# "I Can't Talk Now": Speaking with Voice Output Communication Aid Using Text-to-Speech Synthesis During Multiparty Video Conference

Wooseok Kim
Department of Industrial Design, KAIST
Daejeon, Republic of Korea
ooooseok@kaist.ac.kr

Sangsu Lee
Department of Industrial Design, KAIST
Daejeon, Republic of Korea
sangsu.lee@kaist.ac.kr

## ABSTRACT

COVID-19 has resulted in the rapid popularization of video conferencing. A growing number of users have become obligated to find suitable places for video conferencing, but sometimes they inevitably participate in unsuitable conditions such as noisy or too silent public spaces. However, the video conference experience according to the environment users are in has not been sufficiently discussed. In particular, there is no conducted research on the occasions where video conferencing participants feel unable to speak with their voice due to spatial factors and how to address these situations. In this study, we propose a voice output communication aid (VOCA) for video conferencing which allows users to chat without making a sound. We made a technology probe and conducted a user test. Users who feel unable to speak orally could participate more actively with VOCA. Based on the results, we described the effects and potential of VOCA for video conferencing.

## CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **Empirical studies in collaborative and social computing**.

## KEYWORDS

Video Conference, Voice Output Communication Aid(VOCA), Means of communication

**ACM Reference Format:**
Wooseok Kim and Sangsu Lee. 2021. "I Can't Talk Now": Speaking with Voice Output Communication Aid Using Text-to-Speech Synthesis During Multiparty Video Conference. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '21 Extended Abstracts), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3411763.3451745

## 1 INTRODUCTION

Video conferencing has continuously evolved over the years as the hyperconnected society progresses. Significantly, the COVID-19 pandemic expanded the need for non-face-to-face communication around the world [6], and the number and range of users have increased dramatically. However, are there as many suitable rooms for video conferencing as there are users? With video conference becoming a commonplace, growing number of people cannot participate in video conferences in personal spaces for various reasons, such as economic or time constraints. However, most of the research related to video conferencing so far has been was conducted in the context of users in independent and controlled environments. There has not been enough discussion about how spatial factors affect users and the video conferencing experience. Kimura et al. explained that using smart devices via voice in public places can be annoying to other people and that stating private information aloud is risky [14]. This means that spatial factors may have a psychological effect on the user.

Current video conferencing services use noise cancelling [16] and background image processing [15, 22] to prevent users' surroundings from disturbing other participants in the conference. However, such surrounding factors not only interfere with other participants, they also interfere with users of that surrounding space. For example, if a café visited by a user is too noisy for video conferencing, the user may be unwilling to turn on the microphone and speak out loud. At a contrasting situation, when users are in a silent space with surrounding people, the users also feel restricted because they could possibly attract unwanted attention or distract other people in the area. In both cases, the user cannot participate properly, and the quality of the meetings may be deteriorated. We thought about how those users can participate and speak without making a sound.

In the current video conferencing interfaces, text chat can be used as an alternative. However, communicating via text chat in a video conference can distract attention of attendees and interfere with audio-visual interaction [7, 13]. Since the lacking ability to grab attention for extended period is a frequently mentioned disadvantage of video conferences, it is problematic for an attendee to continue texting during a multiparty video conferencing (MPVC) while others are speaking out loud. Geerts et al. [9] reported that text chat is slower than voice chat in information synchronization. The differences in information synchronization can confuse conversations; therefore, a user should avoid communicating only in text chat in MPVC sessions where voice chat mainly occurs.

Since we found out that text chat does not serve as a good alternative for speaking through existing research, we considered another non-verbal input to voice output interface for users who are hesitant to speak out loud with their voice. Augmentative and alternative communication (AAC) systems are defined as systems


*CHI '21 Extended Abstracts, May 8–13, 2021, Yokohama, Japan*

ACM ISBN 978-1-4503-8095-9/21/05…$15.00
https://doi.org/10.1145/3411763.3451745

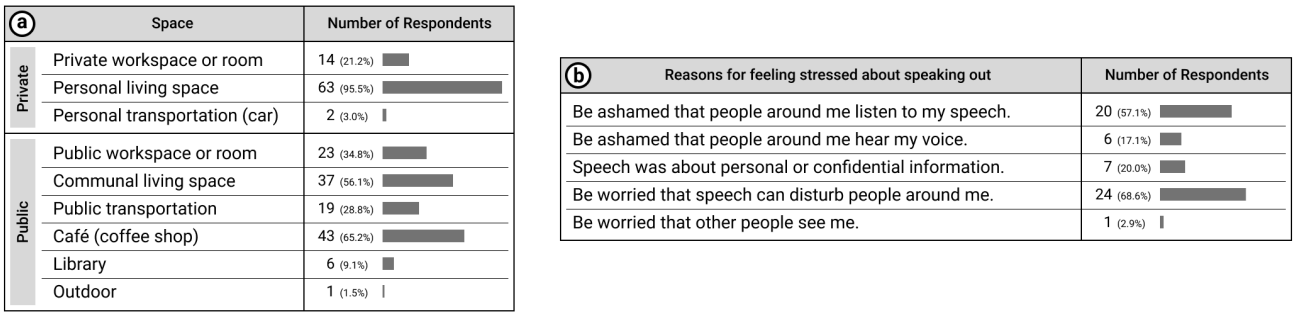


**Figure 1: Results of preliminary survey: (a) Spaces in which respondents had participated in video conferences; (b) Reasons for feeling stressed about speaking out when participating in a video conference in a space with others around.**

that help individuals with complex communicational needs [3]. A voice output communication aid (VOCA), also known as a speech generating device (SGD), is a kind of AAC system that provides natural speech production [4]. Researches on VOCA has typically been conducted only for people with disabilities. However, Beukelman et al.[3] described AAC as a system that supports effective communication for all people who have temporary or permanent communication difficulties. Pino et al. claimed that AAC could support communication among people of all ages, cultures, and languages in the "Design for All" approach [19]. Therefore, we regarded users who felt afraid to speak out due to surrounding factors as those experiencing temporary communication difficulties. In other words, able-bodied people are also eligible to be aided by AAC and VOCA.

VOCA can be produced in several ways, one of which is a silent speech interface (SSI). SSI refers to an interface that allows communication by generating speech without any sound input. It is being studied based on various technologies to provide advanced AAC for people with speech impairments [11]. However, these technologies require a lot of mechanical equipment or require a long time for calculations, and face many commercialization problems[5, 11, 14]. On the other hand, text-to-speech (TTS) has grown rapidly based on various uses [21], and many platforms support the TTS APIs. Nowadays, it is possible to synthesize voices similar to those of a real person and even express various emotions. Voice cloning technology, which analyzes and mimics a specific person's voice, has also been studied [23]. Fiannaca et al.[8] pointed out that various TTS technologies have not yet been applied to AAC. Therefore, we have determined that using TTS is now more appropriate than SSI.

In this context, we conducted a preliminary study to identify users' specific problems in situations where they feel unable to speak. A simple VOCA probe using TTS was developed based on the findings of a preliminary study. A user test was then conducted to compare the experiences according to communication means of the users who felt psychologically stressed. We observed whether a user could participate in MPVC smoothly when the user communicates through text chat and when the user speaks with VOCA. The result of this study identified the limitations of the current video conferencing interface. We found that VOCA not only helps hesitant users, but it can also be used for additional purposes(e.g., re-emphasizing particular contents, or energizing the conversation). Based on our results, we discussed the effectiveness of VOCA for video conferencing and derived 5 research agendas for a further design of VOCA.

## 2 PRELIMINARY STUDY

We conducted a brief survey of video conferencing experiences with 66 users in their 20s who have experienced video conferencing after the COVID-19 pandemic. The survey confirmed that users participated in video conferences in various spaces (Figure 1a). Fifty-two out of 66 respondents said they had participated in video conferences in a place where others were present, and 35 of them answered that they had felt constrained from speaking in those situations for various reasons (Figure 1b). Factors in people's surroundings influenced the frequency of utterances as well as their content. Fourteen of the 35 respondents who felt burdened replied they had experiences of not saying what they wanted to say. Sixteen respondents answered they had experiences in which they said only a part of what they wanted to say. Respondents said they have used text chat as an alternative when there were restrictions on speaking out, but 22 respondents said their text chats became buried because other attendees never checked it. Sixteen respondents said they had to mention their text chats again because others had not checked it. In conclusion, the current video conferencing interfaces do not effectively consider users who face restrictions to articulate vocally, and as a result, these users are unable to communicate appropriately in video meetings.

## 3 MAIN STUDY

In the preliminary study, we confirmed that users feel a psychological burden depending on their surrounding space and that this situation affects the conference's quality. It is a significant problem when a user wants to say something but decides not to say it, and some dialogues do not reach other participants. To check whether VOCA can help these situations, we developed a technology probe [12] and conducted a user test to observe the feasibility of the probe. We organized multiparty video conferences to further observe the situations found in the preliminary study. We formed 4 teams of 4 participants each, and 3 conferences per team were observed. We designated one participant per team as PX (participant X) and controlled PX's experimental conditions in the 3 conferences as follows:

- First conference: PX was loacted in a public space in which it was undesirable to speak out loud.
- Second conference: VOCA was used in the same space as the first conference.
- Third conference: VOCA was available, and PX could speak freely in a private conference room.

We asked each team to elect a leader among the participants to ensure that the 3 meetings were held in a consistent manner. PX was excluded as a leader candidate to accurately compare experiences. Therefore, each team consisted of one PX, one leader, and 2 participants(e.g., Team 1: $PX_1$[participant who controlled the environment], $PA_1$[team leader], $PB_1$, $PC_1$).

## 3.1 Technology Probe : TTS based VOCA for Video Conferencing

The VOCA technology probe was developed as a web app (Figure 2). In a preliminary study, respondents pointed out the long wait time between each delivered text as a drawback of text chats, so we introduced *Speak Immediately* mode on the user's *Input Page* (Figure 2a). With this mode activated, each time the user presses the space bar in the input field, the entered text up to that point are queued for speech synthesis. In other words, every typed word is queued immediately for speech synthesis and cannot be modified, which is similar to voice speech. Also, to take advantage of the text input, we made it possible to review the speech as a text format any time on the *Log Page* (Figure 2b). The difference in experience according to the types of voices was not the object of observation, so we provided one voice according to the participants' gender. The TTS API was provided by NAVER [17].

## 3.2 Participants

Not speaking out loud during video conferencing could require some tolerance from other participants, so we decided that it would be more appropriate to conduct the experiment with people who knew each other. We wanted to observe situations where a user felt constrained, so we carefully selected PXs. They were chosen among the candidates who responded in the preliminary study that they had previously felt constrained by speaking in an environment where other people were around. We were able to recruit 4 PXs from a nearby area, and we selected 12 participants among the volunteers who were acquaintances of PXs. We composed 4 teams consisting of a PX and 3 acquaintances of each PX. One team was a group already conducting regular video conferences, and one team was a group conducting regular face-to-face meetings. The other two teams were made up of colleagues taking classes together (undergraduate program team and master's program team). All 16 participants were students between the ages of 19 and 27 who were enrolled in an undergraduate program or a graduate program of the same department.

## 3.3 Procedure

Before the user test, we provided PXs guidance on how to use the VOCA probe and informed them of the experiment's overall flow. We also gave them instructions to let other participants know that they were in an environment where it was difficult to speak out with their voice during the first and second conferences. We instructed

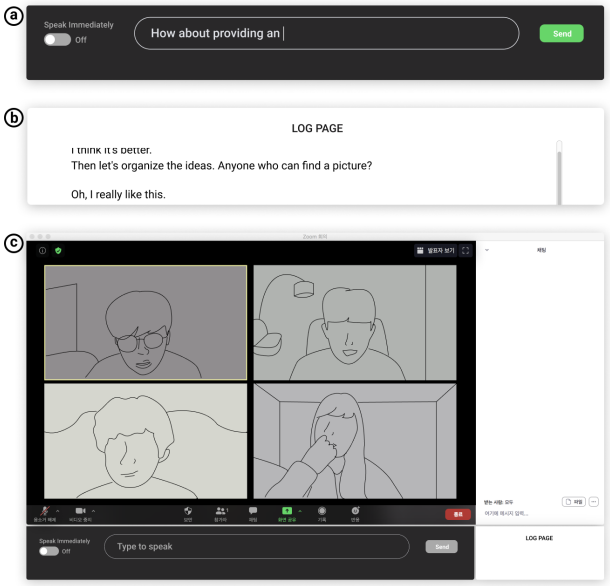


**Figure 2: The VOCA technology probe: (a) On the *Input Page*, PX can send the content to be spoken and use the *Speak Immediately* mode. (b) The *Log Page* outputs the received speech in both text and synthesized voice. (c) Screen of PX when using probe. The look and feel followed Zoom's design.**

PXs to hand over the leader to someone else if PXs were elected as the leader. We gave PXs some practice time to get used to VOCA and develop their own speech strategies. Other participants were asked to wear earphones and participate in the video conference in a quiet private room.

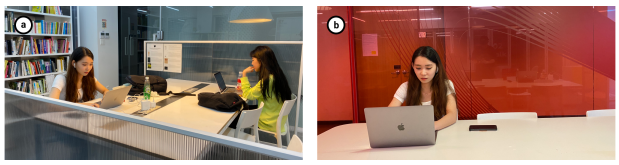


**Figure 3: PX's controlled environment: (a) An environment in which PX(the left person) feels psychological restraint to speak out for the first and second conferences (a quiet public space); (b) An environment in which PX could speak freely for the third conference (private conference room).**

The experiment consisted of an introduction, 3 conferences (up to 30 minutes each), and group interviews. The whole process was done using Zoom [24]. At the first conference, PXs were instructed not to speak with their voices because they were in a quiet library (Figure 3a). We secretly hired someone to study here in advance to make PXs feel a psychological burden. The researcher asked participants to elect a leader through referrals or through a simple game before starting the first meeting. At the first conference, all the teams came up with a discussion strategy. Since the leaders

were decided in advance, this process did not take a long time. After the first conference, we explained VOCA to participants and allowed PX to use VOCA. The participants followed a brief tutorial beforehand for prior familiarization. After the second conference, PX was moved to a private conference room (Figure 3b) for the third conference.

### 3.4 Task

We chose conference topics that all participants shared in common to prevent differences in participants' prior knowledge and interests from affecting the conference experience. The 3 topics covered different casual areas to minimize the reuse of the discussion points. The topics are presented in random order to reduce the presentation order effect.

- Planning an event for department students who had not met in person for a while due to COVID-19.
- Planning a new space in the department building for students who stay in the department building for an extended period.
- Planning a departmental program supporting students who live in the suburbs and want more design inspiration.

We asked for a simple result to observe the overall process. The work form and example were provided through Google Slides. The form requirements are as follows: 1) title of the idea, 2) background, 3) summary, and 4) details.

### 3.5 Data Collection and Analysis

Three types of data were collected: recorded video of the entire process, peer reviews of each conference, and qualitative interview data. We determined that the perceived contribution and participation felt by each team member could help interpret the overall experience, so we conducted peer reviews at the end of each meeting using a 7-points Likert scale. Since the meeting experiences were interactive, group interviews were conducted to share and discuss the experiences. While recording the interview contents, qualitative data were collected through a semi-structured interviews [2].

All interview data and meeting records were transcribed. We analyzed the collected data with a thematic coding approach [10]. We compared the characteristics of each utterance means based on the user's behavior and the content of the interview. As a result, a total of 5 themes of VOCA usage were derived.

## 4 FINDINGS

### 4.1 Comparison of participation and contribution scores

According to the results of the participants' peer evaluation, non-PX participants' scores did not show a significant difference at the three meetings. However, the PXs' scores showed a relatively noticeable difference (Figure 4). Assuming that the third conference—where everyone vocalized freely in an independent environments—is ideal, the scoreboard could mean that the context of second conference supported a more equitable participation than the context of first conference. In the interview, some participants mentioned that speaking via text chat does not seem like participating in a meeting. $PA_3$, who was the leader of Team 3, said that the $PX_1$ seemed to be almost non-existent at the first conference. Some participants also said that they generally believe that someone who speaks only by text chat where others speak in their own voice is not sincere. Unlike other PXs, Team 1's PX scored very well in all conferences while the score of PA1 - who was the leader of Team 1- gradually decreased. In regards to this outcome, $PC_1$ explained that as $PX_1$ became progressively more liberated to join the conversation, he showed a strong growth in leadership, which then resulted in a decreased participation and contribution from $PA_1$((the leader of the group).

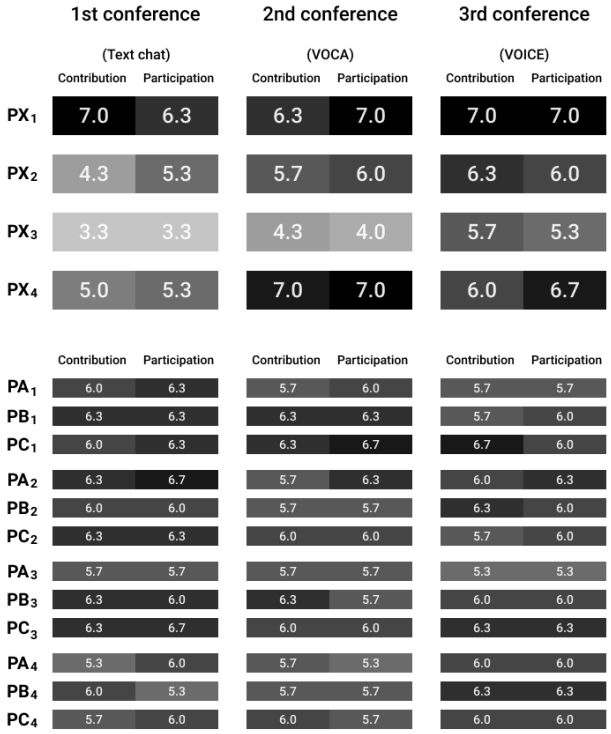


**Figure 4: Average of contribution and participation scores received from peer review.**

### 4.2 Characteristics and issues of speaking through VOCA using TTS

*4.2.1 Synchronous transmission of speech.* When the VOCA probe was available, all PXs actively used it. Users described that the greatest advantage of VOCA compared to text chat is its clear delivery.

> *VOCA was great when I couldn't speak. When I used the text chat, there was no way to know if a message was delivered. But VOCA definitely delivered my message by voice.* ($PX_2$)

*4.2.2 Voices overlap due to a lack of clues as to the beginning and end of the utterance.* Our VOCA program required about 1 second to synthesize and deliver a speech. Since the program had no way of

informing others that PX was typing an utterance, this 1-second delay increased the probability of overlapping voices. Another cause of the delay was PX's typing speed. This made it difficult for other participants to determine whether PX had finished speaking or not. Because of this problem, $PX_1$ had to say "Over!" after finishing his speech to clearly declare he was actually done speaking. The *Speak Immediately* mode was introduced for a faster speech communication, however this mode made it more difficult to determine the beginning and end of the utterances.

*4.2.3 Strong turn-taking.* There was no way to stop the audio from being output using our VOCA probe. Therefore, other participants were forced to hand over their turn to PX when their speech is overlapped with VOCA audio. Some participants ($PA_2$, $PC_2$, $PA_4$) stated that this was not annoying because it is a natural situation that occurs nomally between people. Still, other participants ($PA_1$, $PA_3$, $PC_4$) stated that it was uncomfortable.

*4.2.4 Absence of nuance.* Our probes always speak in the same tone and at the same speed. The synthetic voices often did not convey the user's intended nuances and interfered with others' understanding. This problem was very noticeable in empathetic utterances or simple responses because these types of speeches caught unnecessary attention.

> *It was regrettable that I couldn't control the strength of my speech because I am a person who enjoys sympathetic words.* ($PX_3$)

*4.2.5 Flawless voice like an announcer.* The synthesized voice was a clean third-party voice without any noise. There were neither echo nor reverberation phenomena. Some participants said they felt as though they were talking to a conversational agent such as Siri [1]. They answered it was awkward to exchange informal conversation with a flawless voice. Some others stated that VOCA's unfamiliar voice felt more important than it really was because it was like an announcer's voice. Thus, stating less important content with VOCA caused a hindrance. However, the $PX_3$ and $PX_4$ used VOCA to give a farewell or to energize the conference atmosphere in the last session, taking advantage of this flawless characteristic.

# 5 DISCUSSION AND IMPLICATIONS

This section discusses how existing video conferencing interfaces do not guarantee user participation in certain situations, and how VOCA can provide a solution for the users who are temporarily disabled to communicate vocally in those situations. Finally, based on our findings, we propose future discussions on VOCA design to support human-to-human communication in video conferencing.

## 5.1 Does the current video conferencing interface always support sufficient user participation?

Based on the preliminary study results, we confirmed that people participate in video conferencing in various environments, and they are sometimes psychologically constrained by environmental factors. Some users did not say what they wanted to say. Our user test has shown that users do not prefer present-day video conferencing text chat system when voice chat occurs simultaneously. In the first conference, participants who communicated vocally struggled to check text chats, so three of four PXs gave up on using Zoom's chat system and wrote down what they wanted to say on Google Slides. However, the problem was not resolved. Also, since text chat is delivered asynchronously, a process to synchronize all participants is essential for further conversation. Participants had to read the text chats repeatedly every time they wanted to reply to the chat. Some text chats were not delivered on time and cluttered the conversation. These issues imply that the text chat system did not fully support users' participation in meetings in specific environments.

## 5.2 Effectiveness and potential of VOCA in video conferencing

Our VOCA probe delivered the speaker's utterances immediately in both text and audio formats. Therefore, it was possible to deliver the utterances to all participants in the conversation simultaneously and reconfirm their content. This solved most of the problems with text chat. We also found that the synthesized, clean voice led to the utterances being recognized as crucial information. Using this characteristic, some participants used VOCA to refresh the atmosphere even when they were free to vocalize. This demonstrates the additional potential of VOCA, which can be used to more effectively transmit utterances with special-purpose.

## 5.3 Future research agenda of VOCA design for video conferencing

Our VOCA probe was effective for video conferencing users who felt restrained from speaking out due to factors in their surroundings. Therefore, we determined that it is worth evolving VOCA for video conferencing based on the issues we found, and we propose the following 5 future agendas for a better VOCA design.

*5.3.1 Adjustable speech rate.* In Team 1, where people's speech rate was very fast, VOCA's slow speech rate caused a hindrance. Therefore, a speech rate control interface is required. Users could control this, but we can also consider a way for the system to automatically recognize the average rate of other attendees and adjust its speech rate.

*5.3.2 Adjustable tone and nuances to suit the various purposes of speech.* The VOCA used in this study could not convey the nuances of speech, creating difficulties for other participants to understand the contents. In particular, VOCA's calm and refined voice was not suitable for empathetic or jocular speech. Therefore, there is a need for an interface that can adjust its tone according to the purpose of speech. The Voicesetting interface proposed by Fiannaca et al. [8] can be a solution. However, we have found that VOCA can be used for more varied purposes(e.g., managing meetings or energizing the atmosphere). Therefore, further discussion is needed for those various purposes.

*5.3.3 Signal user input status and beginnings and endings of utterances.* In voice utterance, turn-taking is performed simultaneously at the start of an utterance and includes various non-verbal means of communication. Since a video conferencing environment is inadequate for users to recognize these non-verbal methods, various studies have sought to improve this [18, 20]. However, speaking

with VOCA makes these nonverbal communications more difficult. Therefore, it is necessary to think about transmitting cues about the user's input status and the beginning and end of the speech.

*5.3.4 Cope with situations of voice overlap.* Providing cues about the beginning and end of speech is just a preventive measure against voice overlap. There is a need for more fundamental solutions to cope with the problem. We can consider interfaces that mimic the mutual feedback from human-to-human voice overlapping situations, such as pausing or cancelling the output speech.

*5.3.5 Feeling of a human.* Participants said that the synthesized voice felt like an announcer or a voice agent's voice, and its speech was perceived as special content. Some participants mentioned that the usability would be better if the VOCA spoke with the actual speaker's voice. So, we can investigate the usability of VOCA with a speech synthesizer trained with a user's vocal timbre or intonation. we can take advantage of voice cloning technology[23] that mimics a user's voice. It would also be worthwhile to see whether it is preferable to add natural echo and white noise to the synthesized speech for more natural and equal communication. However, it should be noted that listening to realistic sounds while watching a user's mouth not move can produce the uncanny valley effect.

## 6 CONCLUSION AND LIMITATIONS

In this research, we found that users' participation in video conferencing can be affected by external environmental conditions and found that VOCA can help users who face temporary speaking constraints. As video conferencing becomes more prevalent, more and more users will join video conferencing in situations in which it is difficult to speak out loud. Those situations may arise not only due to disabilities but also from temporary difficulties in users' external environments. Reduced speech can even happen if users catch a cold. Kimura et al. [14] stated that SSI could be used to control smart devices—that is, AAC or VOCA is not just for people with disabilities. If more AAC studies target both the disabled and the able-bodied, communication barriers between them will become more blurred. In that sense, we hope that this study contributes to broadening the scope of AAC research.

The limitation of this study is that we only observed for those in their 20s and not for various age groups, and the experiences may differ according to the user's age range. Another limitation is that we did not investigate the experience of more diverse types of meetings or greater numbers of participants. We will continue to conduct more studies under diversified conditions in the future, because we only observed situations in which users felt burdened by external factors.

## REFERENCES


[1] Apple. Accessed 2021. Siri does more than ever. Even before you ask. https://www.apple.com/siri/.

[2] K Louise Barriball and Alison While. 1994. Collecting data using a semi-structured interview: a discussion paper. *Journal of Advanced Nursing-Institutional Subscription* 19, 2 (1994), 328–335.

[3] David R Beukelman, Pat Mirenda, et al. 1998. *Augmentative and alternative communication.* Paul H. Brookes Baltimore.

[4] DOREEN BLISCHAK, LINDA LOMBARDINO, and ALICE DYSON. 2003. Use of speech-generating devices: In support of natural speech. *Augmentative and alternative communication* 19, 1 (2003), 29–35.

[5] Catarina Botelho, Lorenz Diener, Dennis Küster, Kevin Scheck, Shahin Amiriparian, Björn W Schuller, Tanja Schultz, Alberto Abad, and Isabel Trancoso. 2020. Toward silent paralinguistics: Speech-to-emg–retrieving articulatory muscle activity from speech. *Small* 61 (2020), 12.

[6] Kevin G Byrnes, Patrick A Kiely, Colum P Dunne, Kieran W McDermott, and John Calvin Coffey. 2021. Communication, collaboration and contagion:"Virtualisation" of anatomy during COVID-19. *Clinical Anatomy* 34, 1 (2021), 82–89.

[7] Christine Develotte. 2012. L'analyse des corpus multimodaux en ligne: état des lieux et perspectives. In *SHS Web of Conferences*, Vol. 1. EDP Sciences, 509–525.

[8] Alexander J. Fiannaca, Ann Paradiso, Jon Campbell, and Meredith Ringel Morris. 2018. Voicesetting: Voice Authoring UIs for Improved Expressivity in Augmentative Communication. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173857

[9] David Geerts, Ishan Vaishnavi, Rufael Mekuria, Oskar van Deventer, and Pablo Cesar. 2011. Are We in Sync? Synchronization Requirements for Watching Online Video Together.. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 311–314. https://doi.org/10.1145/1978942.1978986

[10] Graham R Gibbs. 2007. Thematic coding and categorizing. *Analyzing qualitative data* 703 (2007), 38–56.

[11] J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. M. Martín Doñas, J. L. Pérez-Córdoba, and A. M. Gomez. 2020. Silent Speech Interfaces for Speech Restoration: A Review. *IEEE Access* 8 (2020), 177995–178021. https://doi.org/10.1109/ACCESS.2020.3026579

[12] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology Probes: Inspiring Design for and with Families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 17–24. https://doi.org/10.1145/642611.642616

[13] Richard Kern and Christine Develotte. 2018. *Screens and scenes: Multimodal communication in online intercultural encounters.* Routledge.

[14] Naoki Kimura, Michinari Kono, and Jun Rekimoto. 2019. SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300376

[15] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother. 2005. Bi-layer segmentation of binocular stereo video. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 2. 407–414 vol. 2. https://doi.org/10.1109/CVPR.2005.91

[16] S. M. Kuo, Y. C. Huang, and Zhibing Pan. 1995. Acoustic noise and echo cancellation microphone system for videoconferencing. *IEEE Transactions on Consumer Electronics* 41, 4 (1995), 1150–1158. https://doi.org/10.1109/30.477235

[17] Naver. Accessed 2020. Clova Speech Synthesis API. https://developers.naver.com/docs/clova/api/.

[18] David Nguyen and John Canny. 2005. MultiView: Spatially Faithful Group Video Conferencing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA) *(CHI '05)*. Association for Computing Machinery, New York, NY, USA, 799–808. https://doi.org/10.1145/1054972.1055084

[19] Alexandros Pino and Georgios Kouroupetroglou. 2010. ITHACA: An open source framework for building component-based augmentative and alternative communication applications. *ACM Transactions on Accessible Computing (TACCESS)* 2, 4 (2010), 1–30.

[20] Md Tahsin Tausif, RJ Weaver, and Sang Won Lee. 2020. Towards Enabling Eye Contact and Perspective Control in Video Conference. In *Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 96–98.

[21] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *arXiv preprint arXiv:1803.09017* (2018).

[22] C. Zhang, Y. Rui, and L. He. 2006. Light Weight Background Blurring for Video Conferencing Applications. In *2006 International Conference on Image Processing*. 481–484. https://doi.org/10.1109/ICIP.2006.312498

[23] Yu Zhang, Ron J Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, RJ Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran. 2019. Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. *arXiv preprint arXiv:1907.04448* (2019).

[24] Zoom. Accessed 2021. Video conferencing. https://zoom.us/.