Long-Term, in-the-Wild Study of Feedback about Speech Intelligibility for K-12 Students Attending Class via a Telepresence Robot

Matthew Rueben
New Mexico State University
Las Cruces, New Mexico, USA
mrueben@nmsu.edu

Mohammad Syed
San Jose State University
San Jose, California, USA
mohammad.syed@sjsu.edu

Emily London
University of Southern California
Los Angeles, California, USA
elondon@usc.edu

Mark Camarena
Eunsook Shin
markcama@usc.edu
eunsooks@usc.edu
University of Southern California
Los Angeles, California, USA

Yulun Zhang
Timothy S. Wang
yulunzha@usc.edu
wangtimo@usc.edu
University of Southern California
Los Angeles, California, USA

Thomas R. Groechel
Rhianna Lee
Maja J. Matarić
groechel@usc.edu
rhiannal@usc.edu
mataric@usc.edu
University of Southern California
Los Angeles, California, USA

ABSTRACT

Telepresence robots offer presence, embodiment, and mobility to remote users, making them promising options for homebound K-12 students. It is difficult, however, for robot operators to know how well they are being heard in remote and noisy classroom environments. One solution is to estimate the operator's speech intelligibility to their listeners in order to provide feedback about it to the operator. This work contributes the first evaluation of a speech intelligibility feedback system for homebound K-12 students attending class remotely. In our four long-term, in-the-wild deployments we found that students speak at different volumes instead of adjusting the robot's volume, and that detailed audio calibration and network latency feedback are needed. We also contribute the first findings about the types and frequencies of multimodal comprehension cues given to homebound students by listeners in the classroom. By annotating and categorizing over 700 cues, we found that the most common cue modalities were conversation turn timing and verbal content. Conversation turn timing cues occurred more frequently overall, whereas verbal content cues contained more information and might be the most frequent modality for *negative* cues. Our work provides recommendations for telepresence systems that could intervene to ensure that remote users are being heard.

CCS CONCEPTS

• **Human-centered computing** → *Natural language interfaces*; *Field studies*; • **Applied computing** → *E-learning*.

KEYWORDS

telepresence, mobile remote presence, spoken dialogue systems, K-12 education

1 INTRODUCTION

Telepresence robots are a promising option for K-12 students who are homebound and must attend class remotely [14, 37, 51]. A telepresence robot, also called a mobile remote presence (MRP) system [27], allows the user to navigate and perform two-way videoconferencing in a remote environment [25]. Research has explored the use of telepresence robots in a wide variety of contexts including homes [1, 48], hospitals [50], offices [27, 49], museums [38], and professional conferences [36]. In 2001, in-school work showed the promise of telepresence robots for helping remote students to keep up with course material and contribute to classroom discourse [14, 51]. Subsequent research demonstrated that telepresence robots can minimize the effects of physical separation during school absences of K-12 students [37]. Multiple studies have also shown ways for telepresence robots to provide better experiences than other distance learning technologies, albeit in university classrooms [2, 17, 19].

Hearing each other clearly is a problem for users of telepresence systems, especially in the K-12 classroom setting. In a field study
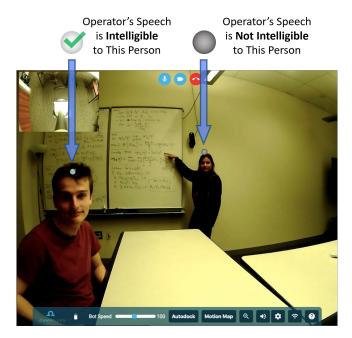
Figure 1: Screenshot of the Ohmni telepresence robot operator interface with enlarged views of the elements added by the speech intelligibility feedback system. The model has estimated that the closer person (seated at left) can understand what the operator has said, but the farther person (standing at center) cannot.

of a telepresence robot in a K-12 classroom, designers operating the robot mentioned audio and video communication quality as a key challenge [7]: it took a great deal of effort to hear conversations over distance or through ambient noise, and the robot's volume setting was often at the wrong level. Speech intelligibility in classrooms is subject to classroom acoustics [3] where voices can interfere with one another, making the target voice harder to hear [5]. The student's home environment also impacts the intelligibility of their speech; not all students have access to a quiet space with a good Internet connection. The student's speech intelligibility also depends on the quality of their microphone and the robot's speaker. Telepresence robots have had audio-related issues in other applications as well; in the workplace, for example, operators can inadvertently speak too loudly [27, 39].

Specifically, it is difficult for the telepresence robot operator to tell how loudly they should speak. Their voice is competing with other sounds in the classroom that are conveyed to them indirectly through the audio connection, and may even be filtered out as background noise. Additionally, viewing the distorted feed from a wide-angle robot camera on a relatively small screen makes it challenging for the robot operator to estimate the size of the classroom and the distance to a potential listener. As a result, the telepresence robot operator's speech might not be appropriately loud and well-enunciated for the classroom context, and they must rely on their listeners for signals about whether they are speaking at the appropriate volume to be heard and understood.

Telepresence systems could help the robot operator to speak at an appropriate volume via several types of interventions. Sidetone, a technique used in the past in telephones [26], appears not to work as well with telepresence robots [39]. Dynamic volume control techniques [48] such as audio compression and ambient noise compensation [39] would require adjusting to the audio characteristics of the operator's speech and all the background sounds that could occur in their home. *We evaluated a different approach: giving the robot operator visual feedback about their speech intelligibility.* Our system [43] displays visual feedback on the operator's interface about multiple potential listeners based on a data-driven model of speech intelligibility. This system was deployed with two homebound K-12 students for several weeks—the first deployment of a speech intelligibility feedback system in the K-12 context. We present findings from subjective interview data and objective behavioral measures.

To provide the visual feedback described above, the telepresence system needs an estimate of the robot operator's speech intelligibility to a particular listener. A listener comprehension model providing such an estimate should include comprehension cues from the listener. People monitor their own comprehension levels from an early age [18] and use various types of verbal [21, 34], prosodic [21, 22], and nonverbal [35, 44] cues (such as gaze and facial expressions) to signal incomprehension. *We performed the first analysis of the types and frequencies of comprehension cues given by listeners to telepresence robot operators in K-12 classrooms.* We collected data from four in-the-wild deployments, including the two mentioned above.

This work makes two complementary contributions—an evaluation of a speech intelligibility feedback system and an analysis of listeners' comprehension cues—as part of a larger study of how telepresence robots can help homebound K-12 students stay connected to their classrooms [7, 8, 15–17, 43]. Both contributions are firsts for telepresence in the K-12 classroom, and our findings also provide insights for the other telepresence applications mentioned above: homes, hospitals, workplaces, museums, and conferences.

2 RELATED WORK

2.1 Feed-Forward Models of Listener Comprehension

Dialogue systems need to estimate multiple variables about a user during a conversation, including their goal [9], the likelihood that they are about to end their turn [46], and their emotional state [31]. Models producing those estimates often use multimodal data [6].

A speech intelligibility feedback system requires a model for estimating listener comprehension. Several such models have been developed. In the context of human-robot interaction, Mead and MatariÄ‡ [32, 33] developed a model of both listener comprehension and typical response volume based on user distance from the robot. Hayamizu et al. [20] and Takahashi et al. [47] developed additional models of listener comprehension that included ambient noise in addition to distance. Comprehension might also depend on the listener's head orientation and prosodic features of the speaker's voice [10, 11].

This work presents an evaluation of a simple feed-forward listener comprehension model we developed specifically for K-12

students that uses both distance and ambient noise [43]. The model was trained using data collected from the same telepresence robot system we used in this work, capturing its microphone, speaker, and audio transmission characteristics. The model does not estimate to whom the robot operator is speaking; instead, listener comprehension is calculated for each potential listener in the classroom.

2.2 Speech Intelligibility Interventions for Telepresence Systems

There are several ways to intervene when a listener comprehension model determines that a listener is unlikely to have understood the telepresence user's utterance.

One approach is to automatically control the robot's output volume. Hayamizu et al. [20] and Takahashi et al. [47] trained two listener comprehension models to control a telepresence robot's output volume based on distance to the listener: one for "normal conversation" and another for "secret talk", so others besides the listener could not hear. Paepcke et al. [39] used audio compression to soften loud utterances, but this approach is parameterized by the expected duration of the sounds needing attenuation so its effectiveness would depend on the noise characteristics in each operator's home. EscottÃ¡ and Beccaro [13] used fuzzy logic to increase the system's volume in response to background noise.

A telepresence system can also provide enhanced audio immersion to the operator. Liu et al. [29] provided binaural auditory scene reproduction to the operator using the operator's head orientation and the 3D locations of sound sources. Paepcke et al. [39] evaluated sidetone—playing the robot operator's voice back to them with a slight delay—on telepresence robots and found that it did cause operators to talk more quietly, but the effect was smaller than in studies of audio-only interactions. They hypothesized that the visual component of the telepresence system interferes with the desired effect of sidetone.

A third approach is to provide visual feedback to the operator. Paepcke et al. [39] evaluated using graphical sound pressure level (SPL) feedback in the user interface, but robot operators tended to overlook this feedback. Kimura et al. [24] calibrated a similar feedback system to the distance to the listener, and additionally projected the same feedback onto the floor where the listener could see it.

This work presents an evaluation of an intervention described by Rueben et al. [43]: an icon positioned above the face of each potential listener that indicates whether the robot operator's speech is loud enough to be intelligible to them. In contrast to the visual feedback used by Paepcke et al. [39] and Kimura et al. [24], our system displays feedback for each potential listener, and does so above each of their faces in the video image for visual salience. Unlike the prototype by Kimura et al. [24] however, our system gives no feedback to the listener.

2.3 Listener Comprehension Cues

Models for estimating listener comprehension could be improved by detecting cues from the listeners about their comprehension levels and using that information to update the model's predictions.

Research on establishing mutual understanding in human-robot interaction [23]—i.e., establishing *common ground* via *grounding*—has primarily focused on enabling the robot to understand the human [40, 42, 45]. In conversations via a telepresence robot, however, the opposite is also important: the robot should be gathering evidence about whether its operator's utterances are being understood by the remote listeners who are physically co-present with the robot. Roque and Traum [41] developed a grounding model for a virtual agent equipped with a spoken dialogue system; for evidence of grounding their model used cues that included repeating an utterance, using the provided information, and failing to respond.

Other types of listener cues have been used to detect speech recognition failures (i.e., incorrect speech transcriptions) in spoken dialogue systems. Some failure detection models use the content of the user's response [21, 34], including repetition of the same utterance, disconfirmations using "no" or "not", and making corrections by repeating an utterance with modifications [34]. Other work uses prosodic features such as tempo and duration of utterances and pauses [21, 22]. Face-to-face interactions with embodied virtual agents and physical robots offer additional cues such as gaze, gestures (e.g., head nods), facial expressions, and body motion [35, 44]. Some systems have combined multiple modalities of information—e.g., dialogue and gestures [30] or prosody and sentiment analysis [28].

Our study is the first to catalog the types of multimodal comprehension cues given by listeners to telepresent students in the K-12 classroom context. We estimate which cue modalities are most common in this context to inform the design of telepresence systems that can provide timely interventions when the remote student is not being heard.

3 MODEL-DRIVEN FEEDBACK ABOUT SPEECH INTELLIGIBILITY

The development of our speech intelligibility feedback system has been described in detail by Rueben et al. [43]. The system provides the robot operator with real-time visual feedback designed to minimize distraction from the ongoing conversation. Determining to whom the robot operator is attempting to speak is avoided by providing feedback for every potential listener in the camera's field of view. As shown in Figure 1, feedback is provided in the form of a small icon positioned over the top of each detected face in the camera frame.

The model operationalizes speech intelligibility as speaking volume appropriateness to determine which of the two icons to display: intelligible (loud enough) or not intelligible (not loud enough). The model was trained on speech intelligibility data from three participants in a laboratory setting. Two variables were controlled and systematically varied: listener-robot distance and ambient noise volume. The following planar model was found to be a good fit for the training data:

$$MinimumIntelligibleVolume = 0.01168 * AmbientNoise + 6.90635 * ln(Distance) + 49.40575 \tag{1}$$

The model outputs the minimum intelligible speaking volume (in decibels) for a given ambient noise level (in the native units of

the robot's microphone) and listener-robot distance (in meters). We used an additional model trained on data collected in the lab to estimate the listener-robot distance from the listener's face height in pixels.

4 LONG-TERM DEPLOYMENTS WITH HOMEBOUND K-12 STUDENTS

This work is part of a larger project studying how telepresence robots can reduce the impact of extended school absences of homebound K-12 students. We deployed a telepresence robot system with four homebound K-12 students, referred to as P1, P2, P3, and P4, who used the robot to attend classes for several weeks while we collected a multimodal dataset. In this section we present the data collection process and then describe our multimodal dataset.

4.1 Data Collection

To obtain meaningful data beyond novelty effects in both homes and classrooms, we collected data for at least two weeks for each participant (see Table 1 for exact durations) to assess how student operators received feedback about speech intelligibility and how students and teachers in the operators' classrooms provided it.

Recruitment was extremely challenging because it required finding participants who were homebound from recruitment through to the end of the deployment. Nevertheless, it was important to capture data that reflected the diverse reasons students become homebound, along with their associated accessibility challenges and other individual differences.

The telepresence platform was the Ohmni robot by OhmniLabs; we enhanced the web-based interface with several modifications, including our speech intelligibility feedback system. Participants accessed the interface via Macbook Air laptops we deployed to their homes. The interface included a slider for adjusting the volume the robot used to play the participant's speech. Participants either used earbud headphones with built in microphones or the laptop's built-in microphone and speakers.

Participants and their families were trained in their homes on how to use the telepresence robot, including the speech intelligibility feedback system. P1 and P3 needed family members to set up and log in for them, while P2 and P4 could do so independently. After the training, the robot was removed from the participant's home and placed in their classroom. The homebound students then attended one or more classes using the robot, subject to their class schedules and health constraints.

Various technical issues arose, especially at the start of each deployment. P1, P2, and P4 experienced times when they could hear the classroom but could not be heard; for P1 this lasted the first few days of the deployment. Some audio-visual recordings were not started or ended correctly, causing data to be lost. For example, one of the two recordings was missing for 2 hours and 15 minutes of the 14 hours of recordings that were annotated.

Most participants attended entire class sessions via the robot, except for P3, who usually attended for at most an hour, for health reasons. P2 and P4 attended multiple classes via the robot (see Table 1). Most participants also had some telepresence interactions outside the classroom, e.g., en route to the next classroom.

At the end of the deployment period, the robot was retrieved from the school and the laptop from the participant's home. Participants, their families, and the teachers were compensated for participating in the study. Classmates were given a small gift for their participation. The project was approved by the Institutional Review Board (IRB), study number UP-18-00129.

Table 1 provides the the details of the four participants' deployments. The speech intelligibility feedback system was developed during the P1 and P2 deployments and subsequently deployed with P3 and P4. Comprehension cues were annotated and analyzed for all four participants. P1 and P3 were difficult to understand: P1 was on a ventilator and spoke in brief bursts, and P3 had significant difficulty enunciating due to their medical condition.

4.2 Multimodal Dataset

We collected multiple types and modalities of data.

Semi-structured interviews were conducted with participants and their caregivers at least once per week. P3 and P4, to whom we deployed the speech intelligibility feedback system, were asked about it at every interview. At the final interview, they were also asked for ways to improve the feedback system, e.g., by making it clearer or easier to use.

Behavioral data were collected whenever the participant was logged into the robot via audio and video recordings from two perspectives. The camera and microphone on the homebound participant's laptop recorded the homebound participant's face and voice as presented to the classroom by the robot. The robot's camera and microphone recorded the classroom as it was presented to the homebound participant on their laptop screen. It was often difficult to understand what people in the classroom were saying, due to ambient noise or distance from the robot.

Sixty-seven hours of audio-visual recordings were collected: 14 hours from P1, 10 hours from P2, 15 hours from P3, and 28 hours from P4.

5 DATA ANALYSIS

The audio and video recordings described in the previous section were manually annotated for comprehension cues sent to the participant from their classmates and teachers. We also analyzed the recordings for speech intelligibility feedback system use and its effect on participants' use of the robot's volume slider.

5.1 Annotation Personnel and Process

The annotation team annotated a subset of the audio-visual recordings because of the time-intensive nature of our annotation process: 14 hours and 20 minutes were annotated, or about 21% of the total duration. Of these, 2 hours and 20 minutes (about 16%) were co-annotated—i.e., annotated by both annotators to evaluate inter-rater reliability. Annotation assignments were distributed approximately evenly across the four participants and then across the weeks within each participant's deployment. Co-annotations were also distributed approximately evenly, except that none were assigned for P1, who spoke very seldom and therefore received very few comprehension cues.

We used the free annotation software ELAN. The two audio-visual recordings for each class session—of the participant from

Table 1: Details of the four deployments.

Participant	P1	P2	P3	P4
Duration	5 weeks	2 weeks	8 weeks	5 weeks
Calendar Days	Mar 26–Apr 17, 2019	Jun 3–Jun 13, 2019	Oct 29–Dec 17, 2019	Feb 21–Mar 13, 2020
Grade	High School	High School	High School	Middle School
Classrooms	Spanish	History, English	Intensive Studies	Math, Science, English, History, Lunch
Feedback System?	No	No	Yes	Yes
Days Attended	10 days	7 days	21 days	13 days
Recorded Time on Robot	14 hours	10 hours	15 hours	28 hours
Difficulties Using the Robot?	Yes: physical and speech	None	Yes: physical and speech	Minor physical

their laptop and of the classroom from the robot—were synchronized and displayed simultaneously in ELAN for the annotators. Additionally, a volume analysis of the participant's audio and an optical flow analysis of the robot's video feed were used to eliminate sections of the recording that did not contain participant speech or robot movement, respectively.

The recordings were annotated for participant speech, robot movement, comprehension cues, and classroom context. The annotation team consisted of two annotators, two annotation supervisors and protocol developers, and a team leader. For training, the annotators participated in protocol development and annotated test recordings. The two annotators' work was compared and then reviewed by both supervisors.

The annotation supervisors continued to review all annotations after the annotators completed their training and began annotating their assigned recordings. The annotators then re-watched any sections of video flagged by a supervisor and revised the annotation file in response to the comments. The responses were reviewed to ensure that all comments were properly addressed.

5.2 Annotation of Comprehension Cues

A comprehension cue was defined as any action or inaction by a potential listener (the cuer) in the classroom from which a reasonable robot operator would gather information about whether (or to what extent) that person heard them. We intended for the annotators to notice more cues than the participants themselves did, in order to catalog as many cues as possible. In addition to a text description of each cue, the annotators recorded: the cuer (teacher, TA, or classmate); whether the cue was verbal or nonverbal; whether the cue was positive (i.e., indicating that the cuer understood the participant), negative (i.e., indicating that they did not), or mixed (e.g., partial understanding); and an initial cue category that was later replaced by a final cue category, subcategory, and modality.

5.3 Development of Cue Categories and Labeling by Modality

After the annotation process was complete, cue categories and subcategories were developed to fit the dataset of cues. The cue categories were developed iteratively between two of the authors: categories were added, merged, or split until every cue fit in at least one category and categories had minimal conceptual overlap.

Each cue was assigned to only one category; cues that fit into multiple categories were assigned to the category that contained the most information about listener comprehension. Next, all the cues assigned to each category were reviewed again together, and some but not all of the categories were split into subcategories to better fit these cues. The subcategorization process resulted in the merging, splitting, deleting, and adding of categories as well as subcategories.

We specified which modality of data would be needed to detect and interpret the cues in each subcategory. Focusing on modalities kept the analysis agnostic to different choices of algorithm or software implementation. The first three top-level modalities we chose were audio, video, and robot movement and settings. Audio was subdivided into verbal and nonverbal voice data (no non-vocal audio cues were found in this study). Next, the verbal data submodality was further subdivided into cues based on key words or phrases and cues based on one or more sentences. Video was subdivided into four submodalities: (1) head, eye, and face (e.g., head movements, eye gaze, and facial expressions); (2) hands; (3) other body pose (requiring the trunk or legs, such as leaning towards the robot); and (4) activity (i.e., higher-level actions such as walking or picking something up).

The fourth and last top-level modality was conversation turn timing. While turn-taking in spoken dialogue is often analyzed using audio or video data, we chose to consider it as a separate modality because it focuses on the timing of the conversation instead of its content.

6 EVALUATION OF THE SPEECH INTELLIGIBILITY FEEDBACK SYSTEM

This section presents our findings about the speech intelligibility feedback system from data collected in deployments to P3 and P4.

6.1 Interview Findings

Both P3 and P4 reported that the speech intelligibility feedback system did not distract from participating in classroom activities, but did not use it very often for several reasons. One was misunderstanding: as late as Week 6, P3's parent (who helped P3 use the robot's interface) reported an incorrect belief that the feedback indicates who is speaking in the classroom. Also, both P3 and P4

reported a belief that the feedback system increased network latency. This sometimes led P3 and their parent to leave the feedback system turned off.

Nonetheless, both participants felt the speech intelligibility feedback system was well-suited for certain applications. P3's parent said it could help to ensure that the teacher could hear P3 over other conversations in the classroom—a common problem for P3, who spoke a lot and often wanted the teacher's attention. P4 wanted to use it to have small, private conversations with friends without being overheard by teachers or other students. Late in Week 6, P4 successfully used the feedback system in this way during a conversation with a classmate they were befriending. P4 also commented that the system could be useful for giving class presentations.

6.2 Behavioral Findings

Analysis of the video recordings revealed that adjustments to the volume slider in the robot's web interface by participants or their family members were very rare. P3 and their family members adjusted the volume slider 8 times and viewed it 2 additional times in the 10 hours of recorded video that we analyzed (out of 15 hours total). P4 adjusted the volume slider 9 times, 6 of which were unrelated to any utterance from or to P4, in the 23-½ hours of recorded video that we analyzed (out of 28 hours total).

We did not find evidence that the speech intelligibility feedback system encouraged more volume slider adjustments. In fact, for both participants there were fewer adjustments or views per hour when the feedback system was on than when it was off. P3 had 6 adjustments or views in the 4-½ hrs of video we analyzed when the feedback system was off and 4 in the 5-½ hrs when it was on. P4 made 8 adjustments in the 17-½ hrs of video we analyzed when the feedback system was off and just 1 in the 6 hrs when it was on.

7 ANALYSIS OF COMPREHENSION CUES

As described in Section 5.1, 14 hours and 20 minutes of audio-visual recordings from all four participants were annotated for comprehension cues given by listeners. In this section we present (1) the agreement levels between the two annotators and results from (2) the initial annotation, (3) categorization, and (4) grouping by modality of the 774 cues we identified.

7.1 Inter-Annotator Agreement

Of the 774 total cues, 231 were in the sections of the recordings that were annotated by both annotators so we could compare their annotations. Only 74 (32%) of these 231 cues were recorded by both annotators; 89 were only recorded by the first annotator and the remaining 68 were only recorded by the second annotator. This low agreement is not problematic if each annotator was attentive to different types of cues, thereby together finding more cues than separately. It might also suggest, however, that annotating for all types of cues at once was too difficult for the annotators.

There is evidence that our annotators had lower agreement when it was difficult to tell what actions and inactions were related to a participant's utterance. This difficulty was made worse by any network latency causing delays between participant utterances and responses from the classroom, which also could have made it difficult for the annotators to assess conversation turn timing.

Additionally, many of P3's utterances were very difficult to understand and spoken to nobody in particular or out of turn (e.g., when the teacher was speaking to the whole class). These factors often made it difficult to tell whether an action or inaction by a teacher or student was related to an utterance by P3. In the co-annotated sections of video for P3, 41 cues that we later placed in the "Beginning to Speak at an Appropriate Time" category (see Table 3) and 19 that we placed in the "Silence When a Response was Warranted" subcategory were recorded by the first annotator but not the second annotator. These 60 cues account for about two thirds of the cues that only the first annotator recorded in the co-annotated video sections across all participants. Inspection of these cues reveals that it is often difficult to tell whether the cuer's appropriately timed utterance or lack of response is related to the most recent utterance by P3.

Subcategories that more obviously met our definition of a comprehension cue (see Section 5.2) had much higher levels of agreement, such as "Repeating What the Participant Said" (7/8 = 88% agreement) and "Answering the Participant's Question" (7/9 = 78% agreement). Agreement was also high for labeling cues as positive, negative, mixed, or of uncertain valence: 60 (81%) of the 74 cues recorded by both annotators were given the same label.

Utterance counts—used in the next analysis to calculate cue frequency as average cues per utterance—also had relatively high agreement: 318 (80%) of the 398 utterances recorded by either annotator were recorded by both. Most of the disagreement was for P3, who often laughed or made other sounds that may or may not have warranted a response.

7.2 Cue Frequency; Positive and Negative Cues

Table 2 summarizes the comprehension cue annotation results before categorization. The large majority of the cues were for utterances by P2 or P3, who spoke much more often than P1 or P4 did. Compensating for this difference revealed that we did not notice more than 1 cue *per utterance* on average for any of the four participants.

We noticed more positive cues (indicating comprehension) than negative cues (indicating a lack of comprehension) for all four participants. However, P1, who had audio issues for several days, and P3, who was difficult to understand and often spoke to the entire class out of turn, had higher proportions of negative cues than P2 or P4 did.

	P1	P2	P3	P4	Total
Total	45	256	378	95	774
per Utterance	0.79	0.81	0.47	0.96	0.61
Positive	19	182	196	60	457
Negative	17	49	148	27	241
Mixed	5	8	27	8	48
Uncertain or Blank	4	17	7	0	28

Table 2: Summary of annotated comprehension cues.

7.3 Final Categories and Subcategories for Cues

Table 3 shows the categories and subcategories we developed to fit the cues. There were 15 categories and 31 subcategories. Seven of the categories had no subcategories, so there were 38 cue types (i.e., 7 categories without subcategories and 31 subcategories) that were labeled by modality. Some subcategories contained just a few cues, and two contained none—"Cuer Action Does Not Match Participant's Utterance" and "Reducing Classroom Noise to Hear Better"—but were included because they are possible.

7.4 Cue Modalities

For both positive and negative cues, the most frequent cue modality was conversation turn timing, followed by audio, and then video; the least frequent was robot movement and settings (see bottom of Table 3).

Conversation turn timing cues may have been the most frequent by an even wider margin if we had allowed each cue to be placed in more than one subcategory. The cues in every subcategory in the audio modality were likely to also belong in the "Beginning to Speak at an Appropriate Time" subcategory—i.e., to include information about listener comprehension in their timing as well as in their verbal content. As described in Section 5.3, we categorized these cues according to their verbal content because it contained more information about listener comprehension than timing did.

If P3 is excluded, however, then the audio modality had the most *negative* cues, which are important for indicating when the telepresence system needs to intervene. As noted earlier, P3 often spoke unintelligibly or out of turn, prompting their teacher and classmates to ignore these utterances. This inflated the number of negative conversation turn timing cues: 105 (85%) of the 124 *negative* conversation turn timing cues were categorized as "Silence When a Response was Warranted" or "Cuer's Utterance Ignores Participant's Utterance" in response to an utterance from P3. The audio modality had the most negative cues for each of the other three participants, totaling 48 (52%) of the 93 negative cues they received.

8 DISCUSSION

8.1 Interventions for Telepresence Operators

Our participants may have simply spoken more loudly or softly as the need arose, removing the need to frequently adjust the robot's volume slider. The characteristics of the audio system may have still needed adjustment, however, for participants to achieve the desired range of volumes without difficulty. Personalization might also be needed—e.g., students with difficulties speaking loudly (like P1) or clearly (like P1 and P3) could have their voices amplified. The best calibration for a particular student might include nonlinear volume scaling, equalization, and compression.

Future work may address the development of a system for such calibration, to be used less frequently by adjusting more parameters than a single, master volume slider does. Related research would study how telepresence operators adjust their *speaking* volume while interacting in the classroom and using other interventions such as sidetone.

Turn	Verbal (keywords)	Verbal (sentences)	Nonverbal	Head, Eye, Face	Hands	Other Body Pose	Activity	Robot Movt, Settings	Cue Categories and Subcategories	Total	Positive Cues	Negative Cues
									Cue Modalities			
v									**Beginning to Speak at an Appropriate Time**	249	*230	2
									No Response	159	7	*141
v									Silence When a Response was Warranted	134	5	*121
		v							Cuer's Utterance Ignores Participant's Utterance	25	2	*20
	v								**Repeating (To Others or Participant) What Participant Said**	18	18	0
									Talking About the Audio	22	6	15
	v								Talking About Whether the Participant's Audio is Intelligible	15	5	9
	v								Talking About Whether the Cuer's Audio is Intelligible	5	0	5
	v								Talking About the Delay (i.e., network latency)	2	1	1
									Talking About Robot Settings	16	2	12
	v								Talking About Volume, Mute, or Other Problems	9	1	7
		v							Implies that a Problem Is Being Fixed	7	1	5
									Asking For Clarification	47	4	29
	v								Saying, "What?" or, "Huh?"	28	1	22
	v								Asking Participant to Repeat	5	0	4
		v							Repeating the Participant's Utterance for Confirmation	6	1	0
		v							Repeating Their Own Utterance	5	0	2
		v							Asking a Question	3	2	1
									Appropriate Response	132	122	0
	v								Backchanneling And Short Acknowledgments	20	19	0
		v							Reactions to Participant's Utterance	39	36	0
	v								Accepting or Rejecting Participant's Utterance	20	20	0
		v							Answering Participant's Question	24	23	0
							v		Fulfilling Participant's Request	2	2	0
	v								Reciprocating (e.g., returning a greeting)	6	6	0
		v							Butting In	8	6	0
	v								Other Appropriate Responses	13	10	0
									Inappropriate Response	23	0	17
		v							Speaking As If Participant Did Not Speak	14	0	12
							v		Acting As If Participant Did Not Speak	2	0	1
	v								Cuer Speech Does Not Match Participant's Utterance	4	0	3
							v		Cuer Action Does Not Match Participant's Utterance	0	0	0
v									Cuer Speech Overlaps Participant's	3	0	1
				v					**Facial Expressions (e.g., smiling, raising an eyebrow)**	5	4	1
									Gestures	37	32	2
				v					Head Gestures (e.g., nodding, turning ear towards robot)	16	14	1
					v				Hand Gestures (e.g., waving, beckoning, thumbs up)	17	15	0
				v	v				Head and Hand Gestures (e.g., shushing with finger to lips)	4	3	1
				v					Gaze Changes (i.e., looking towards or away)	11	5	2
									Body Movement	43	18	19
						v			Leaning	29	8	19
						v			Turning	9	7	0
						v			Walking	5	3	0
								v	**Moving the Robot**	6	6	0
								v	**Adjusting Audio Settings on the Robot**	4	1	1
									Reducing Classroom Noise to Hear Better (--- no cues ---)	0	0	0
									Cue Modalities			
^									**Conversation Turn Timing**	386	235	124
									Audio	276	152	91
	^								Vocal :: Verbal (keywords)	132	71	51
		^							Vocal :: Verbal (sentences)	144	81	40
			^						Vocal :: Nonverbal	0	0	0
									Video	104	64	26
				^					Head, Eye, and Face	36	26	5
					^				Hands	21	18	1
						^			Other Body Pose	43	18	19
							^		Activity	4	2	1
								^	**Robot Settings and Movement**	10	7	1

Table 3: Comprehension cue counts by category and subcategory (top part of the table), and by modality (bottom part). Assignment of subcategories (and categories without subcategories) to modalities is shown on the left. Categories with subcategories and modalities with submodalities are colored gray, as is the column for the Vocal :: Nonverbal submodality because no subcategories were assigned to it. Numbers preceded by a star (*) were affected by P3's unique behavior, as discussed in Sections 7.1 and 7.4.

Both P3 and P4 spent time talking with teachers or classmates about the network latency while using the robot. The latency level could instead be displayed to the operator and their teacher or classmate as part of another feedback system. For example, the system could display when a speaker's utterance finishes playing on the other end of the remote connection.

8.2 Detecting Comprehension Cues

We developed 38 subcategories (and categories without subcategories) to describe the diversity of cues in our K-12 classroom setting, but 54% of the cues were in just three of them: "Beginning to Speak at an Appropriate Time" (32%), "Silence When a Response was Warranted" (17%), and "Reactions to Participant's Utterance" (5%). In this section we discuss how a telepresence system could detect a large share of the available comprehension cues with a relatively simple detection system.

8.2.1 Choosing the Modalities to Detect. Cues in different modalities can carry different amounts of information. The subcategories in the audio modality, for example, often communicate *how much* the cuer understood, whereas cues based on conversation turn timing only communicate whether the cuer could tell when the participant's turn ended.

A single natural language understanding (NLU) system could detect cues in both the audio and conversation turn timing modalities. Even an NLU system that can only perform keyword detection and end-of-turn prediction could detect a large portion of the negative cues identified in our study (see Table 3).

8.2.2 Detecting Cues in the Most Common Modalities. Our work identifies the most common modalities of comprehension cues used during our telepresence deployments in K-12 classrooms: conversation turn timing cues, keyword-level verbal cues, and sentence-level verbal cues.

Cues in these modalities could be detected by real-time telepresence systems. Specifically, designers could draw upon work towards fluent turn-taking in spoken human-robot interactions as reviewed by Skantze [46]. Other work used timing information such as tempo and duration of responses and silences to directly predict misunderstandings [21, 22]. Our work revealed that it can be difficult to determine whether a potential cue is in response to a particular utterance, however, especially when the the utterance is out of turn or difficult to understand (see Section 7.1).

Models have also been trained using key words and phrases to detect features such as repetition of the same thing, disconfirmations via "no" or "not", and making corrections by repeating something with certain words changed [34]. Finally, spoken dialogue systems can understand utterances at the sentence level [12] and there has been recent work on understanding more difficult utterances such as indirect speech acts (e.g., "could you get me a coffee?") [4].

8.3 Improving the Cue Annotation Process

Inter-annotator agreement for comprehension cues can be improved in multiple ways. Our cue annotation process was very difficult for our annotators. A better approach would involve multiple passes through the recordings, focusing on one cue modality

(e.g., hands, face, posture, verbal content) at a time, for one potential listener at a time. However, this would take even more time than our already exceedingly time-intensive annotation process did. Alternatively, researchers could focus on cues in just one modality (e.g., conversation turn timing) or on dyadic interactions with only one potential listener.

8.4 Limitations

The evaluation of our speech intelligibility feedback system was based on a single prototype system trained on data from three people in the lab; the two participants who experienced it reported that they rarely attended to the feedback. While the self-reports showed that our system could be useful for private conversations and class presentations, alternative designs merit exploration.

The low inter-annotator agreement suggests that some cues, or even entire new categories of cues, were missed by our annotators. The relative frequencies we reported for different types of cues may also have been impacted, especially for certain cue types during P3's deployment, as discussed in Section 7.1.

Additionally, the relative frequencies of different types of cues likely depend on contextual factors like the number and physical locations of potential listeners, and whether it is acceptable for the operator to speak. Therefore, the relative frequencies of cues in different categories and modalities could differ significantly between K-12 classrooms and compared to non-classroom environments.

9 CONCLUSIONS

We presented the first findings about speech intelligibility feedback to homebound K-12 students using a telepresence robot to attend class remotely. We analyzed a multimodal dataset collected from four long-term, in-the-wild deployments.

Our analyses revealed that the two participants who evaluated the speech intelligibility feedback system rarely attended to it; future work should study how students speak at different volumes instead of adjusting the robot's audio settings. Future feedback systems should visualize network latency and help students calibrate more than one volume parameter.

By annotating and analyzing over 700 multimodal comprehension cues, we developed 15 cue categories and 31 subcategories. For our four participants, conversation turn timing and verbal content were the most frequent cue modalities. Conversation turn timing cues occurred more frequently overall, whereas verbal content cues contained more information and might be the most frequent modality for negative cues. Future telepresence systems would benefit from detecting and interpreting cues in the one or two richest modalities and using them to ensure that remote users are being heard, either by automatically adjusting the audio settings or by providing other interventions via the user interface.

REFERENCES

[1] Jenay M. Beer and Leila Takayama. 2011. Mobile remote presence systems for older adults: acceptance, benefits, and concerns. In *Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. ACM, Lausanne, Switzerland, 19–26.

[2] John Bell, William Cain, Amy Peterson, and Cui Cheng. 2016. From 2D to Kubi to Doubles: Designs for student telepresence in synchronous hybrid classrooms. *International Journal of Designs for Learning* 7, 3 (2016), 19–33.

[3] John S. Bradley and Hiroshi Sato. 2008. The intelligibility of speech in elementary school classrooms. *The Journal of the Acoustical Society of America* 123, 4 (2008), 2078–2086.

[4] Gordon Briggs, Tom Williams, and Matthias Scheutz. 2017. Enabling robots to understand indirect speech acts in task-based interactions. *Journal of Human-Robot Interaction* 6, 1 (2017), 64–94.

[5] Adelbert W. Bronkhorst. 2000. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica* 86, 1 (2000), 117–128.

[6] Erik Cambria, Newton Howard, Jane Hsu, and Amir Hussain. 2013. Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics. In *IEEE Symposium on Computational Intelligence for Human-Like Intelligence (CIHLI)*. IEEE, Singapore, 108–117.

[7] Elizabeth Cha, Samantha Chen, and Maja J. Matarić. 2017. Designing telepresence robots for K-12 education. In *26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Lisbon, Portugal, 683–688.

[8] Elizabeth Cha, Jillian Greczek, Ao Song, and Maja J. Matarić. 2017. My classroom robot: Exploring telepresence for K-12 education in a virtual environment. In *26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Lisbon, Portugal, 689–695.

[9] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter* 19, 2 (2017), 25–35.

[10] Ambre Davat, Véronique Aubergé, and Gang Feng. 2018. Integrating Socio-Affective Information in Physical Perception aimed to Telepresence Robots. In *5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC)*. IEEE, Kaohsiung, Taiwan, 24–29.

[11] Ambre Davat, Véronique Aubergé, and Gang Feng. 2020. Can we hear physical and social space together through prosody?. In *10th International Conference on Speech Prosody*. ISCA, held virtually, 715–719.

[12] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. International Foundation for Autonomous Agents and Multiagent Systems, Paris, France, 1061–1068.

[13] A. T. Escottá and W. Beccaro. 2021. Controle Automático de Volume em Tempo Real Utilizando Inferência Fuzzy em um Sistema Embarcado. *Trends in Computational and Applied Mathematics* 22, 1 (2021), 41–60.

[14] Deborah I. Fels, Judith K. Waalen, Shumin Zhai, and Patrice L. Weiss. 2001. Telepresence under Exceptional Circumstances: Enriching the Connection to School for Sick Children. In *International Conference on Human-Computer Interaction (INTERACT)*, Michitaka Hirose (Ed.). IOS Press, Tokyo, Japan, 617–624.

[15] Naomi T. Fitter, Elizabeth Cha, and Maja J. Matarić. 2018. Comparing Remote Learning Technologies. In *Proceedings of the 2018 RSS Workshop on Fundamental Issues in Symbiotic Human-Robot Interaction, held in conjunction with the Robotics Science and Systems (RSS) Conference*. RSS Foundation, Pittsburgh, PA, USA.

[16] Naomi T. Fitter, Yasmin Chowdhury, Elizabeth Cha, Leila Takayama, and Maja J. Matarić. 2018. Evaluating the effects of personalized appearance on telepresence robots for education. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Chicago, IL, USA, 109–110.

[17] Naomi T. Fitter, Nisha Raghunath, Elizabeth Cha, Christopher A. Sanchez, Leila Takayama, and Maja J. Matarić. 2020. Are we there yet? Comparing remote learning technologies in the university classroom. *IEEE Robotics and Automation Letters* 5, 2 (2020), 2706–2713.

[18] John H. Flavell, James Ramsey Speer, Frances L. Green, Diane L. August, and Grover J. Whitehurst. 1981. The development of comprehension monitoring and knowledge about communication. *Monographs of the Society for Research in Child Development* 46, 5 (1981), 1–65.

[19] Benjamin Gleason and Christine Greenhow. 2017. Hybrid education: The potential of teaching and learning with robot-mediated communication. *Online Learning Journal* 21, 4 (2017), 159–176.

[20] Akira Hayamizu, Michita Imai, Keisuke Nakamura, and Kazuhiro Nakadai. 2014. Volume adaptation and visualization by modeling the volume level in noisy environments for telepresence system. In *Proceedings of the 2nd International Conference on Human-Agent Interaction*. ACM, Tsukuba, Japan, 67–74.

[21] Jun-Ichi Hirasawa, Noboru Miyazaki, Mikio Nakano, and Kiyoaki Aikawa. 2000. New feature parameters for detecting misunderstandings in a spoken dialogue system. In *6th International Conference on Spoken Language Processing*. ISCA, Beijing, China.

[22] Julia Hirschberg, Diane Litman, and Marc Swerts. 2004. Prosodic and other cues to speech recognition failures. *Speech communication* 43, 1-2 (2004), 155–175.

[23] Sara Kiesler. 2005. Fostering common ground in human-robot interaction. In *IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Nashville, TN, USA, 729–734.

[24] Atsunobu Kimura, Masayuki Ihara, Minoru Kobayashi, Yoshitsugu Manabe, and Kunihiro Chihara. 2007. Visual feedback: its effect on teleconferencing. In *12th International Conference on Human-Computer Interaction*. Springer, Beijing, P.R. China, 591–600.

[25] Annica Kristoffersson, Silvia Coradeschi, and Amy Loutfi. 2013. A review of mobile robotic telepresence. *Advances in Human-Computer Interaction* 2013 (2013), 1–17. https://doi.org/10.1155/2013/902316

[26] Harlan Lane, Bernard Tranel, and Cyrus Sisson. 1970. Regulation of voice communication by sensory dynamics. *The Journal of the Acoustical Society of America* 47, 2B (1970), 618–624.

[27] Min Kyung Lee and Leila Takayama. 2011. "Now, I have a body:" uses and social norms for mobile remote presence in the workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, Vancouver, Canada, 33–42.

[28] Yuanchao Li, Carlos Toshinori Ishi, Nigel Ward, Koji Inoue, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2017. Emotion recognition by combining prosody and sentiment analysis for expressing reactive emotion by humanoid robot. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, Kuala Lumpur, Malaysia, 1356–1359.

[29] Chaoran Liu, Carlos Ishi, and Hiroshi Ishiguro. 2019. Auditory scene reproduction for tele-operated robot systems. *Advanced Robotics* 33, 7-8 (2019), 415–423.

[30] Lorenzo Lucignano, Francesco Cutugno, Silvia Rossi, and Alberto Finzi. 2013. A dialogue system for multimodal human-robot interaction. In *Proceedings of the 15th ACM International Conference on Multimodal Interaction (ICMI)*. ACM, Sydney, Australia, 197–204.

[31] Yukun Ma, Khanh Linh Nguyen, Frank Z. Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion* 64 (2020), 50–70.

[32] Ross Mead and Maja J. Matarić. 2014. Probabilistic models of proxemics for spatially situated communication in HRI. In *Algorithmic Human-Robot Interaction Workshop at the 9th ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Bielefeld, Germany.

[33] Ross Mead and Maja J. Matarić. 2016. Perceptual models of human-robot proxemics. In *International Symposium on Experimental Robotics*. Springer, Rappongi, Tokyo, Japan, 261–276.

[34] Raveesh Meena, José Lopes, Gabriel Skantze, and Joakim Gustafson. 2015. Automatic Detection of Miscommunication in Spoken Dialogue Systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Prague, Czech Republic, 354–363. https://doi.org/10.18653/v1/W15-4647

[35] Yukiko I Nakano, Gabe Reinstein, Tom Stocky, and Justine Cassell. 2003. Towards a model of face-to-face grounding. In *Proceedings of the 41st annual meeting of the Association for Computational Linguistics*. ACL, Sapporo, Japan, 553–561.

[36] Carman Neustaedter, Gina Venolia, Jason Procyk, and Daniel Hawkins. 2016. To Beam or not to Beam: A study of remote telepresence attendance at an academic conference. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)*. ACM, Canada (virtual), 418–431.

[37] Veronica Ahumada Newhart, Mark Warschauer, and Leonard Sender. 2016. Virtual inclusion via telepresence robots in the classroom: An exploratory case study. *The International Journal of Technologies in Learning* 23, 4 (2016), 9–25.

[38] ABC News. 2014. Robots allow classrooms to attend virtual field trips. https://www.abc.net.au/news/2014-10-22/robots-help-teach-school-students-about-water-supplies/5831892. Accessed: 2021-05-20.

[39] Andreas Paepcke, Bianca Soto, Leila Takayama, Frank Koenig, and Blaise Gassend. 2011. Yelling in the hall: using sidetone to address a problem with mobile remote presence systems. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST)*. ACM, Santa Barbara, CA, USA, 107–116.

[40] Oliver Roesler, Amir Aly, Tadahiro Taniguchi, and Yoshikatsu Hayashi. 2019. Evaluation of word representations in grounding natural language instructions through computational human-robot interaction. In *14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Daegu, South Korea, 307–316.

[41] Antonio Roque and David Traum. 2009. Improving a virtual human using a model of degrees of grounding. In *21st International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI, Pasadena, CA, USA.

[42] Raquel Ros, Séverin Lemaignan, E Akin Sisbot, Rachid Alami, Jasmin Steinwender, Katharina Hamann, and Felix Warneken. 2010. Which one? grounding the referent based on efficient human-robot interaction. In *19th IEEE International Symposium in Robot and Human Interactive Communication (RO-MAN)*. IEEE, Viareggio, Italy, 570–575.

[43] Matthew Rueben, Thomas Groechel, Yulun Zhang, Maja J Matarić, and Gisele Ragusa. 2020. Increasing Telepresence Robot Operator Awareness of Speaking Volume Appropriateness: Initial Model Development. In *Companion of the 16th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Cambridge, UK, 421–423.

[44] Boran Sahindal. 2020. *Detecting Conversational Failures in Task-Oriented Human-Robot Interactions*. Master's thesis. KTH Royal Institute of Technology.

[45] Mohit Shridhar and David Hsu. 2018. Interactive Visual Grounding of Referring Expressions for Human-Robot Interaction. In *Proceedings of Robotics: Science and Systems (RSS)*. RSS Foundation, Pittsburgh, PA, USA. https://doi.org/10.15607/RSS.2018.XIV.028

[46] Gabriel Skantze. 2021. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language* 67 (2021), 1–26.

https://doi.org/10.1016/j.csl.2020.101178

[47] Masaaki Takahashi, Masa Ogata, Michita Imai, Keisuke Nakamura, and Kazuhiro Nakadai. 2015. A case study of an automatic volume control interface for a telepresence system. In *24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Kobe, Japan, 517–522.

[48] Katherine M. Tsui, Munjal Desai, Holly A. Yanco, and Chris Uhlik. 2011. Exploring use cases for telepresence robots. In *6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Lausanne, Switzerland, 11–18.

[49] Gina Venolia, John Tang, Ruy Cervantes, Sara Bly, George Robertson, Bongshin Lee, and Kori Inkpen. 2010. Embodied social proxy: mediating interpersonal connection in hub-and-satellite teams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, Atlanta, GA, USA, 1049–1058.

[50] Paul M. Vespa, Chad Miller, Xiao Hu, Val Nenov, Farzad Buxey, and Neil A. Martin. 2007. Intensive care unit robotic telepresence facilitates rapid physician response to unstable patients and decreased cost in neurointensive care. *Surgical neurology* 67, 4 (2007), 331–337.

[51] Patrice L. Tamar Weiss, Carolynn P. Whiteley, Jutta Treviranus, and Deborah I. Fels. 2001. PEBBLES: A personal technology for meeting educational, social and emotional needs of hospitalised children. *Personal and Ubiquitous Computing* 5, 3 (2001), 157–168.