darren Gergle. 2016. Gazed and Confused: Understanding and Designing Shared Gaze for Remote Collaboration. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*. ACM, New York, NY, 2492–2496. https://doi.org/10.1145/2858036.2858499

[18] Sarah D'Angelo and Darren Gergle. 2018. An Eye For Design: Gaze Visualizations for Remote Collaborative Work. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*. ACM, New York, NY, 349. https://doi.org/10.1145/3173574.3173923

[19] Charles Darwin and Phillip Prodger. 1998. *The expression of the emotions in man and animals*. Oxford University Press, USA.

[20] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. 2018. Unobtrusive assessment of students' emotional engagement during lectures using electrodermal activity sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–21. https://doi.org/10.1145/3264913

[21] Katsuya Fujii and Jun Rekimoto. 2019. SubMe: An Interactive Subtitle System with English Skill Estimation Using Eye Tracking. In *Proceedings of the 10th Augmented Human International Conference 2019, Reims, France, March 11-12, 2019*. ACM, New York, NY, 23:1–23:9. https://doi.org/10.1145/3311823.3311865

[22] Shkurta Gashi, Elena Di Lascio, and Silvia Santini. 2019. Using unobtrusive wearable sensors to measure the physiological synchrony between presenters and audience members. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (2019), 1–19.

[23] Alexander M Goberman, Stephanie Hughes, and Todd Haydock. 2011. Acoustic characteristics of public speaking: Anxiety and practice effects. *Speech communication* 53, 6 (2011), 867–876.

[24] Philip J. Guo, Juho Kim, and Rob Rubin. 2014. How video production affects student engagement: an empirical study of MOOC videos. In *First (2014) ACM Conference on Learning @ Scale, L@S 2014, Atlanta, GA, USA, March 4-5, 2014*. ACM, New York, NY, 41–50. https://doi.org/10.1145/2556325.2566239

[25] Aman Gupta, Finn L. Strivens, Benjamin Tag, Kai Kunze, and Jamie A. Ward. 2019. Blink as you sync: uncovering eye and nod synchrony in conversation using wearable sensing. In *Proceedings of the 23rd International Symposium on Wearable Computers, UbiComp 2019, London, UK, September 09-13, 2019*. ACM, New York, NY, 66–71. https://doi.org/10.1145/3341163.3347736

[26] Joanna Hale, Jamie A Ward, Francesco Buccheri, Dominic Oliver, and Antonia F de C Hamilton. 2020. Are you on my wavelength? Interpersonal coordination in dyadic conversations. *Journal of nonverbal behavior* 44, 1 (2020), 63–83.

[27] Mariam Hassib, Daniel Buschek, Pawel W. Wozniak, and Florian Alt. 2017. HeartChat: Heart Rate Augmented Mobile Chat to Support Empathy and Awareness. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017*. ACM, New York, NY, 2239–2251. https://doi.org/10.1145/3025453.3025758

[28] Mariam Hassib, Stefan Schneegass, Philipp Eiglsperger, Niels Henze, Albrecht Schmidt, and Florian Alt. 2017. EngageMeter: A System for Implicit Audience Engagement Sensing Using Electroencephalography. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017*. ACM, New York, NY, 5114–5119. https://doi.org/10.1145/3025453.3025669

[29] Elaine Hatfield, John T. Cacioppo, and Richard L. Rapson. 1993. Emotional Contagion. *Current Directions in Psychological Science* 2, 3 (June 1993), 96–100. https://doi.org/10.1111/1467-8721.ep10770953

[30] James D. Hollan and Scott Stornetta. 1992. Beyond Being There. In *Conference on Human Factors in Computing Systems, CHI 1992, Monterey, CA, USA, May 3-7, 1992, Proceedings*. ACM, New York, NY, 119–125. https://doi.org/10.1145/142750.142769

[31] Kate S Hone and Ghada R El Said. 2016. Exploring the factors affecting MOOC retention: A survey study. *Computers & Education* 98 (2016), 157–168.

[32] Stephen Hutt, Kristina Krasich, James R. Brockmole, and Sidney K. D'Mello. 2021. Breaking out of the Lab: Mitigating Mind Wandering with Gaze-Based Attention-Aware Technology in Classrooms. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*. ACM, New York, NY, 52:1–52:14. https://doi.org/10.1145/3411764.3445269

[33] Dongsik Jo, Ki-Hong Kim, and Gerard Jounghyun Kim. 2016. Effects of avatar and background representation forms to co-presence in mixed reality (MR) teleconference systems. In *SIGGRAPH ASIA 2016, Macao, December 5-8, 2016 - Virtual Reality meets Physical Reality: Modelling and Simulating Virtual Humans and Environments*. ACM, New York, NY, 12:1–12:4. https://doi.org/10.1145/2992138.2992146

[34] Juho Kim, Philip J. Guo, Carrie J. Cai, Shang-Wen (Daniel) Li, Krzysztof Z. Gajos, and Robert C. Miller. 2014. Data-driven interaction techniques for improving navigation of educational videos. In *The 27th Annual ACM Symposium on User Interface Software and Technology, UIST '14, Honolulu, HI, USA, October 5-8, 2014*. ACM, New York, NY, 563–572. https://doi.org/10.1145/2642918.2647389

[35] Seungwon Kim, Gun Lee, Nobuchika Sakata, and Mark Billinghurst. 2014. Improving co-presence with augmented visual communication cues for sharing experience through video conference. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, IEEE Computer Society, New York, NY, 83–92. https://doi.org/10.1109/ISMAR.2014.6948412

[36] Masakatsu Kubota, Tomio Watanabe, and Yutaka Ishii. 2019. A Speech Promotion System by Using Embodied Entrainment Objects of Spoken Words and a Listener Character for Joint Attention. In *Proceedings of the 7th International Conference on Human-Agent Interaction, HAI 2019, Kyoto, Japan, October 06-10, 2019*. ACM, New York, NY, 311–312. https://doi.org/10.1145/3349537.3352803

[37] Manu Kumar, Jeff Klingner, Rohan Puranik, Terry Winograd, and Andreas Paepcke. 2008. Improving the accuracy of gaze input for interaction. In *Proceedings of the Eye Tracking Research & Application Symposium, ETRA 2008, Savannah, Georgia, USA, March 26-28, 2008*. ACM, New York, NY, 65–68. https:

//doi.org/10.1145/1344471.1344488

[38] Kai Kunze, Yuzuko Utsumi, Yuki Shiga, Koichi Kise, and Andreas Bulling. 2013. I know what you are reading: recognition of document types using mobile eye tracking. In *Proceedings of the 17th Annual International Symposium on Wearable Computers. ISWC 2013, Zurich, Switzerland, September 8-12, 2013.* ACM, New York, NY, 113–116. https://doi.org/10.1145/2493988.2494354

[39] Raja S. Kushalnagar and Christian Vogler. 2020. Teleconference Accessibility and Guidelines for Deaf and Hard of Hearing Users. In *ASSETS '20: The 22nd International ACM SIGACCESS Conference on Computers and Accessibility, Virtual Event, Greece, October 26-28, 2020.* ACM, New York, NY, 9:1–9:6. https://doi.org/10.1145/3373625.3417299

[40] Celine Latulipe, Erin A. Carroll, and Danielle M. Lottridge. 2011. Love, hate, arousal and engagement: exploring audience responses to performing arts. In *Proceedings of the International Conference on Human Factors in Computing Systems, CHI 2011, Vancouver, BC, Canada, May 7-12, 2011.* ACM, New York, NY, 1845–1854. https://doi.org/10.1145/1978942.1979210

[41] Yi-Chieh Lee, Wen-Chieh Lin, Fu-Yin Cherng, Hao-Chuan Wang, Ching-Ying Sung, and Jung-Tai King. 2015. Using Time-Anchored Peer Comments to Enhance Social Interaction in Online Educational Videos. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015.* ACM, New York, NY, 689–698. https://doi.org/10.1145/2702123.2702349

[42] Jie Li, Xinning Gui, Yubo Kou, and Yukun Li. 2019. Live Streaming as Co-Performance: Dynamics between Center and Periphery in Theatrical Engagement. *Proc. ACM Hum. Comput. Interact.* 3, CSCW (2019), 64:1–64:22. https://doi.org/10.1145/3359166

[43] Shao-Kang Lo. 2008. The nonverbal communication functions of emoticons in computer-mediated communication. *Cyberpsychology & behavior* 11, 5 (2008), 595–597.

[44] Valeria Lo Iacono, Paul Symonds, and David HK Brown. 2016. Skype as a tool for qualitative research interviews. *Sociological Research Online* 21, 2 (2016), 103–117.

[45] Zhicong Lu, Rubaiat Habib Kazi, Li-Yi Wei, Mira Dontcheva, and Karrie Karahalios. 2021. StreamSketch: Exploring Multi-Modal Interactions in Creative Live Streams. *Proc. ACM Hum. Comput. Interact.* 5, CSCW1 (2021), 1–26. https://doi.org/10.1145/3449132

[46] Peter D MacIntyre, Kimly A Thivierge, and J Renée MacDonald. 1997. The effects of audience interest, responsiveness, and evaluation on public speaking anxiety and related variables. *Communication research reports* 14, 2 (1997), 157–168.

[47] Divine Maloney, Guo Freeman, and Donghee Yvette Wohn. 2020. "Talking without a Voice": Understanding Non-verbal Communication in Social Virtual Reality. *Proc. ACM Hum. Comput. Interact.* 4, CSCW2 (2020), 175:1–175:25. https://doi.org/10.1145/3415246

[48] Evelyn Z McClave. 2000. Linguistic functions of head movements in the context of speech. *Journal of pragmatics* 32, 7 (2000), 855–878.

[49] Khadijah Mukhtar, Kainat Javed, Mahwish Arooj, and Ahsan Sethi. 2020. Advantages, Limitations and Recommendations for online learning during COVID-19 pandemic era. *Pakistan journal of medical sciences* 36, COVID19-S4 (2020), S27.

[50] Prasanth Murali, Javier Hernandez, Daniel McDuff, Kael Rowan, Jina Suh, and Mary Czerwinski. 2021. AffectiveSpotlight: Facilitating the Communication of Affective Responses from Audience Members during Online Presentations. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021.* ACM, New York, NY, 247:1–247:13. https://doi.org/10.1145/3411764.3445235

[51] Joshua Newn, Eduardo Velloso, Fraser Allison, Yomna Abdelrahman, and Frank Vetere. 2017. Evaluating Real-Time Gaze Representations to Infer Intentions in Competitive Turn-Based Strategy Games. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play, CHI PLAY 2017, Amsterdam, The Netherlands, October 15-18, 2017.* ACM, New York, NY, 541–552. https://doi.org/10.1145/3116595.3116624

[52] Kotaro Oomori, Akihisa Shitara, Tatsuya Minagawa, Sayan Sarcar, and Yoichi Ochiai. 2020. A Preliminary Study on Understanding Voice-only Online Meetings Using Emoji-based Captioning for Deaf or Hard of Hearing Users. In *ASSETS '20: The 22nd International ACM SIGACCESS Conference on Computers and Accessibility, Virtual Event, Greece, October 26-28, 2020.* ACM, New York, NY, 54:1–54:4. https://doi.org/10.1145/3373625.3418032

[53] Alexandra Papoutsaki, Patsorn Sangkloy, James Laskey, Nediyana Daskalova, Jeff Huang, and James Hays. 2016. WebGazer: Scalable Webcam Eye Tracking Using User Interactions. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI).* AAAI, IJCAI/AAAI Press, New York, NY, 3839–3845. http://www.ijcai.org/Abstract/16/540

[54] Dhaval Parmar and Timothy W. Bickmore. 2020. Making It Personal: Addressing Individual Audience Members in Oral Presentations Using Augmented Reality. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2 (2020), 55:1–55:22. https://doi.org/10.1145/3397336

[55] Iana Podkosova and Hannes Kaufmann. 2018. Co-presence and proxemics in shared walkable virtual environments with mixed colocation. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology, VRST 2018, Tokyo, Japan, November 28 - December 01, 2018.* ACM, New York, NY, 21:1–21:11. https://doi.org/10.1145/3281505.3281523

[56] Isabella Poggi, Francesca D'Errico, and Laura Vincze. 2010. Types of Nods. The Polysemy of a Social Signal. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta.* European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2010/summaries/596.html

[57] Rebecca S. Portnoff, Linda N. Lee, Serge Egelman, Pratyush Mishra, Derek Leung, and David A. Wagner. 2015. Somebody's Watching Me?: Assessing the Effectiveness of Webcam Indicator Lights. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015.* ACM, New York, NY, 1649–1658. https://doi.org/10.1145/2702123.2702164

[58] Stefan Scherer, Georg Layher, John Kane, Heiko Neumann, and Nick Campbell. 2012. An audiovisual political speech analysis incorporating eye-tracking and perception data. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12).* European Language Resources Association (ELRA), Istanbul, Turkey, 1114–1120.

[59] Oleg Spakov, Diederick C. Niehorster, Howell O. Istance, Kari-Jouko Räihä, and Harri Siirtola. 2019. Two-Way Gaze Sharing in Remote Teaching. In *Human-Computer Interaction - INTERACT 2019 - 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2-6, 2019, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 11747).* Springer, 242–251. https://doi.org/10.1007/978-3-030-29384-0_16

[60] Charles Donald Spielberger. 1989. *State-trait anxiety inventory: a comprehensive bibliography.* Consulting Psychologists Press.

[61] Moe Sugawa, Taichi Furukawa, George Chernyshov, Danny Hynds, Jiawen Han, Marcelo Padovani, Dingding Zheng, Karola Marky, Kai Kunze, and Kouta Minamizawa. 2021. Boiling Mind: Amplifying the Audience-Performer Connection through Sonification and Visualization of Heart and Electrodermal Activities. In *TEI '21: Fifteenth International Conference on Tangible, Embedded, and Embodied Interaction, Online Event / Salzburg, Austria, February 14-19, 2021.* ACM, New York, NY, 34:1–34:10. https://doi.org/10.1145/3430524.3440653

[62] Wei Sun, Yunzhi Li, Feng Tian, Xiangmin Fan, and Hongan Wang. 2019. How Presenters Perceive and React to Audience Flow Prediction In-situ: An Explorative Study of Live Online Lectures. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–19.

[63] Gahyun Sung, Tianyi Feng, and Bertrand Schneider. 2021. Learners Learn More and Instructors Track Better with Real-time Gaze Sharing. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.

[64] Jaime Teevan, Daniel J. Liebling, Ann Paradiso, Carlos Garcia Jurado Suarez, Curtis von Veh, and Darren Gehring. 2012. Displaying mobile feedback during a presentation. In *Mobile HCI '12, Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services, San Francsico, CA, USA, September 21-24, 2012.* ACM, New York, NY, 379–382. https://doi.org/10.1145/2371574.2371633

[65] Daniel Shu Wei Ting, Lawrence Carin, Victor Dzau, and Tien Y Wong. 2020. Digital technology and COVID-19. *Nature medicine* 26, 4 (2020), 459–461.

[66] Ha Trinh, Reza Asadi, Darren Edge, and T Bickmore. 2017. Robocop: A robotic coach for oral presentations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–24.

[67] Radu-Daniel Vatavu. 2015. Audience Silhouettes: Peripheral Awareness of Synchronous Audience Kinesics for Social Television. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video, TVX 2015, Brussels, Belgium, June 3-5, 2015.* ACM, New York, NY, 13–22. https://doi.org/10.1145/2745197.2745207

[68] Roel Vertegaal. 1999. The GAZE Groupware System: Mediating Joint Attention in Multiparty Communication and Collaboration. In *Proceeding of the CHI '99 Conference on Human Factors in Computing Systems: The CHI is the Limit, Pittsburgh, PA, USA, May 15-20, 1999*, Marian G. Williams and Mark W. Altom (Eds.). ACM, New York, NY, 294–301. https://doi.org/10.1145/302979.303065

[69] Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview.

[70] Bin Wang, Yukun Liu, Jing Qian, and Sharon K Parker. 2021. Achieving effective remote working during the COVID-19 pandemic: A work design perspective. *Applied Psychology* 70, 1 (2021), 16–59.

[71] Xi Wang, Andreas Ley, Sebastian Koch, David Lindlbauer, James Hays, Kenneth Holmqvist, and Marc Alexa. 2019. The Mental Image Revealed by Gaze Tracking. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019.* ACM, New York, NY, 609. https://doi.org/10.1145/3290605.3300839

[72] Tomio Watanabe. 2007. Human-Entrained E-COSMIC: Embodied Communication System for Mind Connection. In *Human Interface and the Management of Information. Methods, Techniques and Tools in Information Design, Symposium on Human Interface 2007, Held as Part of HCI International 2007, Beijing, China, July 22-27, 2007, Proceedings Part I (Lecture Notes in Computer Science, Vol. 4557).* Springer, 1008–1016. https://doi.org/10.1007/978-3-540-73345-4_114

[73] Andrew M. Webb, Chen Wang, Andruid Kerne, and Pablo César. 2016. Distributed Liveness: Understanding How New Technologies Transform Performance Experiences. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, CSCW 2016, San Francisco, CA, USA, February 27 - March 2, 2016*. ACM, New York, NY, 431–436. https://doi.org/10.1145/2818048.2819974

[74] Marcin Wlodarczak, Hendrik Buschmeier, Zofia Malisz, Stefan Kopp, and Petra Wagner. 2012. Listener head gestures and verbal feedback expressions in a distraction task. In *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog, INTERSPEECH2012 Satellite Workshop*.

[75] Nancy Yao, Jeff Brewer, Sarah D'Angelo, Mike Horn, and Darren Gergle. 2018. Visualizing Gaze Information from Multiple Students to Support Remote Instruction. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*. ACM, New York, NY, 1–6. https://doi.org/10.1145/3170427.3188453

[76] Alfred L. Yarbus. 1967. Eye Movements During Perception of Complex Objects. In *Eye Movements and Vision*. Springer US, 171–211. https://doi.org/10.1007/978-1-4899-5379-7_8

[77] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. 2020. ETH-XGaze: A Large Scale Dataset for Gaze Estimation Under Extreme Head Pose and Gaze Variation. In *Computer Vision – ECCV 2020*. Springer International Publishing, 365–381. https://doi.org/10.1007/978-3-030-58558-7_22

[78] Jilei Zhou, Jing Zhou, Ying Ding, and Hansheng Wang. 2019. The magic of danmaku: A social interaction perspective of gift sending on live streaming platforms. *Electronic Commerce Research and Applications* 34 (2019), 100815.