SpeechBubbles: Enhancing Captioning Experiences for Deaf and Hard-of-Hearing People in Group Conversations

Yi-Hao Peng [1], **Ming-Wei Hsu** [2], **Paul Taele** [3], **Ting-Yu Lin** [1], **Po-En Lai** [2],
Leon Hsu [2], **Tzu-chuan Chen** [2], **Te-Yen Wu** [2], **Yu-An Chen** [2], **Hsien-Hui Tang** [4], **Mike Y. Chen** [5]
National Taiwan University [1,2,5], Texas A&M University [3],
National Taiwan University of Science and Technology [4]
[1] {B03902097,B03902061}@ntu.edu.tw,
[2] {chad1023,mark840308,leonorz123,aldrich1221,teyanwu,ryan149347}@gmail.com,
[3] ptaele@cse.tamu.edu, [4] drhhtang@mail.ntust.edu.tw, [5] mikechen@csie.ntu.edu.tw

ABSTRACT
Deaf and hard-of-hearing (DHH) individuals encounter difficulties when engaged in group conversations with hearing individuals, due to factors such as simultaneous utterances from multiple speakers and speakers whom may be potentially out of view. We interviewed and co-designed with eight DHH participants to address the following challenges: 1) associating utterances with speakers, 2) ordering utterances from different speakers, 3) displaying optimal content length, and 4) visualizing utterances from out-of-view speakers. We evaluated multiple designs for each of the four challenges through a user study with twelve DHH participants. Our study results showed that participants significantly preferred speech bubble visualizations over traditional captions. These design preferences guided our development of SpeechBubbles, a real-time speech recognition interface prototype on an augmented reality head-mounted display. From our evaluations, we further demonstrated that DHH participants preferred our prototype over traditional captions for group conversations.

ACM Classification Keywords
K.4.2. Information Interfaces and Presentation: Assistive technologies for persons with disabilities; H.5.1. Information Interfaces and Presentation: Artificial, augmented, and virtual realities

Author Keywords
Accessibility, text bubbles, word balloons, deaf and hard of hearing, closed captions, augmented reality, hololens.

INTRODUCTION
For members of the deaf or hard of hearing (DHH) community, comprehending speech from face-to-face group conversations with hearing individuals can present challenges. The reason is while speech plays an important role in face-to-face communications, it becomes reduced or lacking as a communications channel between hearing and DHH individuals [16]. As a result, the DHH community has turned to alternative approaches for better comprehending speech from hearing individuals such as writing and keying, gesturing and signing, relying on human interpreters, and interpreting lip movements (i.e., speechreading) [9, 12, 15, 31]. However, these approaches can pose additional constraints that are not as seamless as speech comprehension is for hearing individuals.

Figure 1. A proof-of-concept demonstration of SpeechBubbles. The user (right) equipped with a Microsoft HoloLens views speech bubbles adjacent to two speakers (left and middle).

With their growing ubiquity and reliability, smartglasses have strong potential to address DHH individuals' existing challenges in comprehending speech from hearing users in face-to-face group conversations. Efforts that may benefit this technology include leveraging current advances in speech recognition that translate conversational speech into readable captions and subtitles [5, 6, 21, 23, 25], and specializing computing assistive interfaces for DHH users that better streamline certain conversation interactions (e.g., Ava [1], UNI [3]). However, the former relies on more deliberately-controlled filmed video assumptions in media that do not closely capture users' actual freeform-viewing perspectives,

CHI 2018, April 21–26, 2018, Montréal, QC, Canada.

https://doi.org/10.1145/3173574.3173867

while the latter disregards valuable nonverbal signals or does not target DHH users.

We propose SpeechBubbles, a smartglasses-driven assistive interface for improving DHH individuals' experiences of comprehending speech from hearing individuals in face-to-face group conversations. Our work uses a Microsoft HoloLens head-mounted display (HMD) and a microphone array setup, and displays graphical word balloons—called *speech bubbles*—in an augmented reality (AR) interface. Our aim is to intuitively display and optimally place speech content into dynamic real-time captions. The contributions of our system first include addressing the limitations of existing static and dynamic caption and subtitle systems that are limited to preset filmed videos. Our system also addresses the limitations of existing assistive interfaces for DHH users that do not closely capture seamless face-to-face group conversation experiences with hearing individuals. We explored designs for different display methods, caption presentation content, and graphical out-of-view utterance hint cues, and discovered optimal and preferred visualization options for our desired interface.

RELATED WORK

Our proposed smartglasses-driven interface relies on a novel captioning visualization for DHH individuals in group communications. Therefore, we discuss related works in communication assistive systems, video captioning interfaces, and text bubble visualizations.

Communication Assistive Systems for DHH Users

Technologies for supporting the communication and information needs of DHH individuals have been a focus of researchers since the early 1970s [16]. Gugenheimer et al. [12] proposed a system that investigated the impact of computing devices in real-time conversation settings to enhance face-to-face communication experiences between DHH and hearing individuals, but heavily relied on a human sign language interpreter to support the communication between hearing and DHH individuals. Kawas et al. [22] proposed an approach that improved speech-to-text caption readability for DHH students from speech of their classroom hearing peers, but it is constrained from requiring prior preparation and setup before captioning sessions and is also vulnerable to speech accuracy and latency issues. Kushalnagar et al. [26] proposed a speech-to-text solution that tracked the locations of classroom presenters to more intelligently place captions at their locations, but primarily relies on uni-directional communication assumptions from a speaker that may not be as applicable for the greater immediacy and spontaneous nature of group conversation assumptions in more varied environments.

Assistive systems for DHH users have also leveraged directional microphones for broadening sound awareness such as speech for DHH users, which is helpful for face-to-face group conversations where originating speech can occur outside the direct range of a DHH user's view. McCreery et al. [28] proposed such a system for DHH children of school-age with hearing aids. However, the system focused more on reducing the impact of digital noise reduction, and target users expressed difficulties in hearing sounds originating at their periphery and from behind. Other research works have further proposed the inclusion of informative visual cues in assistive systems, in order to better augment sound awareness for DHH individuals [19, 27, 30]. Jain et al. [19] proposed a head-mounted display that provides optimal visual notifications for DHH users, which involves visual indicators for better identifying the spatial location of sounds and for better presenting that information in real-time. Our proposed system aims to expand on the efforts from that particular work, from the context of more informed spatial sound awareness to the context of more improved speech conversation comprehension.

Video Captioning Interfaces

Conventional captioning places text scripts over video in static locations, and have generally consisted of different visualization forms such as scrolling upward, appearing spontaneously, painting over, and appearing cinematic [18]. This visualization benefits DHH individuals for viewing videos without audio. However, prior studies reported that when DHH individuals used conventional captions, their viewing experience provided little significant information due to needing to quickly associate caption scripts with on-screen content [13, 14]. As a result, researchers have investigated dynamic captioning—or captions over video that vary in position [5, 18]—that more intelligently alleviate information visualization challenges inherent in conventional subtitles. These automated dynamic captioning systems though rely on displaying captions for videos from existing media such as movies or television shows with more director-controlled scenes, while using existing text scripts from these media to drive their caption placement. Such assumptions limit their usability for live face-to-face group conversations with users' more freeform perspectives and greater spontaneous conversational speech.

Text Bubble Visualizations

Word balloons—alternatively, speech bubbles or text bubbles—are visualizations that depict speech in comics since the late 19th century [24]. As "one of the most distinctive and readily recognizable elements in the comic medium", word balloons have consistently represented dialogue in stories represented in this medium [7, 24]. This visualization has similarly been adapted in other media for novel forms of communication. Early work by Kurlander et al. [24] proposed a graphical chat program that adapted word balloons with general rules of comic panel composition automatically onto users' avatars. Chun et al. [29] focused on visual optimal placement of word balloons for conveying text onto static images. Word balloons as a visualization design of choice has also had applications for representing speech in assistive systems for DHH users. Piper et al. [29] displayed word balloons from microphone speech or keyboard typing onto a multitouch tabletop display to better facilitate communication between hearing doctors and DHH patients. The use of word balloons in comics and its successful adoption into computer-mediated communications and DHH assistive systems motivated us

to adopt this visualization into our proposed interface's visualization cues for DHH individuals in face-to-face group conversations.

DESIGN

The core idea of our design was to provide an overall desirable visualization for DHH users, so that they may become more seamlessly engaged in speech-based group communications with hearing individuals whom lack prior sign language knowledge. With captioning systems serving as a conventional visualization solution for supporting DHH individuals in such conversational situations [25], and with the expanding ubiquity of smartglasses and HMDs for AR interactions, we sought to design a system that explored the open design space of AR environments to better optimize captioning visualizations. To accomplish this, we first conducted semi-structured interviews to investigate issues that DHH individuals encountered for group conversations. Secondly, we went through co-design processes that allowed DHH individuals to actively participate in our design flow by sketching out their ideal interface. After problem-exploring and idea-purposing, we configured visualization details (e.g., font size and style) that were adopted from prior works and user preferences that were obtained from the mentioned interviews and processes. Finally, we identified and categorized the participants' design challenges that arose from any issues that they encountered, and provided corresponding visualization designs that better optimized DHH individuals' experiences for group conversations.

Semi-structured Interviews and Co-design Process

In order to more deeply understand issues that DHH individuals encountered in group conversations with hearing individuals from existing captioning resources, we conducted a semi-structured interview and participatory design (co-design) with 8 DHH participants. The eight participants (3 female) ranged in ages from 18 to 39 years (M=24, SD=6.70), and had degrees of hearing loss: 1 minor, 1 moderate, 4 severe, and 2 extremely severe. Since all participants had hearing loss from both ears, all reported having used hearing aids and three reported using cochlear implants afterwards. All participants stated Chinese Mandarin as their native language. We based our semi-structured interview format from prior work by Hong et al. [18]. For each participant, we asked for the difficulties that they encountered when engaged in group conversation, how they acclimated to these challenges, and what ideal visualization solution that they would propose for an ideal captioning system.

Challenges

All participants reported experiencing communication problems in group conversations even when wearing hearing devices. For example, *"I need to read lips for better comprehension and couldn't clearly hear a lot of people at the same time"* (P3), and *"I hate group conversation because I can't clearly hear what others are saying, and would be left hanging as a result."* (P2). Among all the stated difficulties that they faced, we discovered that most participants expressed their concerns in the following categories:

- **Direction of sounds:** Most participants (7) expressed concerns when it came to identifying the direction of sounds. Without reading lips, they had trouble determining the origin of the sounds. Moreover, a deaf person could not hear sounds coming from behind them, because modern hearing aids are often unable to fully retrieve sounds from there. *"Both hearing aids and cochlear implants are more effective for sounds from the front, so I can't hear clearly if someone wasn't talking in front of me."* (P3).

- **Multiple sound sources:** The DHH participants had trouble distinguishing which speaker was talking when several people spoke simultaneously. *"If someone was talking and others chimed in when that person hadn't finished yet, I wouldn't understand what they said."* (P5).

- **Noisy environment:** Five of the participants reported that they would have had even worse experiences for having conversations in a noisy environment. *"Every time we have group discussions in the classroom, I often can't grasp the topic very well due to noise from the other groups."* (P1).

Accommodations

We also received remarks from participants on accommodations that they took, in order to better alleviate issues that they faced in comprehending group conversations.

- **Adaptive:** Most participants (7) reported that they would use verbal accommodations such as asking others to repeat themselves, asking others to speak more slowly or loudly, and requesting someone to serve as a temporary interpreter. Two participants mentioned that captioning services were helpful in group conversations. For example, *"Group conversations were actually a nightmare for me before I started using real-time captioning services provided by the school."* (P4). Although current captioning service helped significantly, participants also mentioned that it somehow limited their autonomy within group conversations. *"Captioning services weren't suitable for every situation such as group activities."* (P4). *"[Captioning services] actually diverted my attention to the display and would be distracting to the conversation."* (P6).

- **Maladaptive:** Half the participants (4) reported that they had taken compensated methods such as requesting someone to summarize what speakers said at the end of the conversation, or that they would sometimes give up trying to understand what was said in the group entirely. *"I pretend to listen with keen interest and would keep silent off to the side."* (P2). Two participants even had a tendency to avoid group conversations. For example, *"I would prefer to discuss online, since that way I don't need to expend a lot of effort trying to understand the conversation."* (P2). According to their feedback, participants seemed to imply that group conversations remained a large challenge for at least some DHH individuals.

Ideal design for real-time captions

After the interviews, we asked participants to describe and sketch their ideal design for an AR interface over an image of a representative group conversation scenario (Figure 2).

- **Placement of captions:** Most participants (7) responded with designs that associated captions to the speaker by their caption location. While two participants placed the caption close to the speakers, five participants drew arrows or bubble-like designs that pointed towards the speaker. The remaining participant remarked about displaying captions in the center of the view and labeling the speaker's name before each utterance. *"I wanted to put all the captions with the speakers' name in the center and display some other information on both sides."* (P2). However, all participants agreed that the visualization should not cover people's faces and speaker's facial expressions for eye contact purposes: *"It's intuitive to put the captions above or below the speaker's face without covering their face, representing the words said from the speaker.* (P8).

- **Advanced feature:** Participants also proposed additional features for their ideal designs. Four participants were concerned about speech originating from their left, right, or rear. As a result, two participants suggested using a hint icon or arrow on the peripheral region as a visual cue. Two other participants preferred displaying captions with the out-of-view speakers' photo on the bottom of the text. In addition, some participants mentioned that visualizing with animated text could help imply the latest utterance in the conversation. *"I think you need some animations in text such as color changing to hint at which sentence we want users to see."* (P2).

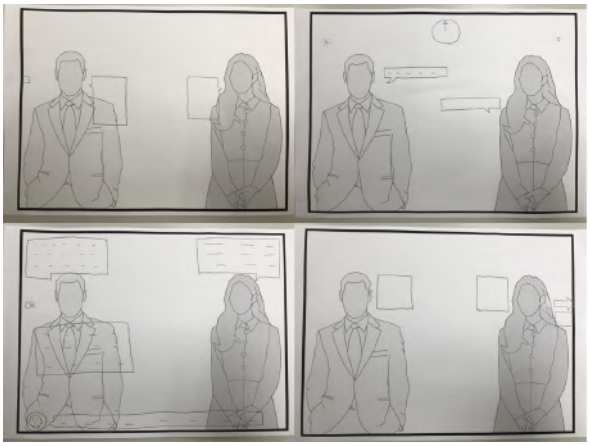

Figure 2. Participants' drawn ideal designs during the co-design process.

Visual Cue Details

After collecting feedback from the DHH participants, we investigated replacing traditional caption visualizations with a bubble-like visualization—based on insights derived from user preferences and prior works—as detailed in the following.

Text bubble behavior

We derived the design of a rectangular text bubble with rounded corner as our design because it is capable of containing the maximum amount of phrases. According to conventional speech bubbles utilized in the comic book medium, we chose a white background with black text as the text display style [29], and set this background with 50% transparency beneath the AR environment. Since Mandarin is the primary language of fluency of the participants in our convenience sampling and one of the most widely-read language in the world, we designed our captions for Mandarin. We also selected Noto Sans Mono CJK TC [2] as our font style, since it was designed for scripts in languages such as Mandarin. We limit the number of characters in a single line to 12, since exceeding this selected limit would exceed a typical individual's average reading speed [32, 33], and display the caption lines for 3 to 5 seconds as suggested by [10].

Sound location awareness

From our co-design task, we mapped the speech bubble's edge with the speaker's location, and also placed the bubble adjacent to the speaker's face based on [18]. For sound originating from outside the user's view, we implemented a bubble-like hint at the user's peripheral region as a visual cue [11, 29]. We also chose an egocentric perspective to display this visual information, since prior work showed that it was easier to locate direction compared to an exocentric perspective [19]. The bubble-like visualization was able to demonstrate both direction and captioning simultaneously, which opened up more design possibilities for indicating out-of-view speakers.

Design Goal and Proposal

Since we desire having DHH individuals better engage with hearing individuals in group conversations, we based our design purpose on issues that the DHH participants expressed facing in these situations, and on challenges that they brought up while drawing proposed captioning solutions during the co-design process. From the steps conducted in our study, we first define and categorize these issues into five main design challenges, and then propose recommendations to directly address these difficulties (Figure 3).

- **Speaker Association:**
  - When talking in group conversation, DHH individuals may not identify who is immediately talking, especially when several people are talking simultaneously.
  - We developed a speech bubble-like visualization which—compared with traditional captioning—allows DHH users to more easily recognize the source of the speech.
- **Amount of content:**
  - The appropriate amount of the displayed caption for each utterance needs to be determined to present adequate amount of content.
  - We provided different speech bubble sizes with scrolling text—from single line to multiple lines—for determining the most comfortable length to present the appropriate number of content on the HMD.
- **Order of utterances:**
  - When displaying conversations with captions, it may be very difficult for people to distinguish between the sequence of utterances.

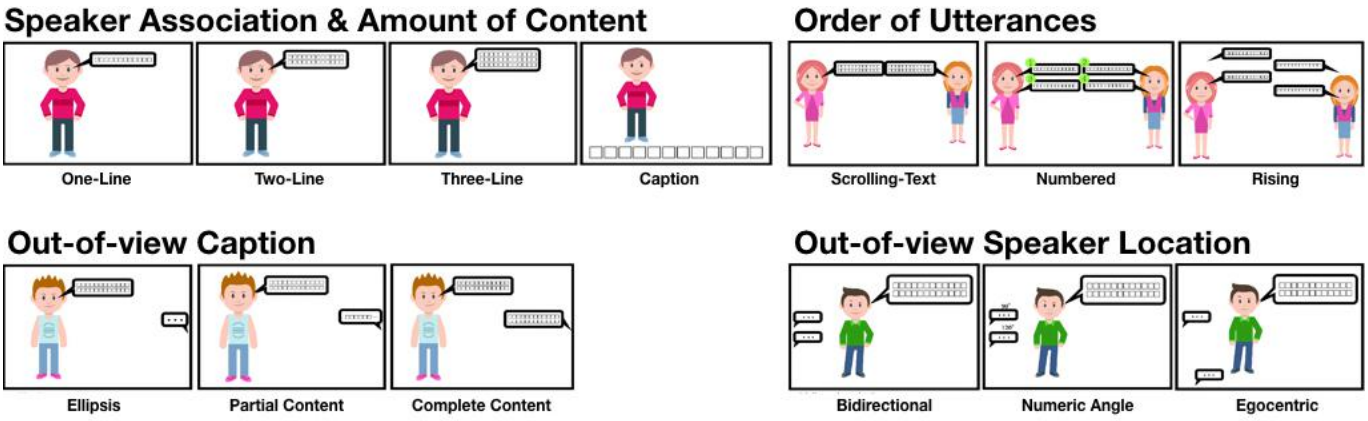

Figure 3. Designs explored for each of the five design dimensions: 1) speaker association 2) amount of content, 3) order of utterances, 4) out-of-view caption, and 5) out-of-view speaker's location.

  - In addition to the normal static bubble with scrolling texts, we provided an enumerated bubble that numbered each bubble according to its temporal order in the conversation, and a rising bubble that moves directionally upward after displaying on the user's view and chooses the dialogue's order by the bubble's height.

- **Out-of-view caption:**
  - Being aware of information that is communicated outside of the peripheral view is still a major challenge for DHH individuals.
  - We offered three design types for providing hints of words uttered by speakers: 1) ellipses for indicating when someone is talking, 2) partial content of what the speaker said, and 3) complete content of what the speaker said.

- **Out-of-view speaker's location:**
  - Distinguishing out-of-view speaker's location is also a problem when it comes to communication occurring outside of the DHH individual's view.
  - We provided three types of designs for hinting at the relative locations of out-of-view speakers: 1) *bidirectional* for determining the direction of the out-of-view speakers by only indicating whether the speaker is located to the left or right of the user, 2) *egocentric* for converting the plane parallel to the user into the plane of the real world where the user is standing and where the bottom of the view denotes positioning behind the user, and 3) *numeric angle* for presenting the angle between the speakers and the user from 0° to 180°.

USER STUDY

In order to understand DHH users' preferences for the different types of visualization designs, we conducted a series of user studies that evaluated our previously-described design efforts.

Study Method and Procedure

To evaluate the various speech bubble designs, we adopted the same design and evaluation methodologies as [19]. We chose to show videos on a 24-inch LCD monitor instead of the HoloLens, since it better reflected the field of view of future AR devices and so that the findings would better benefit the research community. The reason is that the current HoloLens has an extremely narrow 35° diagonal field of view (FOV), compared to second-generation headsets (e.g., the Meta 2) that can provide up to 90° of diagonal FOV and to human vision that provide 210° horizontal FOV.

Study Participants

We recruited 12 participants (3 female) whose ages ranged from 20 to 55 years (mean=28.42, SD=18.11) from a university campus and via the internet. Our participants had varying degrees of hearing loss: minor (1), moderate (4), severe (6), and extremely severe (1). All of them have binaural hearing loss. The participants also experienced hearing loss at different stages of their lives: diagnosed congenital hearing loss at birth (4), early childhood hearing loss (6), and acquired hearing loss (2). In regards to employed assistive devices, all our participants used digital hearing aids such as cochlear implants (5) and personal FM receivers (3). Furthermore, since our study included watching videos, we confirmed that all participants had non-disabled vision for daily activities: without glasses (10) and with glasses only when needed due to weak amblyopia (2). Finally, most participants (10) employed lipreading during conversation.

Study Procedure

We instructed the participants to perform a series of video-watching tasks to correspond to different visual designs. During the study, participants saw the designs divided into four parts (Figure 3). More specifically, we provided two conversation scenarios for the first two parts of our study: with overlapped utterances that are similar to daily conversations or without.

In total, participants performed $[(4 + 3) \times 2] + (3 + 3) = 20$ video-watching tasks. We conducted our study using Latin-

square designs to counterbalance, alternating between visual cues for the conversations. After participants completed each video-watching task, we instructed them to rank all three main design aspects from a 5-point Likert scale. In addition, we used indices from [8, 25] to evaluate our design.

- **Comprehension:** The level of understanding for both the content of the conversation and the meaning of the hint in the video.
- **Comfort:** The level of comfort if they were engaged in conversation under the particular design.
- **Preference:** The ranking of design in the same dimension under different conversational scenarios. If there were *n* items for comparison, design scores would range from *0* to *n-1* points.

With 12 participants and 20 video-watching tasks, participants performed a total of $12 \times 20 = 240$ video-watching tasks.

Study Results

From our user study, we describe our study results in terms of participants' main thoughts of the presented visualization, expanded discussions on our study's quantitative and qualitative feedback such as ranking the different visualization options for each design dimension, and suggestions for potential improvements.

Participants' thoughts on the visualization

All participants (12) found our bubble-like visualization very useful for AR devices when conversing with others in a group. Most participants (11) expressed this preference because they considered the visual information with graphic symbols easier to interpret than with sound cues. Moreover, half the participants (6) found it useful for the design to directly indicate the voice source. *"The bubble-like design provided distinct signs of the speakers so that I could quickly distinguish between who was talking at that moment. This visualization would be even more helpful when more than one person was talking simultaneously."* (P7). However, two participants (P9, P10) perceived potential problems with our visual cues. *"The main problem is that it might be a little bit annoying for me to chat with others alongside a lot of bubbles due to [the bubbles] interfering with my view.* (P10). Although P10 had strong concerns about the visual design, the participant still agreed with its helpfulness.

	Nonoverlap		
	Comprehension	Comfort	Rank
Static Caption (c)	3.75	2.58	4
One-Line (1)	4.00	3.83	3
Two-Line (2)	4.83	4.17	1
Three-Line (3)	4.58	4.17	2
	Overlap		
	Comprehension	Comfort	Rank
Static Caption (c)	3.58	2.75	4
One-Line (1)	3.58	3.58	3
Two-Line (2)	4.17	4.17	1
Three-Line (3)	4.08	4.08	2

Table 1. The summary of our 5-point Likert scale evaluation for our user study and user preference rank, in terms of amount of content and comparison with traditional captioning visualizations.

Associating speech utterances with speakers

The results for identifying speakers were derived from the first part of our user study. The data in Table 1 show that regardless of conversational scenario, the bubble-like visualization was more greatly preferred than traditional captioning using the Friedman test with the Wilcoxon-Nemenyi-McDonald-Thompson post-hoc test (Friedman & WNMT tests) [17]: $p_{c1} = 0.002$, $p_{c2} < 0.001$, $p_{c3} < 0.001$. In general, the participants pointed out that the main issue with traditional captions was the lack of information associating dialogue with speakers. *"It was very hard for us to identify who was talking only by watching the caption. In contrast, the bubble-like design could quickly indicate the speaker so that it was better for us to understand the dialogue."* (P8).

Appropriate amount of content to display

We also derived the optimal caption length for the user's view from the first part of our user study. In Table 1, we did not observe statistical significance in terms of preference for each design. Meanwhile, the multi-line bubble was more preferred than the single-line bubble on average but was not statistically significant using the Friedman & WNMT tests: $p_{12} = 0.0569$, $p_{13} = 0.1030$, $\alpha_{adj} = 0.0167$. One possible reason may be related to presentation of the chat history, where one participant mentioned the benefit of multi-line bubbles for providing chat history content. *"I think two-line bubbles and three-line bubbles are way better than single-line ones. Since the bubble is bigger, the past [utterance] spoken from each speaker would stay in the bubble for awhile, which allowed us to more easily trace back the dialogue that we missed."* (P3). Furthermore, some participants reported that multiple-lines bubbles were capable of completely presenting whole sentences. *"Since multi-line bubbles contained more space for characters, they could present longer sentences in several lines that provide complete conversation content."* (P12).

	Nonoverlap		
	Comprehension	Comfort	Rank
Scrolling-Text (s)	4.25	3.92	2
Numbered (n)	4.75	3.83	1
Rising (r)	4.5	3.33	3
	Overlap		
	Comprehension	Comfort	Rank
Scrolling-Text (s)	4.50	4.33	2
Numbered (n)	4.08	3.25	3
Rising (r)	4.58	3.92	1

Table 2. The summary of our 5-point Likert scale evaluation of our user study and user preference rank, in terms of the display methods for ordering.

Displaying order of the conversation

Table 2 shows the results for representing the order of dialogue with different display methods, which indicated differences between user preferences for conversations with and without overlapped utterances. For conversations with non-overlapped utterances, numbered bubbles as a display method scored the highest on average as a preference, while the result was nearly statistically significant on the Friedman & WNMT tests: $p = 0.0569$. However, for conversation views with overlapped utterances, the numbered bubble design would become too visually complex for participants due to

the additional enumerated information. On the other hand, participants considered the rising bubble design as the most comprehensible and intuitive design in comparison. The design received the highest average ranking score from users, although it was not statistically significant on the Friedman & WNMT tests: $p_{rn} = 0.0072$, $p_{rs} = 0.1103$). Some participants expressed how the rising bubble design helped in conversations occurring at a rapid pace: *"The rising bubble was somehow useless when we talked at normal speed. Interestingly, the bubble turned out to be of great help in letting us keep up with faster dialogue due to intuitive representing the order of the conversation."* (P7). As for improvements, one participant offered a suggestion for the visual design in cases of new people joining an existing group conversation, where the bubble-like design assigned to different people could be distracting. *"I suggest the chatroom-like display approach for concentrating information in a smaller range of view. Similar with a chatroom for multiple people, there should be a name or some kind of identification for every word spoken by the speaker."* (P3). Another participant offered color suggestions for the text and its background in the design. *"I think that I would choose black for the text background instead of white, since white is uncomfortable for me to stare at for a long time."* (P9).

Intuitive hint cues for out-of-view dialogue

Taking into account utterances from out-of-view speakers, we discuss two different hint aspects. For indicating out-of-view utterances, Table 3 shows that people preferred hints that were displayed completely or partially compared to symbolic ellipses using the Friedman & WNMT tests: $p_{ep} = 0.0048$, $p_{ec} = 0.0003$). *"I think that showing words in the hint bubble let me realize that there might be communication taking place outside my view. On the contrary, if there were only ellipses, it would be hard to associate them with the dialogue."* (P8). However, some participants considered hint as ellipses as the best design in comparison. *"I think that hint bubble with ellipses make me feel that I'm playing a game and will soon encounter a non-player character (NPC), which was so intriguing that I couldn't wait to discover what was happening outside of my view."* (P2). To imply locations of out-of-view speakers, we discovered that bidirectional hints were more significantly preferred than angle ones. However, the bidirectional design was nearly as preferred compared to the egocentric one using the Friedman & WNMT tests: $p_{ba} = 0.0124$, $p_{be} = 0.0765$. Among the feedback from all participants, one participant offered suggestions for the angle hint design. *"In my opinion, the angle one was the most intuitive since I knew that I could find speakers when the angle moved down to zero. However, I was considering that you don't need that high of precision. That's to say, you only need to show a rough relative degree for indicating the speaker's location."* (P11). Another participant (P12) thought that none of our designs were intuitive for him. *"In my view, to give users a clearer view, the compass-like design would be a better choice. In fact, this kind of design was more comprehensible in that it could be commonly seen in games, which often provide some sort of dotted hints for where other people were on the small compass."* (P12).

Out-of-view Caption			
	Comprehension	Comfort	Rank
Symbol (s)	3.42	3.75	3
Partial Content (p)	4.08	4.00	2
Complete Content (c)	4.42	4.00	1
Out-of-view Speaker's Location			
	Comprehension	Comfort	Rank
Bidirectional (b)	4.17	4.08	1
Numeric angle (a)	3.50	3.33	3
Egocentric (e)	4.08	3.67	2

Table 3. The summary of our 5-point Likert scale evaluation of our user study in terms of the hint indicators for out-of-view speakers, including the out-of-view caption and out-of-view speaker's location hint.

IMPLEMENTATION AND EVALUATION

In order to collect initial feedback from users directly experiencing our interface design, we developed SpeechBubbles, a prototype that uses a Microsoft HoloLens device for visualization and an Aputure Lavalier microphone for sound processing. We implemented our rising bubble visualization as our selected display method, and the directional hint bubble with complete utterances displayed to indicate events occurring externally from the user's view.

Interface Implementation

Our system consists of two main parts: the recognition side and the interface side. For recognition, we utilized the Google Speech API for recognizing speech after receiving the voice signal, which sends the processed data by socket. For the interface side, the data sent from the recognition side displays the dialogue along with our visual cues on the HoloLens using Unity [4]. Since the objective of our implementation was to mainly test our visual design, we used external speakers for recognition instead of a microphone array used in prior works, since the latter had greater processing delays that would negatively affect our prototype's evaluation. We sent all collected data from the different speakers to a single server, and broadcasted the data—which was collected from both the conversational and directional information—directly to the Hololens via SimpleHTTPServer. To visualize the data, we followed the results from our initial study and set up the corresponding settings from each of our design choices (e.g., text style). We chose a rising bubble display over other display choices and a directional hint bubble with complete utterance displayed from other hint methods, since these approaches had evaluated better from overall comparisons.

Evaluation Methodology

To evaluate our interface design for DHH individuals in group conversations, we conducted a user study (Figure 4) that compared the background text style and preferable display type between SpeechBubbles and traditional captioning. We recruited six participants (2 female) whose ages ranged from 20 to 25 years (mean=23, SD=2.16) from the internet. Our six participants had varying degrees of hearing loss: moderate (2), severe (2), and extremely severe (2). Half the participants (3) were born into hearing loss, and all participants (6) had binaural hearing loss.

Before each study session, we introduced the Microsoft HoloLens and the hint functionality to the participant. After

the participant became familiarized with the device, we seated them at a chair across from two speakers, located 0° and 25° relative to the participant, respectively. Each study session lasted about one hour. To understand the participant's preferable display background [20], we first allowed participants to choose from two display choices (i.e., either black background and white text, or white background and black text), and then displayed their choice for the study. Our study consisted of two parts (i.e., traditional captions and SpeechBubbles), where the order of presenting the two parts were counterbalanced. We introduced two topics to the participant for conversing in a group with two other speakers: vacation and favorite food. The two speakers initiated the conversation, and we encouraged the participant to engage in conversation with them. Each conversation lasted about one minute. We then displayed the conversation on the screen for the participant to view in real-time. We followed up each conversation session with a questionnaire based on the participant's experience with either the traditional caption display or the SpeechBubbles display, depending on which approach was presented to them at the time.

Figure 4. An external view and a live demo view from the HoloLens.

Preliminary Interface Feedback

After the participant completed both parts of the study, we asked them to provide us with feedback on both approaches and to select their preferable display type.

Background preference

All participants (6) agreed that the black background was the more suitable display. Nonetheless, two participants also expressed preference for the white background with black text, since it matched the look of popular social media interfaces such as Line and Facebook.

Quantified analysis

We asked participants to rate comprehension and comfort on a 5-point Likert scale, as well as their overall preference for caption vs. SpeechBubbles. Participants rated captions with 3.67 (SD=0.52) in comprehension and 3.17 (SD=0.75) in comfort, and rated SpeechBubbles with 4.00 (SD=0.64) in comprehension and 2.83 (SD=0.98) in comfort. There was no statistical significance between the ratings of the two conditions. In terms of overall preference, 83% of the participants preferred SpeechBubbles over captions.

Qualitative feedback

All participants (6) considered SpeechBubbles to be helpful in conversations for some contexts. Most participants responded that SpeechBubbles made it is easier to identify speakers in the conversation. *"It was easier with SpeechBubbles to identify who was talking, whereas not knowing who the speaker was in caption system could misconstrue the conversation."* (P5). We identified two problems that most of the participants addressed during the feedback section. Firstly, three of the participants reported that the HoloLens's screen size was too small. Secondly, all participants felt like they had difficulty keeping up with conversations due to delays in the voice recognition, especially in the captioning system. *"[The captioning] only has one line of caption, plus there were delays in the voice recognition, so I wasn't able to identify who was speaking."* (P1). However, participants in the SpeechBubbles study were able to identify the speaker in conversations. *"I felt more satisfied with SpeechBubbles, because I was able to identify who was talking during the conversation."* (P2).

DISCUSSION AND FUTURE WORK

From our results, participants had offered design suggestions for improving future assistive interfaces specific to DHH users in group conversations.

- **Speaker association:** The bubble-like style was favored over traditional captioning. Captioning as a display method should be replaced with more intuitive, symbolic graphic design along with content.
- **Amount of content:** Bubble-like designs should contain more than one line of text. For captioning in Mandarin, there should be at most 12 Chinese characters per line.
- **Order of utterances:** Bubble-like designs with content rising vertically upwards on the screen is the most comprehensible and desirable visualization design for users in daily casual conversations.
- **Out-of-view speaker**: Most users preferred displaying full or partial speech content with bidirectional indications as hint cues. With display methods that are more intuitive than those including additional numerical information, users can more quickly discover and locate out-of-view speakers and their speech utterances.

We also address further topics outside the focus of this current work, but which we consider worth discussing as next step considerations: captioning with different languages, emotion behind captions, and potential applications for hearing people.

Captioning with Different Languages

Although our work primarily focused on captioning for Mandarin, we also explored captioning content in other languages. For Mandarin, written sentences consist of logograms, where utterances correspond to Chinese characters displayed in the caption. However, English sentences are composed of phonograms, where utterances instead correspond to letters. The writing scripts of these two languages would directly determine how much space can be allocated for each single caption line. On average, there are more language symbols allocated in daily-use English sentences compared to daily-use Chinese sentences. Therefore, designers need to consider adjustments in our specifications for employing bubble-like designs in other languages. For example, the Japanese and Korean languages may require

more complex adjustments for their writing scripts, since those languages can combine logograms and phonograms.

Emotion behind Captions

Emotions that people convey behind utterances is a valuable communication trait to more deeply engage in conversations. That is, people usually express themselves with different tones to convey their moods. However, DHH individuals who rely on reading visuals cues may misunderstand certain conversational situations, due to not being able to infer emotions conveyed from speech through these visual cues alone. Therefore, our studies can be expanded to investigate people's emotions for captioning interfaces, such as different caption fonts and speech bubble styles that map to different emotions.

Contributions to Hearing Individuals

When we initially conceptualized designs for HMDs, we developed our study around issues that DHH individuals faced. However, some of our study findings could potentially be applied to hearing individuals for certain scenarios. For example in situations with unfamiliar spoken languages, our design could be applied to HMD interfaces to assist hearing individuals with real-time machine translations.

Additional Next Steps

We also envision at least several more potential next steps. For wider group conversation situations, we would like to evaluate with hearing participants whom can sign and with DHH participants whom rely on captions in other languages. For interface robustness, we would like to evaluate in actively-noisy environments. For broader user access potential, we would like to expand our study to more widely include those with less severe hearing loss. For improved device usage, we would like to explore smartglasses with wider viewing areas and that can reliably integrate visualizations with a wearable microphone array. Lastly, we would like to explore support for visualizing emotions expressed in utterances for richer conversation experiences.

CONCLUSION

We propose SpeechBubbles, a real-time captioning interface with a bubble-like display to enhance DHH individuals' group conversation experiences. We interviewed eight DHH individuals and discovered their group conversation issues, and also asked them to co-design ideal visualizations for potential captioning solutions. To better understand user preferences for the prototype design, we conducted a 12-person user study—using comprehension and comfort as factors—to explore several ideal designs for the speech bubble display and hint cues. We evaluated our prototype from the design choices on six participants. Their feedback provided potential ideas to further expand our design perspective for enhancing DHH individuals' group conversation experiences.

ACKNOWLEDGMENT

We offer our deepest gratitude to Professor Richard E. Ladner from the University of Washington's Paul G. Allen School of Computer Science & Engineering for his valuable insights in the development of our work's participatory design.

REFERENCES

1. Ava. http://www.ava.me/.
2. Google Noto Fonts. https://www.google.com/get/noto/help/guidelines/.
3. UNI. http://www.motionsavvy.com/.
4. Unity. https://www.unity3d.com/.
5. Andy Brown, Rhia Jones, Mike Crabb, James Sandford, Matthew Brooks, Mike Armstrong, and Caroline Jay. 2015. Dynamic Subtitles: The User Experience. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video (TVX '15)*. ACM, New York, NY, USA, 103–112.
6. Delia Chiaro, Christine Heiss, and Chiara Bucaria. 2008. *Between Text and Image: Updating Research in Screen Translation*. John Benjamins Publishing Company, Amsterdam, Netherlands.
7. Bong-Kyung Chun, Dong-Sung Ryu, Won-Il Hwang, and Hwan-Gue Cho. 2006. An Automated Procedure for Word Balloon Placement in Cinema Comics. In *Advances in Visual Computing: Second International Symposium, ISVC 2006 Lake Tahoe, NV, USA, November 6-8, 2006. Proceedings, Part II*. Springer Berlin Heidelberg, Berlin, Heidelberg, 576–585.
8. Michael Crabb, Rhianne Jones, Mike Armstrong, and Chris J. Hughes. 2015. Online News Videos: The UX of Subtitle Position. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS '15)*. ACM, New York, NY, USA, 215–222.
9. Marilyn E. Demorest and Sue Ann Erdman. 1986. Scale Composition and Item Analysis of the Communication Profile for the Hearing Impaired. *Journal of Speech and Hearing Research* 29, 4 (dec 1986), 515–535.
10. Gilbert C.F. Fong. 2009. Let the Words Do the Talking: The Nature and Art of Subtitling. In *Dubbing and Subtitling in a World Context*. The Chinese University of Hong Kong, 91–106.
11. Benjamin M. Gorman. 2014. VisAural: A Wearable Sound-localisation Device for People with Impaired Hearing. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS '14)*. ACM, New York, NY, USA, 337–338.
12. Jan Gugenheimer, Katrin Plaumann, Florian Schaub, Patrizia Di Campli San Vito, Saskia Duck, Melanie Rabus, and Enrico Rukzio. 2017. The Impact of Assistive Technology on Communication Quality Between Deaf and Hearing Individuals. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 669–682.
13. Stephen R. Gulliver. 2002. Impact of Captions on Deaf and Hearing Perception of Multimedia Video Clips. In *Proceedings of the 2002 IEEE International Conference on Multimedia and Expo (ICME '02)*, Vol. 1. IEEE, Washington, DC, USA, 753–756.

14. Stephen R. Gulliver and George Ghinea. 2003. How Level and Type of Deafness Affect User Perception of Multimedia Video Clips. *Universal Access in the Information Society* 2, 4 (nov 2003), 374–386.
15. Richard S. Hallam and Roslyn Corney. 2014. Conversation Tactics in Persons with Normal Hearing and Hearing-impairment. *International Journal of Audiology* 53, 3 (mar 2014), 174–181.
16. Marion A. Hersh and Michael A. Johnson. 2003. *Assistive Technology for the Hearing-impaired, Deaf and Deafblind*. Springer-Verlag London, London, England, UK.
17. Myles Hollander, Douglas A. Wolfe, and Eric Chicken. 1999. *Nonparametric Statistical Methods*. Wiley, New York, NY, USA.
18. Richang Hong, Meng Wang, Mengdi Xu, Shuicheng Yan, and Tat-Seng Chua. 2010. Dynamic Captioning: Video Accessibility Enhancement for Hearing Impairment. In *Proceedings of the 18th ACM International Conference on Multimedia (MM '10)*. ACM, New York, NY, USA, 421–430.
19. Dhruv Jain, Leah Findlater, Jamie Gilkeson, Benjamin Holland, Ramani Duraiswami, Dmitry Zotkin, Christian Vogler, and Jon E. Froehlich. 2015. Head-Mounted Display Visualizations to Support Sound Awareness for the Deaf and Hard of Hearing. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 241–250.
20. Jacek Jankowski, Krystian Samp, Izabela Irzynska, Marek Jozwowicz, and Stefan Decker. 2010. Integrating Text with Video and 3D Graphics: The Effects of Text Drawing Styles on Text Readability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 1321–1330.
21. Fotios Karamitroglou. 1998. A Proposed Set of Subtitling Standards in Europe. *Translation Journal* 2, 2 (1998), 1âĂŞ15.
22. Saba Kawas, George Karalis, Tzu Wen, and Richard E. Ladner. 2016. Improving Real-Time Captioning Experiences for Deaf and Hard of Hearing Students. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '16)*. ACM, New York, NY, USA, 15–23.
23. Cees M. Koolstra, Allerd L. Peeters, and Herman Spinhof. 2002. The Pros and Cons of Dubbing and Subtitling. *European Journal of Communication* 17 (sep 2002), 325–354.
24. David Kurlander, Tim Skelly, and David Salesin. 1996. Comic Chat. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '96)*. ACM, New York, NY, USA, 225–236.
25. Kuno Kurzhals, Emine Cetinkaya, Yongtao Hu, Wenping Wang, and Daniel Weiskopf. 2017. Close to the Action: Eye-Tracking Evaluation of Speaker-Following Subtitles. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 6559–6568.
26. Raja S. Kushalnagar, Gary W. Behm, Aaron W. Kelstone, and Shareef Ali. 2015. Tracked Speech-To-Text Display: Enhancing Accessibility and Readability of Real-Time Speech-To-Text. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS '15)*. ACM, New York, NY, USA, 223–230.
27. Tara Matthews, Janette Fong, and Jennifer Mankoff. 2005. Visualizing Non-speech Sounds for the Deaf. In *Proceedings of the 7th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '05)*. ACM, New York, NY, USA, 52–59.
28. Ryan W. McCreery, Rebecca A. Venediktov, Jaumeiko J. Coleman, and Hillary M. Leech. 2012. An Evidence-Based Systematic Review of Directional Microphones and Digital Noise Reduction Hearing Aids in School-Age Children With Hearing Loss. *American Journal of Audiology* 21, 2 (dec 2012), 295–312.
29. Anne Marie Piper and James D. Hollan. 2008. Supporting Medical Conversations Between Deaf and Hearing Individuals with Tabletop Displays. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW '08)*. ACM, New York, NY, USA, 147–156.
30. Ruiwei Shen, Tsutomu Terada, and Masahiko Tsukamoto. 2012. A System for Visualizing Sound Source Using Augmented Reality. In *Proceedings of the 10th International Conference on Advances in Mobile Computing & Multimedia (MoMM '12)*. ACM, New York, NY, USA, 97–102.
31. Nancy Tye-Murray, Suzanne C. Purdy, and George G. Woodworth. 1992. Reported Use of Communication Strategies by SHHH Members: Client, Talker, and Situational Variables. *Journal of Speech, Language, and Hearing Research* 35 (jun 1992), 708–717.
32. Y. Wang. 2006. Discussion on Technical Principle for Handling with Translation of Captions of Movies and Televisions. *Journal of Hebei Polytechnic College* 6, 1 (2006), 61–63.
33. Huayong Zhao. 2000. *The Approach & Research of Foreign Film Dubbing*. Broadcasting Corporation of China, Beijing, China.