Accessible Video Calling: Enabling Nonvisual Perception of Visual Conversation Cues

LEI SHI[*], Microsoft Research & Information Science, Cornell Tech, Cornell University, USA
BRIANNA J. TOMLINSON[*], Microsoft Research & School of Interactive Computing, Georgia, USA Institute of Technology
JOHN TANG, Microsoft Research, USA
EDWARD CUTRELL, Microsoft Research, USA
DANIEL MCDUFF, Microsoft Research, USA
GINA VENOLIA, Microsoft Research, USA
PAUL JOHNS, Microsoft Research, USA
KAEL ROWAN, Microsoft Research, USA

Nonvisually Accessible Video Calling (NAVC) is a prototype that detects visual conversation cues in a video call and uses audio cues to convey them to a user who is blind or low-vision. NAVC uses audio cues inspired by movie soundtracks to convey *Attention*, *Agreement*, *Disagreement*, *Happiness*, *Thinking*, and *Surprise*. When designing NAVC, we partnered with people who are blind or low-vision through a user-centered design process that included need-finding interviews and design reviews. To evaluate NAVC, we conducted a user study with 16 participants. The study provided feedback on the NAVC prototype and showed that the participants could easily discern some cues, like *Attention* and *Agreement*, but had trouble distinguishing others. The accuracy of the prototype in detecting conversation cues emerged as a key concern, especially in avoiding false positives and in detecting negative emotions, which tend to be masked in social conversations. This research identified challenges and design opportunities in using AI models to enable accessible video calling.

CCS Concepts: • **Human-centered computing → Accessibility**; Empirical studies in accessibility

KEYWORDS

Video calling; blind or low vision; audio cues; assistive technology; computer-mediated communication.

ACM Reference format:

Lei Shi, Brianna J. Tomlinson, John Tang, Edward Cutrell, Daniel McDuff, Gina Venolia, Paul Johns, Kael Rowan. 2019. Accessible Video Calling: Enabling Nonvisual Perception of Visual Conversation Cues. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3, No. CSCW, Article 131, November 2019. ACM, New York, NY, USA. 22 pages. https://doi.org/10.1145/335923.

[*] Both authors contributed equally to this work.

2573-0142/2019/November – ART131... $15.00

https://doi.org/10.1145/3359233

1 INTRODUCTION

People who are blind or low-vision (BLV) have difficulty perceiving the visual cues that often play a central role in communication. The field of kinesics focuses on how humans communicate through body movement and gesture, an important nonverbal communication channel that is largely conveyed visually [3, 4]. One advantage of video calling over audio-only phone calls for remote communication is the ability to convey and visually perceive these kinesic cues [17]. However, these visual conversation cues (VCCs) in video calls are not accessible to people who are BLV, which can lead to conversational asymmetry, particularly with sighted partners.

By leveraging the video and audio streams already used in video calls, computer vision and artificial intelligence technologies can be used to detect a variety of VCCs that naturally occur in video calling. In this research, we explore how we might use these technologies to enable BLV users to perceive and react to VCCs in video calls. Conveying these cues to BLV users provides important context for them to interpret and mediate conversations, potentially providing a richer conversational experience. We present Nonvisually Accessible Video Calling (NAVC), a prototype system that uses musical sounds inspired by movie soundtracks to convey transient VCCs, including facial gestures such as surprise and happiness or nodding/shaking the head in agreement/disagreement, as well as more static cues such as focused attention. Fig. 1 diagrams how NAVC augments a regular video call by detecting VCCs and conveying them to a user through audio cues. Conceptually, NAVC could be added to any video call without requiring the installation of any additional equipment by the end user.

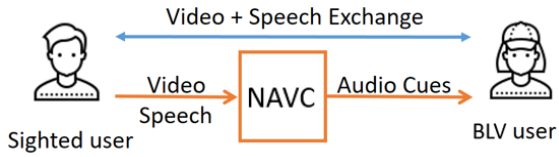

Fig. 1. Besides the typical video and speech exchange in video calls (in blue), NAVC analyzes the video and speech from a conversation partner to send audio cues to convey the visual conversational cues to a user who is blind or low-vision (in orange).

We designed and evaluated NAVC in a user-centered design process. We started with need-finding interviews with 26 BLV participants to understand their experience with video calling and the kinds of VCCs that they felt would be most useful to perceive. Then, as we began the design and implementation of NAVC, we held two design reviews with an agency that provides training to the BLV community. Based on these deep engagements with BLV informants, we built a prototype that surfaces a set of VCCs to BLV callers during video calls. Finally, we performed a user study of the system with 16 BLV participants in video calls, comparing the experience with and without the cues provided by NAVC. We conclude with observations learned from the user study and design implications for future work.

2 RELATED WORK

2.1 Cues in Conversations

We surveyed prior literature exploring the kinds of information that sighted and BLV interlocutors use in conversations. Research by Krishna et al. [18] compared the challenges of

maintaining social situational awareness for sighted, remote collaborators and people who are BLV. They noted similarities between both groups in managing conflict, reading non-verbal cues, and expressing opinions. Fichten et al.'s [11] early interview study compared communication cues that BLV and sighted people use in conversations. They documented how sighted people relied more on visual cues whereas people who are BLV relied more on audible cues. They noted cues of head nodding, eye contact, and facial expressions used by sighted people to show conversational interest, which drew our attention to surfacing these cues to BLV users. More recently, Qiu et al. [26] explored a wide range of in-person conversational needs and signals between sighted and BLV partners, many of which were visual cues that are inaccessible.

It is also helpful to consider the kinds of benefits that video calling provides over audio-only calls. One representative study comparing interaction among remote work team members with and without video [17] found that video enabled demonstrating understanding, managing pauses in the conversation, and especially helping handle conflict and disagreement. Altogether, this research has shown that visual conversational cues provide important supplementary information to conversation partners that an audio-only channel would miss.

2.2 Assistive Communication Devices

The advent of sensing and computing technologies has led to research prototypes exploring sensory substitution, where information available to one sense is conveyed via a different sense [23]. For example, BLV users can sense visual information that is conveyed via audio. Hermann et al. [15] explored interactive sonification to represent head motions to people who are BLV. They attached a motion sensor to a speaker's head which enabled them to associate head gestures with sounds that were conveyed to a person who is BLV. Zhao et al. [30] trained a face recognition system to help BLV users recognize the identity and facial expressions of friends (e.g., to select a conversation partner on FaceBook), but their system was not designed to be used in interactive conversation. The Expression prototype [1] used Google Glass as a video sensor for a conversational aid. The prototype used the camera to recognize smiling, yawning/sleepiness, and gaze direction in a conversation partner, which were conveyed through speaking these announcements to the BLV user. EmoAssist [27] used an Android mobile phone aimed at a conversational partner to visually detect seven behavioral expressions (sleepy, yawn, open lip smile, closed lip smile, looking left/right, looking up/down, head tilt). The recognized behavioral expressions were spoken to the BLV user through earphones. Importantly, these researchers noted that describing these cues via speech could conflict with the speech of the ongoing conversation, a critical consideration for a system such as NAVC.

To avoid linguistic conflict during conversations, other modalities such as touch can be used. For example, Krishna et al. [18] built a social interaction assistant for BLV people. The assistant used a megapixel digital camera attached to a pair of glasses to detect faces, facial expressions, and eye gaze. A tactile belt conveyed the direction and distance of a sighted conversation partner, and vibrotactile patterns on a haptic glove depicted facial expression information. While this haptic feedback avoids competing with the verbal speech, it requires additional devices and learning the mapping between the vibration patterns and detected cues.

2.3 Non-speech Audio Design

Another way to avoid linguistic conflicts during conversations is to convey information using non-speech audio. Non-speech auditory displays have been used across diverse application areas

for both BLV and sighted users, as they enable the quick representation of concepts in an accessible, non-visual way [9]. Auditory notifications and updates can be provided in direct, symbolic, or metaphorical ways. Auditory icons [14] are semantically-linked, environmental sounds, while earcons [6] often utilize designed, abstract sounds to encode information. Auditory icons may be easier to learn, but earcons support a larger range of informational representations, particularly for concepts without a direct mapping to real-life objects [13]. Earcon mappings may invoke embodied representations (e.g., physical movement) [2, 5].

Abstract audio may be particularly effective for communicating emotional affect. Researchers have explored the use of peripheral auditory cues in soundscapes and soundtracks for games and movies to convey emotional content [10, 16, 29]. Soundtracks are designed to convey contextual cues without masking the dialogue [28]. Other research has explored the use of auditory emoticons (non-verbal human sounds such as a laugh or sigh) to convey affective information [8].

Based on the rich prior research in audio displays, we designed a set of background sounds and earcons to convey visual communication cues that BLV callers may find useful. The design of these sounds leveraged inherent emotional representations of the sounds, similar to how soundtracks convey emotion during movies: discordant sounds mapped to disagreement or negative reactions, and harmonic sounds matched with agreement or positive reactions. We aimed to design a set of sounds which could convey these VCCs in an informative and non-interruptive way, paralleling how soundtracks augment the sensory and immersive experience of a movie without distracting from the dialog.

Building on prior work in assistive communication devices, NAVC aims to make video calls more accessible to people who are blind or low vision. Since the camera is already streaming video, NAVC can use this as a sensor to detect important visual conversation cues using cloud-based software without any additional equipment beyond what is needed for a typical video call. NAVC then renders this information to BLV users as non-speech audio in a relatively nonintrusive, intuitive fashion.

3 USER-CENTERED DESIGN PROCESS

The design of NAVC explored two research questions:

RQ1. What are the VCCs that are important to people who are BLV and can be reliably detected, and how should they be conveyed?

RQ2. What attributes of the cues used to convey VCCs while conversing over video calls are useful to BLV users?

The design space of video communication is very rich, ranging from one-on-one conversations, to group calls comprising dozens of individuals or co-located groups. As a starting point for our design, we focused on one-on-one conversations in English in an American context between a sighted person and one who has limited or no vision.

We followed a user-centered design process by beginning with interviews of people who are blind and low vision to identify their needs related to conversations with sighted people and get their reactions to a Wizard of Oz prototype. Then, as we worked to develop the application and build models for detecting different VCCs, we held two design review sessions with an agency that provides skills training to the BLV community. These sessions further guided the design and development of NAVC and kept us anchored to the context and needs of BLV users. Interactions with BLV users for the interviews and user study of the working prototype were reviewed by the institutional review board of our organization.

3.1 Need-finding Interviews

To begin this project, we conducted a set of interviews to explore the common practice and experience of people who are BLV when engaging in conversations. A regional center serving the BLV community helped us recruit 26 (16 female, 10 male) of their adult clients to participate in the interviews. The participants exhibited a range of visual conditions, including total blindness, tunnel vision, and lack of foveal vision. Overall, participants were unable to see facial expressions of a seated conversation partner. Hour-long interviews were conducted in-person at the center's sites in two U.S. cities (plus one conducted at the interviewee's place of work). Each participant received a $75 USD Amazon gift voucher as a gratuity for their participation.

During the interviews, we asked participants to think of a specific, recent, and substantial (about 10 minutes long) in-person conversation with one sighted person. We asked about the conversational cues that they used to interpret and manage the conversation and prompted them to specifically reflect on instances of detecting agreement, interest in the topic, disagreement, confusion, or time anxiety. We asked them to reflect on challenges in perceiving conversation cues given their visual ability.

We also introduced the concept of NAVC by playing sounds meant to convey the VCCs: paying visual attention, nodding head in agreement, shaking head in disagreement, confusion, eye rolling, yawning, raising a hand to talk, smiling, and looking at a watch. This list was drawn from cues conveyed in prior research [1, 11, 17, 23, 27] and suggestions from the participants themselves. We tested a Wizard of Oz (WOz) prototype in a conversation (about 5 minutes) where we manually triggered the sounds to convey the naturally occurring VCCs during the conversation. The researcher focused the conversation on opinions around current events (e.g., sports, news) and tried to express different VCCs when driving the conversation.

All interviews were video recorded and transcribed. Open and axial coding qualitative methods [7] were used to draw out the main themes from the interview data.

3.1.1 Needs in Conversation Cues

Consistent with prior research, participants mentioned how they used vocal cues (tone, inflection, pace, volume) as well as other audible cues (sighs, breath intake, pauses, silence) to detect conversation cues from a sighted conversation partner. These auditory cues were mentioned by 19 participants, highlighting that any generated conversation cues must not overpower the audible cues on which people are already relying. Some examples from the interviews (using their participant number, P0-P25):

P5: So I listened to audible cues. Like pauses or the way questions might be asked.

P7: For the most part, it was her tone of voice. At times, it was more what she was doing with her body, because I can see forms.

P5: I feel like having a conversation with a stranger might be a bit different because I feel like it's easier with people I know because I kind of have a pre-expectation of, maybe, a bias on what to expect here.

In response to a question about what conversation cues they wished they had access to, 17 (65%) explicitly mentioned wanting to see facial expressions:

P24: Absolutely, facial expression and whether they look disappointed or they looked happy. …and yeah, that's mainly because sometimes I can't tell.

P7: What I really miss are facial expressions. ...depending on how well I know the person, there's a lot of ambiguity in what's being said.

Those who could see body movements (like P7) mentioned reacting to them, as well. Even though most people recalled specific conversations with people they knew, some volunteered (as with the comments from P5 and P7) that discerning conversation cues with people they did not know well was more challenging.

3.1.2 Reaction to the NAVC WOz Prototype

Participants had universally positive responses to the WOz prototype of NAVC. On a scale of 1 (useless) to 10 (extremely useful), they rated it 7.8 on average, with only one rating below a 6:

P4: I like the fact that I can instantly tell whether you agreed or disagreed with what I'm saying.

P5: …assuming that it's accurate, it takes out the guess work and kind of tells me quickly. It's kind of telling me directly.

P8: That's basically saying you're agreeing right at that time with having a conversation instead of waiting 'til after the sentence. You agreed right as we were talking, like we were talking face-to-face. We both had vision.

P19: I think this could be helpful in clarifying someone's emotional tone and meaning… you just need a little bit of, like I keep saying, getting used to it.

While people wanted more time to learn the sounds paired with the cues (such as P19), they appreciated the instant feedback they got during the conversation.

3.1.3 Ratings of the Conversation Cues

After experiencing the prototype, participants rated a list of potential cues, most of which they experienced in the conversation, on scale of: 3 very important, 2 somewhat important, and 1 less important. Table 1 shows the average ratings and variance of the list of cues.

Table 1. Average ratings and variance of participants' ratings of the importance of visual conversation cues on scale of 1 (less important) to 3 (very important).

Cue	Average	Variance
Smiling	2.58	0.47
Attention	2.46	0.50
Hand Raise	2.43	0.44
Head nod	2.38	0.39
Head shake	2.38	0.47
Confusion	2.31	0.54
Eye rolling	2.23	0.49
Watch look	1.83	0.54
Yawning	1.31	0.27

The average ratings show an interest in smiling and less interest in looking at a watch or yawning as useful conversation cues. Participants mentioned that they could often detect yawning through audible cues. Except for yawning, which was rated lower with a relatively low

variance, each cue had a substantial variance in their ratings, reflecting a significant lack of agreement on the importance of different cues. People prioritized cues that were specifically important to them and that cue importance could vary greatly depending on their level of vision. For example, those who could see overall body motion did not find it important to detect head nodding and shaking, whereas those who were totally blind tended to rate them as very important. This finding is consistent with prior research on the need to tailor assistive technologies to the individual's needs and capabilities [25].

3.2 Design Review 1

The goal of the first design review was to solicit feedback for the revised WOz prototype. Based on the initial interviews and the encouraging response to the first WOz prototype, we began to develop a working prototype and iterated on sound designs for different conversational cues. To keep us connected to the needs of BLV users and as another approach to get design feedback, we partnered with an agency that provides skills training and technology development services to the BLV community. We convened a group design review with four members of their Access Technology department who were themselves BLV.

A researcher engaged in a conversation with one member using a revised WOz prototype and probed for the group's reactions to whether the VCCs represented were useful and if the design of the sounds evoked the corresponding VCC without being distracting to the conversation. We presented a set of sounds that provided cues for: attention, head nodding, head shaking, smiling, negative emotion (frowning or upset), wanting to take a turn, being expected to take a turn, and eyeroll. The *attention* cue played as a constant sound for as long as the conversation partner looked toward the screen/camera, while the rest were transient sounds triggered by the relevant event (e.g., smiling, head nod). These sounds are included in the Auxiliary Material supporting this paper.

Feedback on the sounds helped us identify important characteristics for the sound cues. Reviewers found the *attention* sound, which is usually playing, to be too high-pitched and intense, almost like a science fiction sound that created anxiety. They also strongly advocated for transient sounds that were short (less than one second) and relatively unobtrusive to avoid cluttering the audio modality. In response to a cartoonish sound for eyeroll, they pointed out that it would evoke too much of an amused reaction, disrupting the conversation, and was thus not a good model for sound cues.

While our initial interviews indicated that cues for turn-taking in group conversations would be very useful, reviewers observed that turn-taking was not a problem for our initial target scenario of one-on-one conversations. The design review also drew attention to the perspective of interpreting silence, a particularly problematic issue for people who are BLV in conversations. Cues that help explain silence or represent reactions that typically do not generate any sound were especially useful. When asked about other cues to represent, reviewers mentioned surprise as an example of a typically silent reaction.

3.3 Design Review 2

The goal of the second design review was to provide design feedback on a working prototype. Just over three weeks later, we convened another group design review with six members of the agency (three of whom overlapped with the first session). We integrated feedback from the first design review into a working prototype that provided cues for: *attention*, *head nodding*, *head shaking*,

smiling, *negative emotion* (frowning or upset), *surprise,* and *thinking* (also included in the Auxiliary Material for this paper).

Similar to the first design review, we held a conversation with a different meeting participant while using the working prototype and probed for feedback on features and sound design. We were particularly interested in opinions on two options for the attention sound—since it would be playing much of the time, we wanted to be careful to avoid making it annoying.

A main learning from this design review was an implication of how the facial gesture recognition technology worked in the context of conversation. We experienced many triggers of the *smile* cue during the conversation and discovered that the mouth movements during conversation would incidentally match the facial profile for smiling, generating a false positive. Thus, we had to suppress triggering facial gesture cues when the person was talking. This approach is consistent with focusing on cues that help interpret silence (when the person is *not* talking) and did not apply to gestures that are not focused on the face (such as *head nodding* and *head shaking*). Reviewers also recommended that it was more important to avoid incorrectly triggering a sound when there is no VCC (false positive) than to miss triggering a sound when there is a VCC (false negative).

Taken together, the two design review sessions with people who have technical, practical, and personal expertise in the needs of BLV users were valuable in accelerating the refinement of the NAVC prototype design. They not only helped us focus on which VCCs to indicate, but also refined the sound design for specific audio cues. These design reviews helped us mitigate the challenge of recruiting BLV participants for user studies in a typical iterative design process and enabled us to develop a prototype ready for a user study in about a month.

4 NAVC prototype

The user-centered design process led to a set of six VCCs that were of interest to BLV informants and could be readily detected from a video stream using our computer vision models. Four of them are based on head position: *Attention* (looking forward directly at the conversation partner), *Agreement* (nodding head), *Disagreement* (shaking head), and *Thinking* (looking up, at the ceiling or to the side). Two VCCs are facial expressions: *Happiness* (smiling) and *Surprise* (raising eyebrows with open mouth). We also iterated the design of audio cues to convey these VCCs. We designed a continuous ambient sound when the *Attention* cue is detected that stopped playing when the person looked away. For *Agreement, Disagreement, Thinking, Happiness,* and *Surprise*, we designed different earcons that played when they were detected. The accompanying video figure demonstrates which sound was triggered by each VCC in the prototype.

Some VCCs that were rated highly in the interviews were not implemented. As noted above, cues for *Turn Taking* (hand raise) are not very relevant for one-on-one communication. Other cues, such as *Eye Rolling* and *Confusion* were omitted because we could not get adequate accuracy from our detector. On the other hand, *Thinking* was added based on the discussions in the design review regarding the interpretation of silence, which can occur when someone pauses to think before verbally responding.

It is important to note that the labels for these expressions (e.g., *Thinking* or *Agreement*) are short-hand interpretations of commonly-occurring VCCs that we could detect, but these VCCs may present for other reasons. For example, a person may look away from the screen to consult a notebook, but this does not necessarily mean a lapse in attention. Similarly, a conversation partner may nod along to a conversation to convey amiability and interest rather than clear agreement.

However, we leave this interpretation of the audio cues to the user who is in the best position to contextualize VCCs for a given partner and conversation.

NAVC analyzes the video and speech information directly from the video call to sense these VCCs and generate audio cues in real-time. NAVC has two major components: VCC Detector and Audio Mixer, described in the next sections along with the design of the audio cues. For more details about the implementation of the sensing framework see [19].

4.1 VCC Detector

VCC Detector comprises the computer vision algorithms used to detect the selected VCCs. For the four VCCs based on head orientation (*Attention, Agreement, Disagreement*, and *Thinking*), the Detector first uses Microsoft's Face API [21] to process the video frames, and then applies different specific techniques to classify the VCCs. For each video frame, the API takes the frame as input and outputs the coordinates of face landmarks and the roll and yaw values of the head of a sighted user.

VCC Detector uses the yaw value of the user's face to determine whether he is looking forward at his conversation partner in *Attention*. If looking away causes the absolute value of the yaw value to exceed a threshold, empirically set to 25 degrees for all users, the user is considered to be looking away and not paying attention.

Hidden-Markov Models (HMMs) are used to calculate the probabilities of the *Agreement* and *Disagreement* VCCs [24]. When a user shakes her head in disagreement, the yaw value of her head changes dramatically. Thus, the HMMs use the yaw value to model the *Disagreement* VCC. Similarly, when a user nods his head in agreement, his head moves up and down. To detect this behavior, the Detector calculates the average Y coordinate of major face landmarks. For both *Agreement* and *Disagreement* VCCs, the HMMs continuously output their probabilities.

For *Thinking*, the Detector compares the length of a user's nose with the distance between her eyebrows. When a user looks up towards the ceiling while thinking, the length of her nose will shorten in the images while the distance between her eyebrows remains the same. The ratio of these two distances is used to calculate the probability of *Thinking*.

For the two facial expression-based VCCs (*Happiness* and *Surprise*), the Detector uses pixel-level, CNN-based models, similar to the Emotion API from Microsoft [20]. For each expression-based VCC, the Detector outputs their probabilities for each video frame.

4.2 Audio Mixer

Audio Mixer takes the information from VCC Detector and plays audio cues to BLV users. The sighted user in a video call runs the VCC Detector, which sends messages to the Mixer regarding the state of *Attention* and the probabilities of the other five VCCs.

The Mixer plays an ambient sound (*Attention*) when the user is looking forward at the camera. When the user looks away during a video call, the ambient sound stops until their focus returns. For the other five VCCs, Audio Mixer plays a short earcon when the probability exceeds a preset threshold customized for each conversation partner.

Audio Mixer also changes the volume of the earcons to reflect the strength of the detected VCC. For example, if the threshold for *Agreement* is 40% probability, the Mixer plays its earcon when that VCC Detector threshold is exceeded. The volume of an *Agreement* VCC at 80% probability is louder than that of a 50% probability. The change of volume is intended to convey different levels of VCCs (e.g., slight vs. vigorous nodding).

When a sighted user is speaking, Audio Mixer suppresses the volume of all earcons. Since BLV callers rely heavily on audio for their conversation cues, we try to avoid interfering with the original speech signals from the sighted user. Based on feedback from the design reviews, when a sighted user is speaking, we mute the *Happiness* VCC and play other earcons at 80% volume.

4.3 VCC Audio Design

Having finalized a set of VCCs to detect and convey through NAVC, we iterated on the design of the sounds representing these cues. Updates to the sounds were made based on feedback from the design reviews.

The *Attention* sound was presented as an ambient tone, which played the entire time the sighted discussion partner was looking at the camera. This sound was created from a C major chord in the 3rd octave using a synthesized string pad: individual notes would fade in and out at random intervals, creating a stable and unobtrusive tone which had some movement.

Each of the other VCCs were conveyed through earcons which were between 0.7 and 1.15 seconds long. Short sounds afforded quick notification and easy repetition if an action continued. The earcons for *Agreement* and *Disagreement* were a harmonic string and disharmonic piano timbre, respectively. *Smiling* was a short set of marimba tones. *Surprise* was a plucked electric string timbre with reverberation. *Thinking* was a cycle of ascending tones from a sustained tone synthesizer (e.g., a pad).

4.4 Implementation

We implemented VCC Detector and Audio Mixer as two separate Windows applications in C# using the .NET framework. The sighted user runs VCC Detector on his computer and uses Service Bus [22] to connect to the Audio Mixer running remotely on the BLV user's computer. Audio Mixer is a lightweight application that can easily run on most Windows laptops. NAVC runs independently alongside whichever video calling software users prefer. The architecture of our solution makes it highly scalable so that the VCC metrics can be securely streamed to anyone in real-time during a conversation.

5 USER STUDY

We conducted a lab study evaluating the user experience of NAVC in a video call for people who are BLV. Two computers were set up in separate locations: a laptop at a remote site where BLV participants were recruited, and a desktop in a lab space in another building with sighted experimental confederates who acted as conversation partners. The computers were connected through a Zoom meeting call [31]. Both VCC Detector and Audio Mixer were run on the desktop, and the audio cues were shared through Zoom to enable the experimenters at both sites to have a synchronized experience for study purposes. BLV participants listened to the audio through an external speaker and used the built-in laptop camera, while the experimenter and confederate in the lab space used a webcam for video and listened to audio over headphones to avoid any echo issues. We videotaped the video and audio streams in both locations and instrumented NAVC to log usage.

5.1 Participants

Sixteen participants (denoted as P202-217, and P201 who was a pilot whose data was removed) with diverse levels of visual ability after losing their sight were recruited from two agencies (eight

from each) serving the BLV community in a large city in the US. While some recruits had some residual vision, none of them could detect facial gestures or movements. Seven considered themselves to be low vision while nine considered themselves blind. Ages ranged from 20 to 69, with the median age range of 40-49. Seven participants were female, and six were currently unemployed. On average, they used video calling once a month or less (no video calling experience was required). Each participant received a $100 USD Amazon gift voucher as a gratuity.

Two sighted experimental confederates were hired to participate in discussions with all 16 participants. Both were ages 20-29, male, and currently employed part-time. They both used video calling a few times a week and did not have experience working with people who are BLV. These sighted confederates were recruited to remove some of the variance associated with different conversational partners. Since every BLV participant would be speaking to the same two conversation partners, we could maintain a certain amount of consistency in conversations and tuning of the VCC Detector.

5.2 Procedure

When a BLV participant arrived in the remote site, they first completed the consent process. The experimenter asked demographic questions, including work experience, level of vision, and current video call usage. Then, another experimenter introduced the discussion task to the participant. Each participant engaged in calls that lasted 12 to 15 minutes with each confederate discussion partner: one regular video call (the control) and a video call augmented with NAVC. The orders of video call conditions and confederate discussion partners were counterbalanced.

Before each video call, the participant was prompted to think about a city where they could give restaurant and activity advice. The purported context was meeting someone they had corresponded with online (e.g., on a message board or a remote colleague) that would be coming to visit in person. During each conversation, the participant discovered their conversation partner's food preferences and then they decided together what restaurants and activities both people might enjoy during the visit.

For the NAVC condition, the sighted discussion partner gave a live demo of each sound to the participant, playing each sound at least twice. Participants were encouraged to ask clarifying questions about the sounds during the conversation task. Confederates tried to incorporate the whole range of reactions into the conversation to make sure that all participants were exposed to the full range of VCCs. However, if any VCC seemed particularly rare, one experimenter would sometimes prompt the confederate to portray specific reactions to increase exposure. While this prompting changed the organic flow of the conversation somewhat, it enabled getting feedback on the range of VCCs within the limited time frame of a user study.

During each conversation, the confederates were notified at the five-, ten-, and twelve-minute landmarks. The confederate's goal was to help keep the participant engaged in a low-stress conversation. The confederate was instructed to let the conversation naturally change from one topic to the next, starting from food recommendations and then moving to other likes, dislikes, and activity preferences. To provide all participants a similar experience with a range of topics, an experimenter occasionally provided prompts to reduce the confederate's cognitive load of directing the conversation topic and identified potential topical areas they could ask about during the discussions (shown in Fig. 2), which also had the side effect of causing momentary lapses of the *Attention* cue.

Fig. 2. An experimenter prompting a confederate about a potential conversation topic.

After each video call, participants answered four questions about their perceptions of their conversation partner (e.g., were they open to the participant's suggestions, did they seem distracted during the call). After the NAVC condition, the participants also responded to a standard user experience scale (the Usability Metric for User Experience, UMUX) [12], seven Likert questions about the audio user experience (UX), provided open-ended feedback about the audio experience, and rank ordered the importance of each of the sounds.

After both conditions were completed, the experimenter conducted a semi-structured interview comparing participants' discussion experiences. First, as an indirect preference measure, participants were asked to choose which condition they would prefer if they were to have a third discussion. They were then asked to give feedback on situations when a system like NAVC would be useful and provide suggestions about other VCCs to incorporate into the system.

5.3 Data Analysis

Prior to analysis, any negatively worded statements from the audio UX questions and UMUX scale were reverse coded. Scoring for each scale was completed by summing the scores for each statement (higher scores are better). We analyzed the data using descriptive statistics and Pearson correlations to understand differences with NAVC compared to the control condition. Open and axial coding [7] were used to analyze free response questions after the video calls and at the end of the study. The recorded log files were processed to identify the frequency each auditory cue was triggered.

6 USER STUDY OBSERVATIONS

The user study provided insight regarding how BLV users may engage with NAVC specifically, and similar conversational aids more generally during an actual conversation. Below, we provide details on 1) The usage and accuracy of NAVC in the context of a real conversations between a person who is BLV and a sighted partner; 2) BLV participants' reactions to the NAVC concept; and 3) How BLV participants perceived, understood, and reacted to the specific audio cues for the VCCs.

6.1 Triggering of VCCs in the NAVC Prototype

We analyzed the NAVC log data using Python scripts to extract features of interest. On average, each discussion session lasted 12:30 minutes (SD = 0:46 seconds) and had 76.81 triggered VCCs (SD = 23.76). Since there was little variation in session length, we report statistics per session.

Fig. 3 shows the mean numbers of triggered VCCs per session for each confederate. The average occurrence of each VCC in a session was very different. Across all conversations, and for both confederates, the most triggered VCCs were *Agreement*, *Happiness*, and *Thinking* (in that order). *Disagreement* and *Surprise* occurred less often. We calculated that *Attention* gaps (looking away, causing gaps in attention) occurred 2.75 times per session on average (SD = 1.26) with a mean duration of 4.2 seconds (SD = 1.6). These data reflect that the sighted conversation partners largely paid attention during the calls, and that they triggered VCCs at different rates. Confederate A triggered VCCs more frequently in the study, both in total and for each VCC, than confederate B. While it is not surprising that individuals evoke VCCs at different rates, counterbalancing each confederate with condition and order should have mitigated any effects on the results.

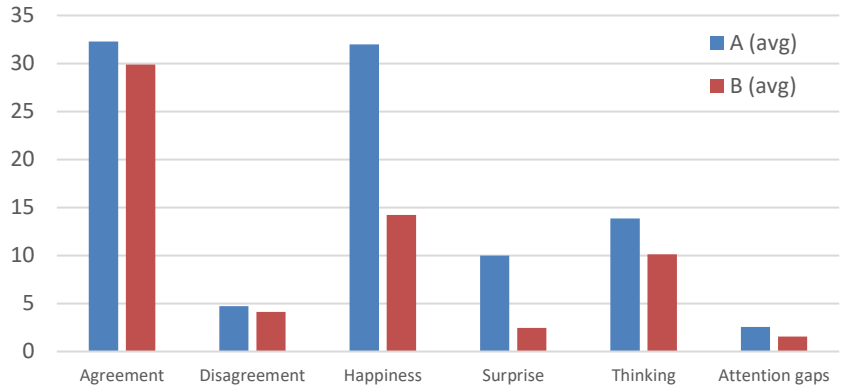

Fig. 3. Mean number of triggered VCCs per session, according to each confederate (A and B).

6.1.1 NAVC Accuracy

We evaluated how accurately the NAVC prototype conveyed VCCs in a conversation. A subset of videos was hand coded to check the accuracy of NAVC's detectors. Two videos for each confederate (four total) were coded, including the first and last discussion they conducted. This selection accounted for learning and familiarization the confederates may have experienced during their week of using NAVC. A coding scheme was generated by two researchers watching a session together to decide how each VCC's correct, false positive, and false negative rates should be marked. One researcher then made three passes through each video: the first without audio (to mark VCCs visually), the second with audio to mark the triggering accuracy of NAVC, and the third to verify accuracy in labeling. These values were used to calculate the sensitivity (percentage of correct triggers that were not false negatives) and positive predictive rate (percentage of correct triggers that were not false positives) for each VCC (Table 2).

Table 2 shows that the accuracy varied across the VCCs between avoiding false positives and negatives. *Attention* was the most accurately recognized VCC. NAVC was better at avoiding false positives than false negatives for *Agreement*, *Disagreement*, and *Surprise*, but was better at avoiding false negatives than false positives for *Thinking* and *Happiness*. Our design strategy, informed by the design reviews, was that it was more important to avoid false positives (have a high positive predictive rate) than miss cues as false negatives (allowing for a lower sensitivity). While our data show room for improvement in the positive predictive rate, it would be especially helpful to focus

on *Thinking* and *Happiness* where the positive predictive rate was lower than the sensitivity, indicating a high potential for false positives. One way to improve the accuracy for all VCCs is more person-specific tuning in recognizing them. While other research prototypes may have shown higher accuracy in detecting conversation cues [1, 27], it is also important to understand how users responded to the way those cues are conveyed.

Table 2. Sensitivity and positive predictive rates per VCC.

VCC	Sensitivity (% not false negative)	Positive Predictive Rate (% not false positive)
Agreement	51%	73%
Disagreement	77%	91%
Thinking	91%	36%
Happiness	93%	72%
Surprise	36%	63%
Attention gap	85%	100%

6.2 User Reaction to the NAVC Concept

Participants' reactions to NAVC were reflected in a variety of factors: choice for a future conversation, ratings of the user experience, and subjective comments. Because our system relies on associating new audio cues with well-established conversation cues, this relatively short exposure to NAVC mostly gives us first impressions on the concept. Indeed, the clearest reaction to NAVC is that people wanted to spend more time experiencing the audio cues to associate them with the VCCs. Nine (out of 16) people explicitly mentioned that they wanted more experience using NAVC to become familiar with the auditory cues. While the participants seemed confident in recognizing the audio cue for each VCC when played by itself during the brief initial training, it was harder to recognize their meaning in the context of a conversation. More cognitive effort was needed to recognize the audio cues while also talking and listening in a conversation, which is different from interpreting the sound effects while watching a movie, where the viewer is not generating conversation at the same time. While we believe that this effort could get easier as people get more familiar with registering the meaning of the audio cues, a longer deployment would be needed to explore this effect. We acknowledge that the short interaction provided by this experiment is a limitation of this research; enabling longer exposure in conversations with people would require developing the NAVC prototype to a state of deploying it in the wild, which we leave to future work.

6.2.1 Choice for Future Conversation

One measure of their reaction to the NAVC concept was which condition they would choose if they were to have a third conversation. Responses were evenly split with eight people choosing each condition. The confederate they spoke with in the NAVC condition did not affect which option they chose, as this was also split exactly evenly.

In explaining their choice, many thought NAVC was distracting from the conversation (12 people) or that it was distracting during their first use but could be useful if there were fewer sounds (4 people).

P202: *I wasn't really in the conversation 'cause I was trying to pay attention to the sounds. But I think after using it after a while, and I get adapted to like, … without thinking basically what the sounds mean, I think it would go like a lot smoother.*

Reflecting on situations when NAVC might be useful, ten people said they would use it with family or friends. Five more explained potential work situations when it could be helpful, and eight people thought it would be useful for meeting new people or dating.

P204: *Pretty much any phone conversation with my sisters. …with my sisters, I used to have to rely a lot more on body language and such things.*

P211: *Definitely in conversations with people that I've never met before, don't really know anything about, don't know as much about how they feel about things or their reactions to things it would be helpful.*

P215: *I think for like work, I would definitely use it, because I have no idea what people are thinking. Unlike thinking about something that people enjoy doing and usually have strong preferences about, with work sometimes it's really hard to read what they really think. So for work I think I would definitely use it.*

Table 3. Mean audio user experience scores and standard deviation (n=16).

Audio UX	Mean Score (SD)
The sounds were interesting	5.06 (1.48)
The sounds were pleasant	4.13 (1.51)
The sounds were easy to understand	5.5 (1.71)
It was easy to match the sounds to their meanings	4.19 (1.80)
It was fun to listen to the sounds	5.13 (1.54)
It was confusing to listen to the sounds	4.5 (1.46)
It was confusing to remember which sounds meant which gesture	4.94 (2.08)
Total Audio UX	30.56 (8.84)

6.2.2 User Experience

After the NAVC condition, participants evaluated the overall usability of the system through UMUX and gave feedback on the auditory cues via audio UX questions (Table 3). The mean UMUX score was 13.38 (SD = 3.22) out of 24, and the average combined audio user experience score was 30.56 (SD = 8.85) out of 49. These scores indicate room for improving the user experience, but also identify some interesting correlations with other metrics.

Pearson correlations explored relationships among audio UX and UMUX scores. The correlation between UMUX and the audio UX was $r=0.57$, $p < 0.02$, suggesting a moderate positive correlation between overall usability of the system and the design of the audio cues. Participants who rated the audio UX positively were also somewhat likely to give a higher UMUX score for

usability of the prototype. No significant correlations were found between participants' level of vision and their usability scores.

The working prototype introduced about a one second lag in triggering the earcons after the VCCs occurred in the conversation, which evoked some contrasting reactions compared to the WOz prototype. In the earlier WOz prototype, people liked how the reactions were instantly correlated with the sounds, whereas people complained about the lag in the user study. This delay bothered participants with some residual vision who could identify body gesture cues like head nods.

P209: *The delay of the sounds – I've already seen him shake his head, and then the sound came afterwards. So, so now I'm thinking 'what happened?'*

6.3 VCC Audio Cue Design Feedback

Reactions to the auditory cue design are based on a number of factors: usefulness, aesthetics, and distinguishability.

6.3.1 Audio Cue Usefulness

Participants rank ordered each of the auditory cues from most (1) to least (6) useful (Table 5). Eleven people ranked all six sounds; the rest ranked only those they felt were the most useful or memorable from the conversation. A lower rank represents a higher level of usefulness.

Table 5. Mean rank, standard deviation, and sample size for each of the audio VCCs.

VCC	Rank (SD)	Sample Size (N)
Attention	2.6 (1.99)	15
Agreement	2.88 (1.11)	15
Happiness	3.33 (2.10)	12
Thinking	3.92 (1.61)	13
Disagreement	4.25 (1.56)	13
Surprise	4.63 (1.12)	11

Through open-ended responses, the three most-liked sounds were *Attention* (8), *Agreement* (4), and *Disagreement* (3). Nine people reported that they did not dislike any of the sounds. One person did not like any of the sounds and found them to be distracting during the conversation. Each auditory cue was mentioned once as being the least favorite and two were listed a few times: *Disagreement* (3) and *Attention* (2).

As shown in Figure 3, the most-triggered cues across all conversations were: *Attention*, *Agreement*, *Happiness*, and *Thinking* (in that order). This ordering matches with the average usefulness rank, suggesting that the amount of familiarity correlated with participants' choices for usefulness of the VCCs. While each cue was designed to be easily differentiable, they were only heard during a short conversation; more experience and exposure to the auditory cues might further influence their overall perceived understandability and usefulness. *Surprise* was the cue which most people reported was hardest to recognize due to confusion with other cues (e.g., *Happiness*). *Surprise* was also infrequently triggered during the discussions, which may have made it harder to recognize, as the participants had less experience interpreting it than the other cues.

Self-reported vision level interacted with the rating of usefulness of the VCCs. Those with low vision (n = 7), in general, reported *Disagreement* (M = 2.20, SD = 1.30) and *Agreement* (M = 2.57, SD = 1.27) to be the most useful cues. Those who considered themselves blind (n = 8) ranked *Attention* (M = 1.63, SD = 0.92) and *Agreement* (M = 2.75, SD = 1.04) as the most useful cues. These individual differences suggest one way that the VCCs indicated could be personalized to fit the abilities of each user.

6.3.2 Notification vs. Soundtrack Aesthetic

Each earcon was designed to naturally convey the VCCs (leveraging patterns found in soundtracks and music). They often functioned as ephemeral notification tones with one exception. The *Attention* ambient sound offered a persistent, background signal layered on top of the conversation. Half of participants reported that it was their favorite: it was the "least intrusive" (P217), "peaceful" (P216), and helpful "in the background" (P205). P211 explained that the biggest distraction was the total absence of sounds when the ambient sound went away: "*Oh no, are they not interested in what I was saying? Or were they not there anymore.*" The *Attention* sound faded into the background and became noticeable when it went away.

P216 and P212 voiced contrasting opinions about how well the notification sounds conveyed their associated VCC:

P216: *I like the thinking sound. I think that goes well with the intent.*

P212: *Initially, looking at camera made really good sense. Nod made sense, Disagree made sense, the rest didn't seem different. Want them to be really different from first two and from each other.*

While the easily understandable audio cues were inspired by soundtrack design, in practice it was distracting for some participants to interpret during the conversation.

P213: *From the context of the conversation some of the sounds weren't matching properly. Or I was just misinterpreting them. So it was throwing me off. They sorta get in the way, or I was beginning to tune them out and not pay attention to them.*

P217 suggested "*Maybe if there was a way to make the volume of the sound different from the speech*," that it might reduce the feeling of distraction or interruption. For this study, we played the audio over an external speaker for the participant; using headsets might allow for lower overall volume of the cues, and better balance of the audio levels may reduce distraction.

6.3.3 Distinguishability of the Earcons

Most participants reported understanding the sounds during the short training given before the NAVC condition, but differentiating them in the context of a conversation was more complicated. Ten people found the sounds to be confusing or distracting when layered on top of the conversation and wanted more time to hear and learn the cues. They felt that more familiarity would help make it easier to tell the sounds apart during the conversations.

P210: *The more the sounds occurred, the easier they were to associate with the emotional state, the facial expression. The less frequent the sound, the harder it was for me to go 'oh yeah that's what that was.' For example, the shaking of the head no, I don't recall hearing that much during the conversation.*

Some participants suggested using a subset of earcons for the VCCs, to make it easier to comprehend them: P216 "*The sounds were distracting, but I do think that maybe with fewer sounds then it kinda would be easier to comprehend or flow with it.*"

Nine participants wanted to choose a particularly relevant subset of VCCs for their level of vision and for their conversation partners. P209: "*Some people will habitually rock [nod] their head when they're talking to you…you need to know that person does that a lot and then you'll just tune that out.*" Selecting a subset of VCCs to indicate should make distinguishability among earcons easier. These observations provide further evidence of how personalizing the VCCs used for each specific user would help NAVC fit a variety of conversational contexts.

Most of the VCCs represented by NAVC are either positive or neutral concepts, making them harder to differentiate (due to similar affect and sound design) than if there were a more balanced split between positive and negative ones. Future work should be done to iterate on positively-valenced auditory cues as well, to make them more distinguishable.

6.3.4 Other Relevant VCCs

Nine people mentioned that negative VCCs, such as frowning, confusion, anger, or sadness, would be helpful, especially for serious conversations or ones where people are less likely to verbally express these reactions. While NAVC's VCC detector has some capability for detecting negative affect, we felt it was not accurate enough to include in the prototype. Participants' desire to detect negative reactions presents an interesting challenge for AI detectors. In social situations, people tend to mask public expressions of negative emotions, making it subtler and more challenging to detect. While it might be easier for an AI to detect a user's anger or frustration when interacting with a computer agent, doing so in a social context with another person is considerably more challenging. Even conversations with people who are BLV, who may not be able to see visual cues of negative affect, are still governed by the social protocol which prompts masking their expression. Reliably detecting negative emotion cues in the context of social conversation needs further research and training.

In this research, we chose the six VCCs to convey in NAVC largely based on the preferences of people who are BLV. In practice, our data show that VCCs will occur at distinctly different frequencies in video call conversations, as shown in Figure 3. VCCs that tend to occur less frequently might need earcons that are notably distinct to help users recognize them despite their infrequent occurrence. While more research in different kinds of conversations is needed, our experience raises another challenge for AI modeling, which depends on large training datasets for accurate detection. While it would be easier to train AI models on frequently occurring VCCs, our research indicates that people may value VCCs that occur less frequently, which can make it more challenging to accurately train AI models.

7 DISCUSSION AND CONCLUSION

Through a user-centered design process, we developed NAVC, which showed promising results in the user study. NAVC demonstrated a novel way to help people who are BLV react to visual conversation cues in video calls and identified design opportunities and challenges in using AI computer vision models to detect VCCs. The NAVC concept does not require any additional equipment at the end user sites, since the AI agent for detecting VCCs can tap into the video and audio already being streamed over the network for the video call. This approach is more convenient than prior research which required adding glasses [1, 18] or a smartphone [25, 27] to sense the conversation partner. NAVC also uses sound to augment verbal communication, leveraging what people are already familiar with in movie soundtracks. This approach avoids issues raised in prior work with spoken feedback [1,27] which can conflict with the conversation, or tactile feedback [18, 25], which requires wearing an additional tactile belt or glove and learning

a new mapping between vibration patterns and conversation cues. Based on our experiences designing and evaluating NAVC, we reflect on our two initial research questions.

While exploring how to detect the VCCs that are important to people who are blind or low-vision (RQ1), we report a ranked list of the kinds of VCCs that our BLV informants believe would be useful in conversation. However, this ranking is highly variable between different people, depending on individual needs and visual capability, and our user study showed that it could even differ according to the conversation partner. Our observations build on prior work that demonstrated the importance of personalization in accessibility [25], and illustrates how users should be able to customize the conversation cues they are interested in according to their abilities and sound preferences and even the attributes of specific conversation partners. While NAVC was adequate as a proof-of-concept system for testing, we believe that higher accuracy is needed before a longer-term deployment. We also learned that participants' interest in detecting negative cues is challenging because negative expressions are often naturally masked in social situations.

In determining whether BLV users would find NAVC useful while conversing over video calls (RQ2), we found that users generally liked and found utility in the idea, but UX scores and other measures suggest that there is much room for improvement in the usability and experience of the prototype. One clear result is that a longer-term exposure is needed to become more familiar with the mapping of sounds to conversation cues and acclimate to their inclusion in the conversation.

One key design question we had was how to convey VCCs without interfering with conversation. We found that people liked the *Attention* sound, which was unique as an ambient sound layered over the conversation, but some of the other notification earcons were hard to distinguish. Several factors can make it hard to distinguish the earcons: sounds could be heard infrequently, VCC detection allowed too many false positives, and VCCs were mostly positive or neutral valence, leading to similarities in the audio cue designs. Furthermore, the NAVC prototype used five ephemeral notification sounds, and one ambient awareness sound. Five notification sounds may be too many to easily distinguish during first use. More research could identify the upper limit of sounds distinguishable in initial conversation and prolonged use. Customization could enable users to pick how many sounds or VCCs they prefer. While sound design is a complex and challenging art, our experiences with NAVC have identified three approaches for making the audio cues more distinguishable.

One design approach is to represent more VCCs through continuous, awareness sounds, especially if it reduces the total number of notification sounds. One design we considered during the second design review was to modulate the *Attention* sound to reflect level of engagement. If the user leaned in closer for more intense engagement, the *Attention* sound would raise to a higher pitch, whereas leaning back with less engagement would lower the pitch. Another possibility is to modify the *Attention* sound to convey *Thinking*, since the act of thinking of a response is a modified state of paying attention. Perhaps *Thinking* could be indicated by modulating the *Attention* sound to a cycle of ascending tones that returns to the low, steady-state tone when the user resumes looking directly toward their conversation partner. It would be interesting to explore how other VCCs could be conveyed more continuously, rather than ephemerally, evoking the inspiration from movie soundtracks.

A second design approach to increase distinguishability is to add more negative valence emotions which could afford dramatically different sounding earcons. The design space of sounds for negative emotions could be broader than what we used for the mostly positive or neutral VCCs implemented in NAVC. While users were interested in representing more negative emotions,

accurately detecting these emotions, especially in a social context where people tend to mask them, is currently a challenge for AI model building.

A third design approach is to reduce the number of VCCs indicated to a smaller subset that is both more distinguishable and more relevant to each user's abilities and interests. Individual differences in visual ability correlated with which conversation cues they wanted help perceiving. People also had different opinions about which sounds were easier for them to distinguish. Enabling users to personalize which VCCs were indicated and what sounds were associated with them could enable creating a tailored sound experience.

While this work looked only at one-on-one conversations between BLV and sighted people, future work could explore how a system like NAVC would work with larger groups or meetings or in conversations where both partners are BLV. Conveying VCCs may also be useful to populations other than those who are BLV. Sighted people could also benefit from auditory conveyance of VCCs during conversations when their eyes are busy (e.g., driving while calling in to a meeting) or doing audio-only calls when they do not have the network connectivity or a device for a video call. For those with Autism Spectrum Disorder, recognizing their conversation partner's emotions, or expressing emotion themselves, may be difficult. Using a system like NAVC could provide these cues real-time during conversations or could be used to train recognition of these cues to prepare for conversations. Future work should also explore other VCCs which may be relevant.

Finally, this research helps to focus future work by identifying relevant conversational cues (e.g., *Thinking*, *Happiness*, other cues for negative affect) for training AI models for better recognition. Beyond recognizing facial expressions based on existing labeled datasets, our research highlights the importance of identifying what cues are important to user populations with diverse abilities. Our experience with sound design should guide future work in audio displays, especially in the context of interfaces to support social interaction. We hope that our experiences combining AI algorithms to detect conversation cues with audio feedback for conveying those cues helps guide future researchers who are interested in developing accessible computer-mediated communication.

ACKNOWLEDGEMENTS

We thank LightHouse for the Blind and Visually Impaired, San Francisco and Vista Center for the Blind and Visually Impaired, Palo Alto, the two agencies who helped us recruit users and participated in our design reviews. We also thank the two confederates who served as the sighted conversation partners in our user study. And we thank the anonymous interview and user study participants who are blind or low vision for their feedback.

REFERENCES

[1] ASM Iftekhar Anam, Shahinur Alam, Mohammed Yeasin. 2014. Expression: A dyadic conversation aid using Google Glass for people who are blind or visually impaired. In Proceedings of the Mobile Computing, Applications and Services (MobiCASE 2014), 57-64. http://dx.doi.org/10.4108/icst.mobicase.2014.257780

[2] Alissa N. Antle, Milena Droumeva, and Greg Corness. 2008. Playing with the sound maker: do embodied metaphors help children learn? In Proceedings of the 7th International Conference on Interaction Design and Children, 178-185.

[3] Ray L. Birdwhistell. 1955. Background to Kinesics. ETC: A Review of General Semantics 13, 1 (Autumn 1955), 10-18.

[4] Ray L. Birdwhistell. 1970. Kinesics and Context: Essays on Body Motion Communication. University of Pennsylvania Press.

[5] Meera M. Blattner, Denise A. Sumikawa, and Robert M. Greenberg. 1989. Earcons and icons: Their structure and common design principles. Human–Computer Interaction, 4, 1, 11-44.

[6] Stephen A. Brewster, Peter C. Wright, and Alistair D. N. Edwards. 1993. An evaluation of earcons for use in auditory human-computer interfaces. In Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems, 222-227.

[7] Juliet M. Corbin and Anselm Strauss. 2008. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory, Sage Publications.

[8] Ádám Csapó and György Wersényi. 2013. Overview of auditory representations in human-machine interfaces. ACM Computing Surveys (CSUR) 46, 2, 19.

[9] Alistair D. N. Edwards. 1989. Soundtrack: An auditory interface for blind users. Human–Computer Interaction 4, 1, 45-66.

[10] Mirjam Eladhari, Rik Nieuwdorp, and Mikael Fridenfalk. 2006. The soundtrack of your mind: mind music-adaptive audio for game characters. In Proceedings of the 2006 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology.

[11] Catherine S. Fichten, Darlene Judd, Vicki Tagalakis, Rhonda Amsel, Kristen Robillard. 1991. Communication Cues Used by People with and without Visual Impairments in Daily Conversations and Dating. Journal of Visual Impairment and Blindness 85, 9, 371-78.

[12] Kraig Finstad. 2010. The usability metric for user experience. Interacting with Computers 22, 5, 323-327.

[13] Stavros Garzonis, Simon Jones, Tim Jay, and Eamonn O'Neill. 2009. Auditory icon and earcon mobile service notifications: intuitiveness, learnability, memorability and preference. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1513-1522.

[14] William W. Gaver. 1986. Auditory icons: Using sound in computer interfaces. Human-Computer Interaction 2, 2, 167-177.

[15] Thomas Hermann, Alexander Neumann, Sebastian Zehe. 2012. Head Gesture Sonification for Supporting Social Interaction. In Proceedings of the 7th Audio Mostly Conference (AM 2012), 82-89 http://dx.doi.org/10.1145/2371456.2371469

[16] Neil Hillman and Sandra Pauletto. 2014. The Craftsman: The use of sound design to elicit emotions. The Soundtrack, 7, 1, 5-23.

[17] Ellen A. Isaacs, John C. Tang. 1994. What video can and cannot do for collaboration: A case study. Multimedia Systems 2, 2, 63-73.

[18] Sreekar Krishna, Vineeth Balasubramanian, Sethuraman Panchanathan. 2010. Enriching social situational awareness in remote interactions: Insights and inspirations from disability focused research. In Proceedings of the 18th ACM international conference on Multimedia, 1275-1284. http://dx.doi.org/10.1145/1873951.1874202

[19] Daniel McDuff, Kael Rowan, Piali Choudhury, Jessica Wolk, ThuVan Pham, Mary Czerwinski, M. 2019. A Multimodal Emotion Sensing Platform for Building Emotion-Aware Applications. arXiv preprint arXiv:1903.12133.

[20] Microsoft Azure. 2018. Emotion API. Retrieved September 14, 2018 from https://azure.microsoft.com/en-us/services/cognitive-services/emotion/

[21] Microsoft Azure. 2018. Face API. Retrieved September 14, 2018 from https://azure.microsoft.com/en-us/services/cognitive-services/face/

[22] Microsoft Azure. 2018. Service Bus. Retrieved September 14, 2018 from https://azure.microsoft.com/en-us/services/service-bus/

[23] Gregory Mone. 2018. Feeling Sounds, Hearing Sights. Communications of the ACM, 61,1 15-17. http://dx.doi.org/10.1145/3157075

[24] Carlos Morimoto, Yaser Yacoob, and Larry Davis. 1996. Recognition of head gestures using hidden Markov models. In Proceedings - International Conference on Pattern Recognition. 461–465. https://doi.org/10.1109/ICPR.1996.546990

[25] Sethuraman Panchanathan, Troy McDaniel. 2014. Person-centered accessible technologies and computing solutions through interdisciplinary and integrated perspectives from disability research. Universal Access in the Information Society 14, 3, 415-426. https://doi.org/10.1007/s10209-014-0369-9

[26] Shi Qiu, Jun Hu, and Matthias Rauterberg. 2015. Nonverbal Signals for Face-to-Face Communication between the Blind and the Sighted. In Proceedings of International Conference on Enabling Access for Persons with Visual Impairment, 157-165.

[27] AKMMahbubur Rahman, ASM Iftekhar Anam, Mohammed Yeasin. 2017. EmoAssist: emotion enabled assistive tool to enhance dyadic conversation for the blind. Multimedia Tools and Applications 76, 6, 7699-7730. http://dx.doi.org/10.1007/s11042-016-3295-4

[28] David Sonnenschein. 2001. Sound Design: The Expressive Power of Music, Voice and Sound Effects in Cinema. Michael Wiese Productions.

[29] Bruce N. Walker, Mark T. Godfrey, Jason E. Orlosky, Carrie Bruce, & Jon Sanford 2006. Aquarium sonification: Soundscapes for accessible dynamic informal learning environments." In Proceedings of the International Conference on Auditory Display (ICAD 2006), 238-241.

[30] Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. 2018. A Face Recognition Application for People with Visual Impairments: Understanding Use Beyond the Lab. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, p. 215. ACM, 2018.
[31] Zoom Meeting. 2018. Retrieved August 21, 2018 from https://zoom.us/.

Received April 2019; revised June 2019; accepted August 2019.