# A Preliminary Study on Understanding Voice-only Online Meetings Using Emoji-based Captioning for Deaf or Hard of Hearing Users

Kotaro Oomori
oomori@digitalnature.slis.tsukuba.ac.jp
University of Tsukuba
Digital Nature Group

Akihisa Shitara
University of Tsukuba
Digital Nature Group

Tatsuya Minagawa
University of Tsukuba
Digital Nature Group

Sayan Sarcar
University of Tsukuba

Yoichi Ochiai
wizard@slis.tsukuba.ac.jp
University of Tsukuba
Digital Nature Group
Pixie Dust Technologies, Inc.

## ABSTRACT

In the midst of the coronavirus disease 2019 pandemic, online meetings are rapidly increasing. Deaf or hard of hearing (DHH) people participating in an online meeting often face difficulties in capturing the affective states of other speakers. Recent studies have shown the effectiveness of emoji-based representation of spoken text to capture such affective states. Nevertheless, in voice-only online meetings, it is still not clear how emoji-based spoken texts can assist DHH people to understand the feelings of speakers without perceiving their facial expressions. We therefore conducted a preliminary experiment to understand the effect of emoji-based text representation during voice-only online meetings by leveraging an emoji-based captioning system. Our preliminary results demonstrate the necessity of designing an advanced system to help DHH people understanding the voice-only online meetings more meaningfully.

## CCS CONCEPTS

• **Human-centered computing → User studies**.

## KEYWORDS

User study, Voice-only online meeting, Emoji-based emotion expression, Deaf, Hard of hearing

**ACM Reference Format:**
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 *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '20), October 26–28, 2020, Virtual Event, Greece.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3373625.3418032

*ASSETS '20, October 26–28, 2020, Virtual Event, Greece*


## 1 INTRODUCTION

As online meetings have become more important during the coronavirus disease 2019 (COVID-19) pandemic, the use of real-time spoken text graphical user interface (GUI) systems in online meetings is increased. However, text-based communication face challenges in conveying information regarding user emotions compared to its speech-based counterpart. The lack of availability of nonverbal communication cues prevents the sender from clearly expressing the mood of a message, hindering recipients from complete understanding of the message [6].

In normal face-to-face telecommunication, Deaf or hard of hearing (DHH) people can obtain affective information of the others from their facial expressions. However, many online communication situations exist where people do not turn on their video (cluttered rooms, no makeup, etc.), and as a result, DHH people participating in voice-only group communications face difficulties in determining the affective state of the communications.

Recent studies have shown the effectiveness of emoji-based representation of spoken text to capture the affective state of the speakers [3]. However, in online voice-only meetings, no work has focused on understanding how emoji-based text representation can assist DHH people to perceive the affective state of speakers and make their conversation effective without seeing their facial expressions.

We, therefore conducted a preliminary experiment for measuring the effect of emoji augmented text representation on understanding voice-only online meetings for DHH users. For this experiment, we implemented an emoji-based captioning system that generates emoji-based spoken text by analyzing speech and emotion of speakers in a conversation by using the state-of-the-art APIs.

## 2 EMOJI-BASED CAPTIONING BASED ON SPEECH AND EMOTION IN A CONVERSATION

In this study, we implemented a system that generates emoji-based spoken text by analyzing speech and emotion of speakers in voice-only online meetings, as shown in Figure 1.

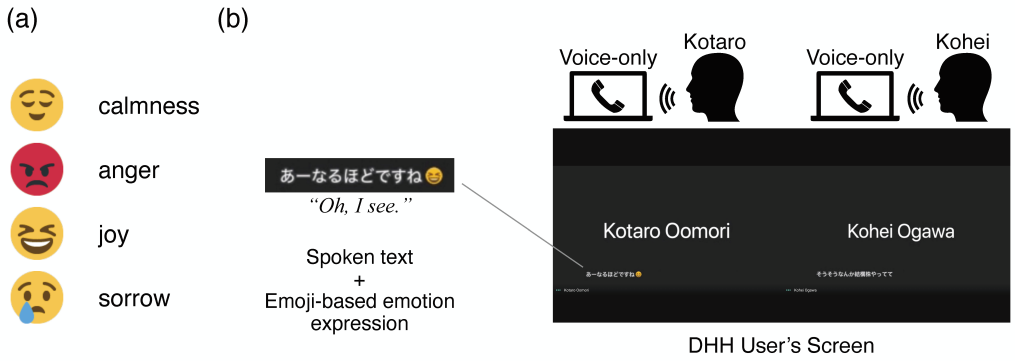


**Figure 1: Our emoji-based captioning system used for DHH users in online voice-only meetings. (a) List of emojis which we used. (b) DHH users can view emoji-based spoken text from other speakers without seeing their facial expressions.**

Our system uses two different APIs together to analyze verbal and nonverbal information. The first is Google's cloud-based speech-to-text API [2], which has a high speech recognition ability [7]. This API analyzes voice inputs from the microphone stream in real-time and returns the speech as text. Here, the recognition continues until silence is detected (as a pause), during which the result of the intermediate process is calculated in real-time. The second API is a voice emotion recognition API. To create a speaker independent system, the web Empath API [4] is used.

The text is displayed in real time, which includes the results being recognized. An emoji is inserted when silence is detected, and the Google speech-to-text API recognition is detached. For the audio "wav" file to be passed to the web Empath API, a recording process run in the background during the speech recognition process. The results are obtained by stopping the recording at the boundary between discrete speech recognition segments. Using Twemoji [12] which is emoji published by Twitter, the highest value is chosen from a set of "calmness," "anger," "joy," and "sorrow" emoji (Figure 1). The emoji is inserted when silence is detected, and the speech-to-text API recognition is detached. It does not show an emoji when an error message is returned from the web Empath API. Following the literature [1], we used a movie-style, two-line, at the bottom captioning style method, which exhibited the best usability scores in the study. The spoken text and emojis were rendered and transmitted to the GUI of the online meeting service through the use of virtual camera application (OBS-VirtualCam [10]).

Our system encountered a few technical difficulties: (1) under the noisy environment, the speech-to-text API recognition does not detach well, (2) the web Empath API cannot analyze the audio file which is longer than 5 seconds, and (3) using the window capturing on a virtual camera, some lag has occurred and the image quality has degraded.

## 3 PRELIMINARY USER STUDY

We conducted a preliminary experiment for measuring the effect of emoji augmented text representation on understanding voice-only online meetings for DHH users.

### 3.1 Participants

We recruited four participants (two females) ranging in age from 24 to 27 years old (M=25.25, SD=1.50). Each participant had binaural hearing loss. The hearing state of each participants is shown in Table 1. We confirmed that all participants had normal vision for daily activities and all participants were able to watch videos during our study.

### 3.2 Procedure

Participants watched three 30 seconds video clips and answered a survey with four questions. Our questionnaire is inspired by [5] (Figure 2). The videos included two students with no speaking disability using two captioning styles—text-only, and text + emoji-based emotion expression. The videos were the recordings of an online conversation from a third-party perspective. The video is in a gallery view, window of two people is in the position of left and right. The imagined situation was that of an online video chat being held with three people, including the participant, who was observing a conversation between the other two. To reduce any misunderstanding regarding the meaning of emoji face, the meaning of each emoji (Figure 1) was communicated to participants before the experiment started. It should be noted that we did not ask how DHH participants feel about the meaning of the emoji.

We have two conditions to test (text-only, text + emoji-based emotion expression). Participants experienced both situations in a random order and counterbalanced across participants. Concerned about the spread of COVID-19 infection, the experiment was conducted online. Participants watched the videos on their own devices. (P1 and P3 watched on a smartphone, P2 and P4 watched on a PC). All of our studies are approximately 15 minutes long. All of our participants are Japanese, thus all of the procedures in our experiment were in Japanese. The experiment design was approved by the university ethics committee.

### 3.3 Result

Figure 2 shows the results of the evaluation of the question items drawn using the Likert scale of five levels, and user preferences. The results were derived using the t-test and we did not find any

**Table 1: Hearing information of participants**

| Participant number | Hearing state | Age of diagnosis | Hearing-aid |
|---|---|---|---|
| P1 | binaural 60 dB | 3 years old | binaural |
| P2 | Sensorineural hearing loss (right 100 dB left 101 dB) | 2-3 years old | binaural |
| P3 | Sensorineural hearing loss (binaural 103 dB) | 1 years old | none |
| P4 | Sensorineural hearing loss (right 96 dB left 98 dB) | about 4-5 years old | binaural |

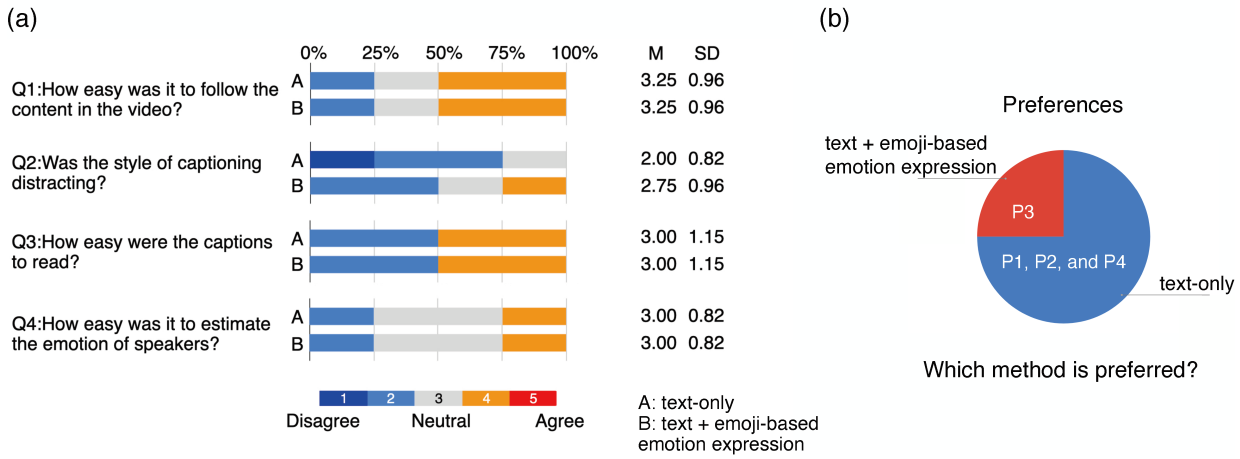


**Figure 2: The results of our study. (a) There is no significant difference between these two visualization methodologies (text-only, text + emoji-based emotion expression) among all 4 questions. (b) P3 prefers text + emoji-based emotion expression, and the others prefer text-only.**

significant difference between two conditions. In Q3, different viewing devices might have an effect, and we will keep this in mind in future experiments.

Through summarizing the subjective feedback, we mainly found three themes as stated below.

1. "The degree of hearing and Deaf culture": P3 prefers the method of text plus emoji-based emotion expression, and the others prefer only-text method. P3 is the only person who does not wear a hearing aid in this study (Table 1). And P1 said, "There is a discrepancy between the emotions inferred from the content of the sentence and the tone of the voice and the emotions expressed by the emoji." The degree of hearing could have an impact. Also, those who do not wear hearing aids are the Deaf who use sign language, and their requirements might be different [8].
2. "The problem caused by a misrecognition of the system and ethics": P1 said, "Emotions are easier to understand with emojis than with only text subtitles, but if there is a mistake in recognition, it is difficult to communicate well." Speech emotion recognition results are not always correct, so the system's emojis can mislead the user into misinterpreting other person's emotions. A misunderstanding of the other person's emotions can have a negative impact on the relationship. It is true that people can misinterpret others' emotions even when emotion recognition systems are not used, even in non DHH people. However, the impact and the ethical problem of support with emotion recognition tools on the misperception of the other person's emotions need to be discussed.
3. "More optimized GUI for online group meetings": P2 said, "Since the two screens are displayed left and right, I moved my viewpoints left and right each time the speaker switches to the other. I found it very hard. If two people are up and down the screen, I think it's still easier." It is also important perspective that optimization for multi-person online meetings. P2 also said, "Emojis are small and it was hard to distinguish between the calm and joy emoji." In this study, we added emoji in same size as text. However, in online meetings, the window size becomes smaller and smaller when the number of attendees is growing. This issue needs to be addressed.

## 4 DISCUSSION AND FUTURE WORK

We found that it is necessary to design emoji expression, GUI, and emotion recognition methodology to assist DHH people in group meetings.

A primary problem associated with using emoji in communications is that perception of emojis depend on culture [9] and varies from person to person [11]. A future study is required to understand the effect of cultural differences towards designing appropriate emoji expressions. A survey on how DHH people feel about each emoji is needed. Moreover, emotions are complex, and

we used only four types of single codes available to express most of them. In future, we plan to use more emoji expressions.

In this study, we analyzed emotions from audio recording. However, there are various methods for emotion recognition, and it is necessary to verify the effectiveness of other methods. Since we usually infer emotions from both verbal and non-verbal information, we have to examine the effectiveness of the combination of these methods (for e.g. [3] combines textual information and speech information).

## REFERENCES


[1] Larwan Berke, Khaled Albusays, Matthew Seita, and Matt Huenerfauth. 2019. Preferred Appearance of Captions Generated by Automatic Speech Recognition for Deaf and Hard-of-Hearing Viewers. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312921

[2] Google. 2020 (accessed July 31, 2020). *Cloud Speech-to-Text - Speech Recognition | Cloud Speech-to-Text | Google Cloud.* https://cloud.google.com/speech-to-text/?hl=en

[3] Jiaxiong Hu, Qianyao Xu, Limin Paul Fu, and Yingqing Xu. 2019. Emojilization: An Automated Method For Speech to Emoji-Labeled Text. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3313071

[4] Empath Inc. 2020 (accessed July 31, 2020). *Empath.* https://webempath.com/

[5] Sushant Kafle, Peter Yeung, and Matt Huenerfauth. 2019. Evaluating the Benefit of Highlighting Key Words in Captions for People Who Are Deaf or Hard of Hearing. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 43–55. https://doi.org/10.1145/3308561.3353781

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

[7] Apostolos Meliones and Cosmin Duta. 2019. SeeSpeech: An Android Application for the Hearing Impaired. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments* (Rhodes, Greece) *(PETRA'19)*. Association for Computing Machinery, New York, NY, USA, 509–516. https://doi.org/10.1145/3316782.3324013

[8] Carol Padden and Tom Humphries. 1989. Deaf in America: Voices from a culture. *Ear and Hearing* 10, 2 (1989), 139.

[9] Jaram Park, Vladimir Barash, Clay Fink, and Meeyoung Cha. 2013. Emoticon style: Interpreting differences in emoticons across cultures. In *Seventh International AAAI Conference on Weblogs and Social Media*.

[10] OBS Open Broadcaster Software. 2020 (accessed July 31, 2020). *OBS-VirtualCam 2.0.4.* https://obsproject.com/forum/resources/obs-virtualcam.539/

[11] Garreth W. Tigwell and David R. Flatla. 2016. Oh That's What You Meant! Reducing Emoji Misunderstanding. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (Florence, Italy) *(MobileHCI'16)*. Association for Computing Machinery, New York, NY, USA, 859–866. https://doi.org/10.1145/2957265.2961844

[12] Twitter. 2020 (accessed July 31, 2020). *Twemoji.* https://twemoji.twitter.com/