HoloSound: Combining Speech and Sound Identification for Deaf or Hard of Hearing Users on a Head-mounted Display

Ru Guo*1, Yiru Yang*1, Johnson Kuang[1], Xue Bin[2], Dhruv Jain[1], Steven Goodman[3], Leah Findlater[3], Jon E. Froehlich[1]

[1]Computer Science and Engineering, [2]HCI + Design, [3] Human Centered Design and Engineering
University of Washington, Seattle, WA, USA
{grgrggtr*, yangy87*, jkuang7, djain, jonf}@cs.uw.edu, {xb285, smgoodmn, leahkf}@uw.edu
(*equal contribution)

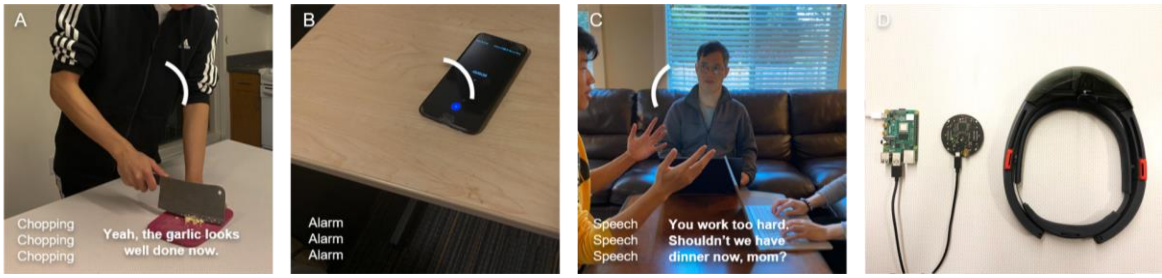

A
Chopping
Chopping
Chopping
Yeah, the garlic looks well done now.
B
Alarm
Alarm
Alarm
C
Speech
Speech
Speech
You work too hard. Shouldn't we have dinner now, mom?
D

Figure 1: Illustrations of HoloSound showing sound identity, source location, and speech transcription. The three most recent sounds are shown at the bottom left of the display, the locations of at most four simultaneous sound sources are shown as circular arcs in the center, and the speech transcription is either shown as subtitles or can be positioned close to the speakers in the 3D space (not shown). See supplementary video.

ABSTRACT

Head-mounted displays can provide private and glanceable speech and sound feedback to deaf and hard of hearing people, yet prior systems have largely focused on speech transcription. We introduce *HoloSound*, a HoloLens-based augmented reality (AR) prototype that uses deep learning to classify and visualize sound identity and location in addition to providing speech transcription. This poster paper presents a working proof-of-concept prototype, and discusses future opportunities for advancing AR-based sound awareness.

CCS CONCEPTS

• Human-centered computing ~ Accessibility ~ Accessibility technologies

KEYWORDS

Augmented reality; head-mounted display; deaf; hard of hearing; speech-transcription; real-time captioning; sound recognition; sound localization; sound awareness.

1 Introduction

Head-mounted displays (HMDs) have the potential to provide glanceable, always available, and private sound and speech feedback to deaf and hard of hearing (DHH) users [4,11,19]. Yet, most prior work on HMD-based feedback has focused on speech transcription [11,13,19]. While this speech-centric work has shown promise in making conversations accessible, other aspects of sounds may also be useful, such as the identity of non-speech sounds [16] and the location of the sound source [7]. Thus, a few researchers have investigated showing sound localization on an HMD using external microphone arrays [8,13]. Finally, although not on HMDs, past systems have identified and conveyed non-speech sounds (*e.g.*, doorbells, knocking) on the smart home displays [16] or a smartwatch [17].

While conveying these individual sound properties was deemed useful [13,14,17], in real life, speech and sound information often co-exist, and must be conveyed simultaneously. Indeed, in a recent large-scale survey, DHH people expressed a strong desire for receiving non-speech sound cues along with the speech transcription [3]. Displaying these multiple sound cues together in an unobtrusive and glanceable manner is a challenging problem that remains to be explored.

To begin investigating this problem, we present an early prototype of an augmented reality (AR) based system, called *HoloSound*, that leverages advances in deep learning and sound sensing to simultaneously provide three key desired sound properties to DHH users in real-time: speech transcription, sound identity, and source location (Figure 1). HoloSound uses a speech-to-text API to generate a transcription that can be positioned in 3D space, a deep-learning engine to display the three most recent non-speech sounds (*e.g.*, doorbell, knocking), and an external microphone array to visualize the direction of at most four sound sources in the vicinity. Though our current user interface is preliminary, we plan to refine HoloSound's design through a design probe study with DHH users.

In this poster paper, we describe the HoloSound system, discuss our plans for future user evaluation with DHH users, and enumerate opportunities for further advancing AR-based sound awareness.

2 The HoloSound System

While not all DHH people want to use sound feedback technologies, past large-scale surveys with DHH people [3,4], show that many would find such technologies desirable and useful in everyday activities. HoloSound is informed by prior AR-based sound awareness work with DHH users [11,13,19], as well as personal (*e.g.,* [12]) and research experiences (*e.g.,* [15]) of one of our authors (Jain) who is hard of hearing. The system consists of two parts: an app running on the HoloLens and an external microphone array, *ReSpeaker* [20], retrofitted on top of the HoloLens (Figure 2). A cloud-based server is used to interface the portable microphone array with the HoloLens and to run the sound recognition engine. Figure 1 shows the preliminary user interface. Below, we detail the three key components of HoloSound, which are also demonstrated in the supplementary video. The system is open sourced on GitHub: https://git.io/JJaHe.

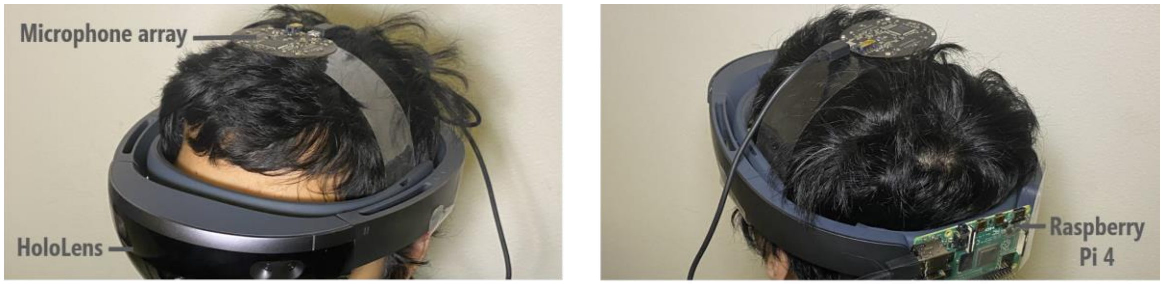


Figure 2: The three components of HoloSound system.

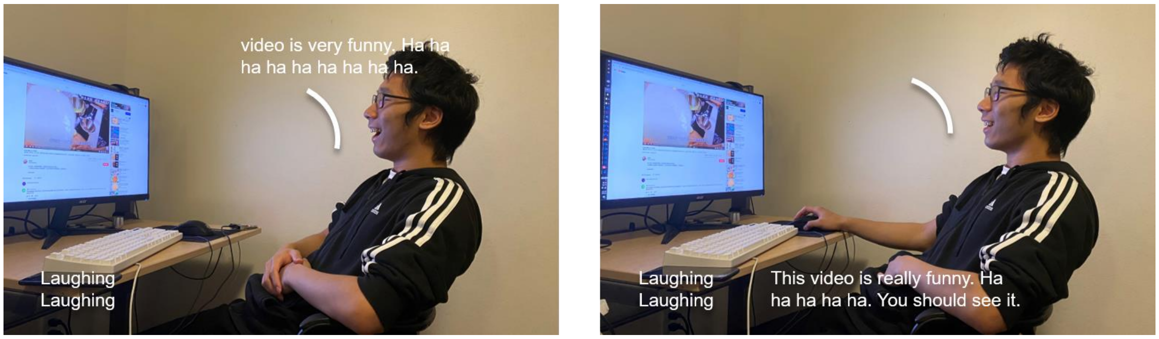
2.1 Speech Transcription

Many DHH users use speech-transcription [6,18], and a recent survey [4] showed that HMDs were the most preferred wearable device for speech feedback. To transcribe speech, HoloSound uses Microsoft Azure's *Speech-to-text* API [21]. To accommodate multiple contexts-of-use, we offer two views for displaying the transcribed text: *windows* and *subtitles*, both informed from prior work [11], which used a human transcriptionist rather than automated methods to generate captions for display on a HoloLens device.

In the windows view, the goal is to reduce the visual split between the transcribed text and the speaker, hence the user can place the text windows on top of the speakers using the HoloLens' pinch gesture (Figure 3 left). This view could be more suitable when the speakers are stationary (*e.g.*, in a group meeting [19]). We use the HoloLens' 3D spatial mapping feature to recognize the environment and automatically position the captions at an appropriate depth near the user's desired spatial location.

In the subtitles view, a single text block appears at a fixed distance in front of the user and moves with the user's head (Figure 3 right). This view is analogous to video captioning and could be preferred when the speakers are moving (*e.g.*, in a lecture setting [11] or while walking [14]).

For both views, the text scrolls up and disappears as the new transcription is appended. Users can also customize the transcribed text's font size (default: 0.75° angular), the number of lines (default: 2), the width of each line (default: 60 characters), and the distance of captions from the eye (2m, 4m, 8m, or the default: projected onto background surfaces (*e.g.*, walls) [14]). To stabilize jitter, the text block stays at the same location and smoothly drifts along when the user's head moves by at least 25°.


video is very funny. Ha ha ha ha ha ha ha ha.

Laughing
Laughing

Laughing
Laughing

This video is really funny. Ha ha ha ha ha. You should see it.

Figure 3: Speech-transcription can either be (A) placed on top of the speakers in the 3D space (*windows* view) or (B) move with the user's head (*subtitles* view).

2.2 Sound Recognition

Besides speech transcription, HoloSound shows the three most recently recognized sound events (*e.g.*, doorbell, knocking) at the bottom-left corner of the display (Figure 1). We use a deep learning-based sound classification engine running on a cloud that continually senses and processes audio in real-time.

To create the classification engine, we followed an approach similar to *HomeSound* [16], which uses transfer learning to adapt a deep CNN-based image classification model (VGG) for sound classification. This model achieved an overall accuracy of 84.9% on sounds recorded in the homes. To train the VGG model, we used sound clips of 19 common sound classes preferred by DHH people (*e.g.*, door knock, fire alarm, phone ring) from online sound effects libraries (*e.g.*, BBC [22], FreeSound [5]). All clips were converted to a single format (16KHz, 16-bit, mono) and silences greater than one second were removed, resulting in 27.8 hours of recordings. We then used the method in Hershey *et al.* [10] to compute the log-mel spectrogram features, which were fed to the model.

To process sounds in real-time, HoloSound uses a sliding window approach to sample the microphone at 16KHz (16,000 samples every second), extract the log-mel spectrogram features, and upload the 1-second buffer to the cloud. After classification, all sounds below 50% confidence and 45dB loudness are ignored.

2.3 Sound Localization

The third key desired property conveyed by HoloSound is sound location. For localization, we use *ReSpeaker* [20], an external portable 4-microphone array (Figure 2), which we ultimately envision could be integrated into future AR devices. Though HoloLens has four onboard microphones [23], these microphones are specifically designed for voice input from the user (*e.g.*, by enhancing audio input from the user's face through beamforming) [1] and thus cannot be used for accurate localization. The ReSpeaker array is coupled to a *Raspberry Pi 4* [24] running a modified 3D Kalman filter sound localization algorithm [9]. After processing, the direction of at most four sound sources in the user's vicinity is sent to the HoloSound app through a WIFI server. To visualize each sound, the continuous 3D direction is projected to one of 12 discrete directions in the horizontal plane; these values are then shown as circular arcs in a top-down view on the HoloLens display's vertical plane (Figure 1).

3 Discussion

In this paper, we introduced an initial AR prototype for visualizing three key components of sound information on HMDs in real-time (sound identity, source location, and speech transcription). However, considerable work remains in studying this system with DHH users and iterating on its design. Below, we discuss our future plans for a user study and further exploration opportunities for AR-based sound awareness.

UI exploration. While our current UI is preliminary, we plan to use HoloSound to prototype multiple UI designs and conduct a design probe study with 12-16 DHH individuals. Informed from a design space outlined in our past work [11], our goal is to explore how these designs may vary with different social contexts (*i.e.*, 1:1 meeting *vs.* a midsize meeting *vs.* a large lecture). Specific research questions of interest, include:

1. In what contexts do the users desire full transcription *vs.* a topical summary?
2. How might the UI design vary with the conversation importance (*e.g.*, with friends *vs.* at workplace)?
3. How many source locations should be shown simultaneously, and how does this vary with context?
4. What non-speech sounds are desired for each context? How should these sounds be visualized?
5. Should the wearer's voice be transcribed?

A key aspect of this study will involve balancing cognitive load with useful information. For example, if the user is involved in an important 1:1 meeting, the system could prioritize speech transcription and show alerts for important non speech-sounds only (*e.g.*, fire alarms).

System exploration. We will also perform accuracy and performance (latency, CPU, and memory usage) testing of our sound recognition and localization systems, which could have a significant impact on the user experience. For localization accuracy, we will place a sound source at different angles in front of our system and measure the mean angular error and standard deviation. For sound recognition accuracy, since the accuracy may vary with audio contexts and background noise, a hearing user will wear the device in different physical locations (*e.g.*, home, in transit, office) for several hours and report on whether the sound recognition is accurate. Finally, for performance testing, we will play recordings of real-life sounds and speech on a computer placed near our system, and measure the HoloLens' CPU and memory usage as well as the end-to-end latencies of our speech transcription, sound recognition, and sound localization features.

Examining complementary haptic feedback. Another potential area of exploration is complementing the visual HMD feedback with haptic notifications delivered through a smartwatch or a custom hardware solution. While haptic feedback provides more limited bandwidth than visual feedback, it can be used to provide complementary information—such as to notify the user of an important non-speech sound [6], or to enhance transcription by providing speech tone [2]. Future work should compare performance tradeoffs in providing haptic notifications to complement visual information. One idea is to conduct a controlled study with varying device combinations (HMD-only, HMD + smartwatch, HMD + custom haptic hardware) and feedback modalities (visual-only, visual+haptic). Participants could be given a distractor task and asked to localize sounds emanating from a circular array of speakers placed around them, while measurements of speed and accuracy of identifying the sound source, as well as self-reported cognitive load are taken.

4 Conclusion

Our work contributes the design and implementation of an augmented reality system, called *HoloSound*, that uses a head-mounted display and an external microphone array for transcribing speech, identifying sounds, and localizing sound sources, which are displayed to the user in the 3D space. HoloSound is open sourced (https://git.io/JJaHe) and can be used to conduct AR-based design investigations with DHH users.

ACKNOWLEDGMENTS

This work was supported by a University of Washington Reality Lab Grant, National Science Foundation Grant no: IIS-1763199, HoloLens devices donation by Microsoft, and a Google Faculty Research Award.

REFERENCES

[1] Manish Sharma, Mallikarjuna Rao Abhijit Jana. *HoloLens Blueprints - Google Books*. Retrieved June 7, 2020 from

https://books.google.com/books?id=_Hc5DwAAQBAJ&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

[2] Edward T. Auer. 1998. Temporal and spatio-temporal vibrotactile displays for voice fundamental frequency: An initial evaluation of a new vibrotactile speech perception aid with normal-hearing and hearing-impaired individuals. *The Journal of the Acoustical Society of America* 104, 4: 2477. Retrieved from http://scitation.aip.org/content/asa/journal/jasa/104/4/10.1121/1.423909

[3] Danielle Bragg, Nicholas Huynh, and Richard E. Ladner. 2016. A Personalizable Mobile Sound Detector App Design for Deaf and Hard-of-Hearing Users. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*, 3–13.

[4] Leah Findlater, Bonnie Chinh, Dhruv Jain, Jon Froehlich, Raja Kushalnagar, and Angela Carey Lin. 2019. Deaf and Hard-of-hearing Individuals' Preferences for Wearable and Mobile Sound Awareness Technologies. In *SIGCHI Conference on Human Factors in Computing Systems (CHI). In Submission.*

[5] Eduardo Fonseca, Jordi Pons Puig, Xavier Favory, Frederic Font Corbera, Dmitry Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. 2017. Freesound datasets: a platform for the creation of open audio datasets. In *Hu X, Cunningham SJ, Turnbull D, Duan Z, editors. Proceedings of the 18th ISMIR Conference; 2017 oct 23-27; Suzhou, China.[Canada]: International Society for Music Information Retrieval; 2017. p. 486-93.*

[6] Abraham Glasser, Kesavan Kushalnagar, and Raja Kushalnagar. 2017. Deaf, hard of hearing, and hearing perspectives on using automatic speech recognition in conversation. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, 427–432. https://doi.org/10.1145/3132525.3134781

[7] Steven Goodman, Susanne Kirchner, Rose Guttman, Dhruv Jain, Jon Froehlich, and Leah Findlater. Evaluating Smartwatch-based Sound Feedback for Deaf and Hard-of-hearing Users Across Contexts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1–13.

[8] 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*, 337–338. https://doi.org/10.1145/2661334.2661410

[9] François Grondin and François Michaud. 2019. Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations. *Robotics and Autonomous Systems* 113: 63–80. Retrieved from =

[10] Shawn Hershey, Sourish Chaudhuri, Daniel P W Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, and others. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 131–135.

[11] Dhruv Jain, Bonnie Chinh, Leah Findlater, Raja Kushalnagar, and Jon Froehlich. 2018. Exploring Augmented Reality Approaches to Real-Time Captioning: A Preliminary Autoethnographic Study. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems*, 7–11.

[12] Dhruv Jain, Audrey Desjardins, Leah Findlater, and Jon E Froehlich. 2019. Autoethnography of a Hard of Hearing Traveler. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 236–248.

[13] Dhruv Jain, Leah Findlater, Christian Volger, Dmitry Zotkin, Ramani Duraiswami, and Jon 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*, 241–250.

[14] Dhruv Jain, Rachel Franz, Leah Findlater, Jackson Cannon, Raja Kushalnagar, and Jon Froehlich. 2018. Towards Accessible Conversations in a Mobile Context for People who are Deaf and Hard of Hearing. In *Proceedings of ACM ASSETS 2018*, 12 pages.

[15] Dhruv Jain, Angela Carey Lin, Marcus Amalachandran, Aileen Zeng, Rose Guttman, Leah Findlater, and Jon Froehlich. 2019. Exploring Sound Awareness in the Home for People who are Deaf or Hard of Hearing. In *SIGCHI Conference on Human Factors in Computing Systems (CHI). In Submission.*

[16] Dhruv Jain, Kelly Mack, Akli Amrous, Matt Wright, Steven Goodman, Leah Findlater, and Jon E Froehlich. 2020. HomeSound: An Iterative Field Deployment of an In-Home Sound Awareness System for Deaf or Hard of Hearing Users. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20), 1–12. https://doi.org/10.1145/3313831.3376758

[17] Dhruv Jain, Hung Ngo, Pratyush Patel, Steven Goodman, Leah Findlater, and Jon Froehlich. 2020. SoundWatch: Exploring Smartwatch-based Deep Learning Approaches to Support Sound Awareness for Deaf and Hard of Hearing Users. In *ACM SIGACCESS conference on Computers and accessibility*, 1–13.

[18] Raja. S Kushalnagar, Walter S Lasecki, and Jeffrey P Bigham. 2014. Accessibility Evaluation of Classroom Captions. *ACM Transactions on Accessible Computing* 5, 3: 1–24. https://doi.org/10.1145/2543578

[19] Yi-Hao Peng, Ming-Wei Hsu, Paul Taele, Ting-Yu Lin, Po-En Lai, Leon Hsu, Tzu-chuan Chen, Te-Yen Wu, Yu-An Chen, Hsien-Hui Tang, and Mike Y. Chen. 2018. SpeechBubbles: Enhancing Captioning Experiences for Deaf and Hard-of-Hearing People in Group Conversations. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)*, Paper No. 293.

[20] ReSpeaker Mic Array v2.0 - Seeed Wiki. Retrieved June 7, 2020 from https://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0/

[21] Speech to Text | Microsoft Azure. Retrieved June 7, 2020 from https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/

[22] BBC Sound Effects. Retrieved September 18, 2019 from http://bbcsfx.acropolis.org.uk/

[23] HoloLens (1st gen) hardware | Microsoft Docs. Retrieved June 7, 2020 from https://docs.microsoft.com/en-us/hololens/hololens1-hardware

[24] Raspberry Pi 4. Retrieved June 7, 2020 from https://www.raspberrypi.org/products/raspberry-pi-4-model-b/