Behavioral Changes in Speakers who are Automatically Captioned in Meetings with Deaf or Hard-of-Hearing Peers

Matthew Seita[1], Khaled Albusays[1], Sushant Kafle[1], Michael Stinson[2], Matt Huenerfauth[1]
Golisano College of Computing and Information Sciences[1], National Technical Institute for the Deaf[2]
Rochester Institute of Technology (RIT), Rochester, NY, USA
mss4296@rit.edu, kla3145@rit.edu, sxk5664@rit.edu, msserd@rit.edu, matt.huenerfauth@rit.edu

ABSTRACT
Deaf and hard of hearing (DHH) individuals face barriers to communication in small-group meetings with hearing peers; we examine generation of captions on mobile devices by automatic speech recognition (ASR). While ASR output displays errors, we study whether such tools benefit users and influence conversational behaviors. An experiment was conducted where DHH and hearing individuals collaborated in discussions in three conditions (without an ASR-based application, with the application, and with a version indicating words for which the ASR has low confidence). An analysis of audio recordings, from each participant across conditions, revealed significant differences in speech features. When using the ASR-based automatic captioning application, hearing individuals spoke more loudly, with improved voice quality (harmonics-to-noise ratio), with a non-standard articulation (changes in F1 and F2 formants), and at a faster rate. Identifying non-standard speech in this setting has implications on the composition of data used for ASR training/testing, which should be representative of its usage context. Understanding these behavioral influences may also enable designers of ASR captioning systems to leverage these effects, to promote communication success.

Author Keywords
Deaf and Hard of Hearing, Accessibility, Automatic Speech Recognition, Speaking Behavior, Communication

ACM Classification Keywords
• **Human-centered computing → Accessibility design and evaluation methods**; *Empirical studies in accessibility; Accessibility technologies*;

INTRODUCTION
People who are Deaf or Hard of Hearing (DHH) are a substantial minority of the world population, e.g. about 20% of Americans report some degree of hearing loss [16]. Of these, approximately 60% participate in education or workplace settings [16]. Elliot et al. [10] surveyed 108 DHH students who had recent work experiences with hearing colleagues and found they were not satisfied with current methods for communicating with hearing peers in small groups. Many commented that they skipped entire meetings and waited for someone to email them a summary. This study highlighted a critical problem faced by DHH people as they enter the workforce: Communication barriers may lead to **isolation**, miscommunication, or reduced productivity or professional outcomes. In fact, there are lower employment rates for DHH individuals when compared to hearing peers [14], and research has found that a limiting factor in DHH individuals' professional success is the communication barrier with hearing peers [19].

Many DHH individuals use sign-language interpreting or real-time captioning provided by a human service provider, e.g. for classroom lectures, meetings, or other events. While these services are beneficial for providing communication access, they are **less available in the workplace or team-based projects in educational settings**. A challenge is that many workplace conversations are impromptu in nature, making it difficult to schedule these services in advance. In addition, some regions of the world face a scarcity of qualified or affordable ASL interpreters or captionists.

Our research focuses on a particular approach to address this challenge: **mobile applications that provide live captions, based on automatic speech recognition (ASR)**. ASR software is far from perfect, and the complex audio environment of certain real-world applications cause ASR to be rife with output errors [3]. However, rapid progress has occurred in recent years: prospects for accurate ASR in real-world meeting environments are promising. While ASR still produces text containing errors, the flexibility afforded by such technology to support impromptu communication has driven significant recent interest in ASR-captioning among the accessibility community: Recent work has investigated the requirements and concerns of DHH users [4], metrics to predict the suitability of ASR for this application [17], or DHH users' reactions to experiences with prototype systems [11, 18, 26].

In this paper, we investigate this technology from a new perspective: **How does the use of an ASR-based tool to communicate with DHH colleagues influence the speaking behavior of a hearing individual?**

Our **motivation** for this research stems from the following:

ASSETS '18, October 22–24, 2018, Galway, Ireland

ACM ISBN 978-1-4503-5650-3/18/10…$15.00
https://doi.org/10.1145/3234695.3236355

- **Representativeness of Corpora:** ASR technology is trained on large datasets (or "corpora") consisting of audio recordings of human speakers accompanied by text transcripts of what was spoken. However, it is well-known that if the genre, speaker characteristics, audio environment, or other aspects of the data used to train an ASR system is not representative of the type of audio data that the system will be later asked to analyze, then we can expect the system to have less accurate performance. The majority of speech data used to train ASR systems, e.g. as in [40], consists of transcribed recordings of telephone conversations or other sources. Recordings of hearing speakers using an ASR-based tool to communicate with a DHH colleague are ***not*** typically included. If the speech characteristics of a hearing individual differ in this context, then: (a) the standard training data used to implement ASR systems should include additional audio content from this new context, and (b) automatic evaluations of ASR accuracy based on their performance on existing training datasets may not be indicative of an ASR system's actual performance for this application.
- **Implications for Design:** If something about the social or technological context of using an ASR-based mobile application to communicate with a DHH person is influencing the speech behavior of hearing people, then designers of such systems could leverage this to their advantage. Specifically, some of these changes may be beneficial for DHH individuals, if it leads to the hearing person slowing their communication rate, thereby making time for the DHH individual to glance between captions and the face of their communication partner, especially if the DHH individual is using speech-reading to supplement their understanding of their partner's message. Furthermore, if designers of such systems can influence the behavior of the hearing individual, e.g. to promote clearer articulation, it may be possible to encourage speech that would be easier for ASR systems to accurately recognize. If the text output of the ASR contains fewer errors, the experience of a DHH user of such a system may improve.

To investigate this question, we have conducted an **experimental study** with a prototype ASR system used during meetings between DHH and hearing individuals, and we have analyzed the speech characteristics of the hearing participants. In this paper, we present prior work on the use of ASR as a communication tool for DHH individuals and on factors known to influence speaking behavior. Subsequently, we present our research hypotheses, and then we describe our collection of recordings of meetings between DHH and hearing individuals (in some cases, using an ASR-based communication application). Next, we present an analysis of the speech characteristics of the hearing individual with/without this application. Finally, we summarize our findings, discuss the implications of this work, and suggest future avenues of research in this area.

PRIOR WORK

Our literature review began by examining prior research on ASR technology to support communication and information access for DHH users. In addition, we examined prior research on how the voice characteristics or speaking behavior of hearing individuals may be influenced by contextual factors (e.g. knowing that they are speaking into an ASR system or that they are speaking with a DHH person); prior work on these speech behavioral changes provided a basis for our research hypotheses, which are presented immediately after this literature review.

ASR as an Accessibility Tool for DHH Individuals

Computing accessibility researchers have examined a variety of methods for **generating real-time captions** of audio information for DHH users, to reduce their reliance on professional captioning/transcription services. Some of this work has examined the use of crowds of non-experts who are asked to transcribe small segments of audio, which is combined and shown to users [22]. Other work has applied ASR technology to identify the words spoken in some audio, with a crowd of non-experts asked to correct errors in the ASR-produced transcript [37, 38]. Researchers have also investigated using ASR to transcribe classroom lectures for students [2] or as a real-time system for classroom captioning [12, 21] or videoconferencing [13]. ASR accuracy has significantly improved in recent years, due to advances in neural machine learning [40]. Given this improvement, recent commercial efforts have deployed ASR as a real-time captioning system in classrooms, e.g. [32]. However, a recent study by Kawas et al. [18] identified shortcomings in how ASR was used in classrooms to support DHH students, and the authors proposed several design guidelines to address this feedback from DHH students. Many DHH university students in their study were frustrated with errors in spelling and grammar of captions provided by ASR in the classroom, and they requested various improvements [18].

Berke et al. [4] discuss how there may be greater promise in the use of ASR to facilitate **small-group or one-on-one communication** between DHH individuals and hearing colleagues. Specifically, they argue that participants in a small-group communication context may be able to monitor the accuracy of the ASR so that they can use alternate communication channels (e.g. typing or writing messages) to correct errors in the ASR or to bypass its use if it is not working adequately. In fact, multiple recent studies have investigated preliminary prototypes of ASR-based captioning systems, used in this small-group context:

- Mallory et al. [26] deployed an ASR-based chat application on smartphones in workplace settings with DHH students. They argued that, despite errors and other challenges with current systems, they believed that the technology had promise for facilitating communication in small groups, and it was generally good enough to be helpful in real-world settings [26].

- Berke et al. [4] analyzed DHH users' perspectives on using ASR during one-on-one meetings by asking DHH users to watch a video simulating a brief business meeting, with automatically generated captions below. Their participants indicated that the errors in the ASR-output text were frustrating, yet the participants expressed an interest in using ASR for accessibility.
- Elliot et al. [11] conducted a study with DHH participants engaging in brief collaborative discussions with hearing peers, using an ASR-based real-time captioning application on tablet computers. The authors also found that their participants expressed an interest in using ASR technology in this context and satisfaction with the prototype they had used [11].

Researchers have examined automatic and user-based methodologies for **evaluating** ASR-based real-time captioning systems. Kafle et al. [17] proposed a new metric to replace the ubiquitous word error rate (WER) metric often used to measure the performance of ASR systems. They found that the predictions of their new metric (about the quality of an ASR-output text) correlated better with the judgements of DHH individuals (than WER did); their metric could be used to evaluate whether specific ASR systems would create useful captions for DHH users. In a recent methodological study, Berke at al. [5] analyzed whether specific question types were effective at measuring DHH users' perceptions of ASR-based caption quality, with participants at various English reading literacy levels.

In summary, prior work has found that DHH users have concerns about accuracy of ASR-based captioning for live interactions with hearing peers, yet they have interest in such systems, e.g. [4, 11, 18]. While studies with a small number of participants, e.g. [11, 26], have found that DHH users are satisfied with prototype ASR-based systems, in larger studies, users suggested design improvements are needed, e.g. [11, 18]. Overall, prior research suggests that ASR-based captioning for meetings is an area with **potential** (given users' interest), yet additional research is needed on how to best design such systems. We identify a gap in the literature: While Berke et al. [4] have argued that one reason why ASR technology may be more suitable for small-group meetings (rather than for large lectures) is that the conversational participants in a small group may adapt their behavior based on the system's performance – we did not encounter prior research focused on the **hearing participant** in these contexts, nor how their behavior may be shaped by the design of an ASR-based captioning tool. Thus, the focus of our current study is on examining how hearing individuals may change their speech behavior when participating in a meeting context with ASR-based captioning tools and DHH conversational partners.

Definitions of Speech Properties Used Below

The discussion below refers to various speech properties; for the reader's convenience, we provide brief definitions below. Other speech-related terminology is defined, as needed, when we present our hypotheses for this study.

- **Intensity** refers to the volume of the voice, i.e. the amount of acoustic energy in the speech signal.
- **Pitch** is the fundamental frequency of the voice, sometimes referred to as F_0, e.g. female speakers tend to have higher F_0 than male speakers do.
- **Formants** refer to harmonic resonances, which appear as dark bands in a spectrogram image of a voice and are due to acoustic resonance in the vocal tract, which speakers articulate to produce vowel sounds. F_1 refers to the lowest resonance in the speech frequency spectrogram, and F_2 refers to the 2nd lowest. Loosely, F_1 frequency indicates tongue height during vowels (higher F_1 indicating tongue lower in the mouth, and vice versa), and changes in F_2 frequency may indicate tongue placement during vowel sounds (with higher F_2 frequency indicating a more front-of-mouth placement of the tongue during vowels, and vice versa). When speakers "hyperarticulate" [30], they tend to emphasize the differences in vowel sounds by moving their tongue to more extreme positions in the mouth, which leads to noticeable shifts in F_1 and F_2 formants. Picart et al. [30] measured changes in formants and other properties in hyperarticulated speech, finding a greater variation in F_1 and F_2 values, as well as F_0 pitch increases.
- **Harmonics-to-Noise Ratio (HNR)** is a measure of the proportion a sound signal's energy that consists of periodic components vs. non-periodic noise. In speech, this is often used as a measure of voice quality: Voices with low HNR are characterized as sounding "hoarse." Prior studies have found HNR decreasing with a speaker's age [25] and that higher HNR correlated with positive subjective judgements from listeners [24].

ASR or DHH-Partner Influencing Speaking Behavior

While we did not identify prior research on the behavior of speakers in a context where they were *both* conversing with a DHH partner *and* speaking into an ASR system, we have identified prior work on such behavioral changes in each of these contexts individually, which we summarize below.

- Prior research on speaking behavior indicates that hearing participants may sometimes adjust their speaking patterns **when speaking to ASR software**. Oviatt et al. [29] found that when users spoke to an ASR system with misrecognition errors, they spoke more slowly and paused longer and more often. Stent et al. [35] found that that speakers increased their articulation (e.g. higher F_2) and decreased their speech rate when presented with errors that indicated that speech recognition software could not understand them. However, speakers returned to a normal speaking style after some time without any errors. Stent et al. [35] found that hyperarticulation did not negatively

impact ASR performance, in contrast to prior work [34] with older ASR systems. Burnham et al. [7] found that when speaking to a computer avatar that appeared to misunderstand, users had greater F_1 and F_2 variation, as well as longer vowel duration, but they did not show higher F_0. (This is similar to the profile of speech directed to a non-native speaker [33], rather than child-directed speech, which has increased F_0 pitch.) In summary, there is evidence of hyperarticulation (slowed speech rate and shifts in F_1 and F_2), especially when users realized that the ASR system is having difficulty (e.g. by displaying misrecognized words).

- There has also been significant research on how speakers modify their behavior when they are **having a conversation with a person who is having difficulty hearing or understanding**. For instance, Buz et al. [8] described how hearing speakers adjusted their pronunciation of words when they were misrecognized, i.e. hyperarticulating, as in the ASR-directed speech research above. Koster [20] found that when speakers address an actor who behaved as if he had difficulty understanding their speech, they hyperarticulated their speech, leading to measureable differences in duration, fundamental frequency, formants and formant bandwidths [20]. Specifically, the duration of vowels increased, fundamental frequency increased, and there were changes in F_1 and F_2 formant frequencies.

Based on this prior research, we can speculate that if the hearing participant engaging in a conversation with a DHH partner while using ASR software has the impression that their message is not being conveyed, they may modify their speech – potentially with changes in voice properties such as speech rate, F_1 and F_2 formants, and possibly F_0 pitch, as have been observed in prior research in related contexts.

RESEARCH QUESTIONS

Given the concerns outlined in our "Introduction" in regard to the representativeness of speech corpora (for training or for evaluation) and that the design of ASR-based captioning systems may influence the speech behavior of hearing individuals (potentially to the benefit of ASR accuracy or comprehensibility for their DHH conversational partner), we investigate: When engaging in a meeting with a DHH partner, do the speech characteristics of a hearing individual change? Does the use of an ASR-based captioning application influence this behavior? Would displaying information on the screen to indicate when the ASR system is having difficulty understanding the speech further influence the behavior of the speaker?

Conditions

As discussed in greater detail in the "Methods" section below, we conducted an experimental study in which pairs or triads of individuals (one DHH participant in each group) held several brief meetings to discuss a topic. The meetings were recorded, and the speech was carefully transcribed (with time-codes for the beginning and ending of each word), which enabled us to analyze the speech of the hearing participants for differences in various properties. Each group of participants engaged in three discussions, to experience each of three communication conditions:

- **No ASR:** Each group of participants had to communicate without using ASR technology. They were advised to use whichever communication method they thought would be most beneficial for them. This could include using voice (with the DHH participant using speech-reading), gesturing, writing on paper, or another approach. During this condition, scrap paper and pens/pencils were provided to the participants for their optional use.
- **ASR:** The use of ASR technology was incorporated into their conversations. Each participant in the meetings was given their own mobile device to use with a networked ASR-based "chat" messaging application installed. Each person had to create a username and log in to the application. The hearing participants would communicate by speaking into the application (pressing a button when they wanted to speak), and their words were transcribed by ASR and appeared as a text message in the chat application, visible to the entire group. DHH participants communicated by typing messages into the application.
- **Markup:** Participants used the ASR application as described above, except this time a "markup" feature was enabled. As discussed in Berke et al. [4], there may be a benefit to modifying the display of particular words in a caption (e.g. using italics or underlining) to indicate when the ASR system had low confidence that it had accurately understood a particular word (so that users may distrust a particular word in the transcript). In our study, if the ASR confidence for a word was below 75%, the word was *underlined and italicized* when displayed in the "chat" window, as seen by both hearing and DHH participants.

Hypotheses

We hypothesize that using the ASR technology will influence hearing participants and cause them to change their communication patterns and behaviors. Specifically:

H1: When the speech characteristics of hearing individuals are compared across the three conditions (No ASR, ASR, Markup), there will be a difference in median **intensity** (i.e. voice volume) across conditions. Specifically, we anticipate that participants will speak more loudly in the ASR or Markup conditions, since they are dictating into a device microphone.

H2: …in the **harmonics-to-noise ratio (HNR)** across conditions. Since we are predicting intensity increases in H1 above, and since HNR typically decreases when someone speaks loudly, we anticipate HNR decreases.

H3: …in median **pitch** (i.e. the fundamental frequency of the voice, F_0) across conditions. We speculate that participants may speak in a higher pitch when speaking to an ASR system, as in [20, 30].

H4: …in **F_1 formant** (i.e. the lowest resonance in the speech frequency spectrum) across conditions. We predict participants will hyperarticulate when speaking to an ASR system, as in [29, 35].

H5: …in **F_2 formant** (i.e. second lowest resonance in frequency spectrum) across conditions. As with F_1, changes in F_2 may also indicate hyperarticulation.

H6: …in **speech rate** (i.e. words per minute) across conditions. We speculate that speakers may slow their speech when using ASR, and they may be especially likely to do so if there is a visualization on the screen indicating when the ASR's confidence in its ability to recognize words is dropping (Markup).

H7: …in the **Speech Intelligibility Index (SII)** across conditions. SII is an automatic metric to predict how understandable an audio recording of speech may be for a human listener [1].[1] Speakers may attempt to boost their speech intelligibility for the ASR system.

H8: Finally, if we compare the ASR system's **accuracy** (word error rate) for the speech audio it processed across the ASR and Markup conditions, we expect to see a higher accuracy in the Markup condition. We speculate that seeing the visual indicator of word recognition confidence may encourage speech behavioral changes (like those above) that could lead the speech to be more intelligible for the ASR system.

METHODS

Our experiment investigated the use of ASR technology as an accessibility tool to facilitate communication between DHH and hearing people in small group collaborative meetings. Participants in small groups, consisting of two to three people, were asked to engage in discussions about specific problem-solving topics. Each experimental session was video recorded for later data analysis.

Participant Information

Participants were all undergraduate or graduate students at Rochester Institute of Technology. Recruitment was done via posting of flyers on campus. A total of 21 participants were recruited from this study, participating in groups of 2-3. Twelve of these participants were hearing, and 9 identified as either Deaf or Hard of Hearing. Of the hearing participants, seven identified as male and five identified as female. Of the DHH participants, six identified as male and three identified as female. All of the DHH participants were fluent in ASL, although some of them were comfortable communicating orally with hearing peers, and none of the hearing participants had knowledge of ASL. Participants' ages ranged from 18 to 29 years old.

Room Set-Up

After welcoming the participants and thanking them for their involvement, they were seated in a private observation and recording room at the National Technical Institute for the Deaf at RIT. Researchers monitored the experiment from an observation room, separated from the meeting room by a one-way mirror (Figure 1 and 2).

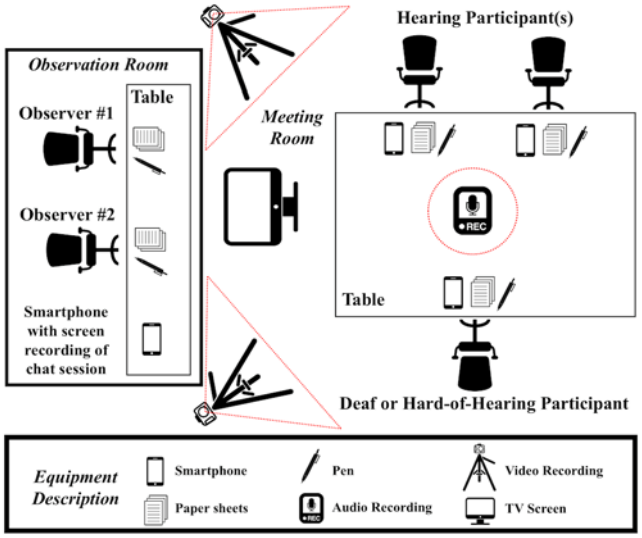

Figure 1: Floorplan of the meeting room and observation room, indicating the location of participants and observers, recording devices, and key equipment used in the study.

The room was set up with a camera and microphone, to record audio and multiple camera angles during the experiment (Figure 1). The DHH participant sat across the table from the 1-2 hearing participant(s), as in Figure 2.

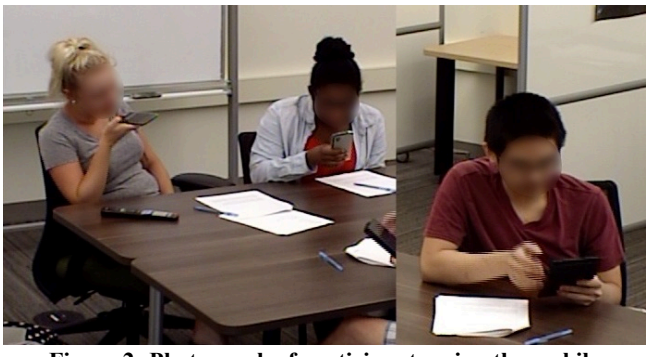

Figure 2: Photograph of participants using the mobile application, in a group meeting with three participants.

The use of a discrete observation room enabled participants to interact more naturally than if observers had been visible in the same room. The various conditions were explained to the participants so they knew how to communicate in each,

[1] There has been significant research on whether SII or other automatic metrics are reliable indicators of understandability for someone who is DHH, e.g. [31], and we are aware of this important limitation. Regardless, we investigate this variable under the premise that the hearing speaker may be intentionally changing their voice to boost its intelligibility (for the ASR system or perhaps for their conversational partner) without awareness of whether those changes specifically may benefit someone who is DHH. With this rationale, we investigate SII changes, but we do not claim that changes in SII necessarily indicate that the speech is more intelligible for the DHH participant in the meeting.

and participants had the opportunity to ask questions after the explanation. Observers entered the meeting room to explain instructions or answer questions; the observers included a hearing person and a Deaf native ASL signer, so that instructions could be explained in English and ASL.

Prototype Captioning System Training

Participants signed an informed consent form and video-recording release before beginning with the experiment. After that, there was a practice session for each participant to help them familiarize themselves with the ASR application. A video with a demo of the mobile application in action was shown to participants.

The application used for this study was developed by Elliot et al. [11] as a prototype for research on communication interaction between DHH and hearing students; those researchers granted permission for the use of this application in our study, shown in Figure 3. The app allows users to log-in with a username of their choice (e.g. their email address) and join a shared "chat room," in which all users in the same chat room see a stream of text messages submitted by other users. Hearing participants submitted audio (pressing a microphone button), which was transcribed by ASR [28]. DHH users typed on an onscreen keyboard to enter and submit text messages[2]. As discussed below, this app was used in the "ASR" and "Markup" conditions of the study; in the "Markup" condition, a feature was enabled that *underlined and italicized* words for which the ASR was not confident in its recognition.

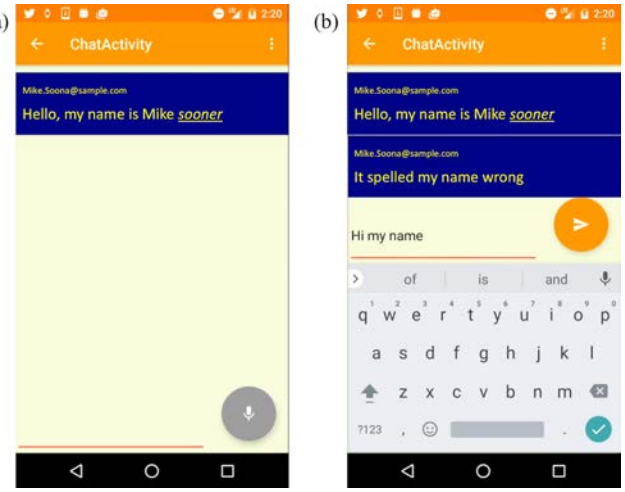

Figure 3: Screenshot of the application used during the study: (a) hearing participants could dictate text by pressing a microphone button to generate text messages, (b) DHH participants could type messages into the chat room.

After watching the introductory training video, each of the hearing participants was asked to log into the ASR application and dictate into the application their name, year in school, and major. They were then asked to dictate five practice phrases into the application. This allowed them to get used to the ASR software and also adjust their voice to make them more easily understood by the application, if necessary. The DHH participants, during the practice session, were asked to type into the application their name, year in school, and major. They were then asked to practice typing five sample phrases into the application. This allowed them to understand how the application works and how their messages would be sent to the other participants. All of the instructions explained in this section were given through both spoken English and ASL.

Prompts for Collaborative Group Work

Each group of participants was given three different scenarios as a basis for discussion. Each scenario asked the participants to discuss a hypothetical survival situation (boat lost at sea, astronaut stranded on the moon, plane crash in desert), and the team was asked to think of 10 items they would want to have in this situation and to agree upon a ranking for those items. These prompts were used in a communication scenario with DHH users by prior researchers [11, 36]; the "moon" scenario is adapted from [15]. For each scenario, groups were given a piece of paper on which they were to write the 10 items they agreed on.

Each group of participants discussed each scenario separately, one after the other. Scenarios were always presented in the order of Boat, Moon, and then Desert. Each scenario was tested with one of the three conditions described previously (No ASR, ASR without Markup, or ASR with Markup), with conditions assigned using a Latin squares schedule. The order in which the conditions were tested was rotated. Participants sat across from each other at a table, and used speech-reading or pen/paper (No ASR condition) or the mobile application (ASR or Markup conditions). For each scenario, participants were asked to spend at least five minutes having a discussion about the topic, using the assigned condition.

DATA COLLECTION AND ANNOTATION

As illustrated in Figure 1, the discussions among each group of the three scenarios (in each of the three conditions) were recorded using two video cameras in the meeting room. One faced the DHH participant, and the other faced the hearing participant(s). A screen-capture from a video is shown in Figure 3; the release signed by participants gave permission for images to be used in research publications.

In addition to the in-room video of the participants, we also recorded a screen-recording of an additional smartphone device connected to the same "chat room" as the participants during the study, to record the conversational stream during the use of the app. The applications running on the smartphones connect over the network to a server that hosts the chat-room and which passes audio-analysis requests to a cloud-based ASR service [28] (details in [11]). The server maintains a log of every text message and ASR

[2] The app permits any user to use both methods of entering text, but we asked hearing participants to use ASR only (unless they believed they must type a few words to correct an ASR mistake). We asked DHH participants to use keyboard input only.

transaction. This log file was also retained after each experimental session, and this data was used to calculate ASR accuracy in the analysis below.

A table-top teleconference-style microphone was positioned on the table between the participants; the audio from this microphone was the basis for the audio analysis below.

Data Annotation

Prior to the analysis, all video recordings were loaded into the ELAN annotation software for data cleaning and preparation [9]. ELAN is an open-source annotation software that allows researchers to annotate and transcribe video or audio recordings. It also provides features to analyze spoken language, sign language, or gesture. ELAN has a tier-based data model that supports multi-participant annotation of time-based media.

Prior to annotating the videos, a template was created to define the "tiers" of annotation (the parallel timeline tracks of information to record for each). We established a separate timeline tier for each participant, where we could label every word spoken by that individual, with start-time and stop-time for each word. We also created a tier where we could indicate if there was more than one person speaking simultaneously during a span of time; this information was needed so that we could omit those sections of audio from our subsequent audio analysis. In addition, the annotation captured when portions of the speech were spoken to the mobile device (for ASR-based captioning).

For each of these five-minute time-segments (for each of the three scenarios, for each group of participants), a pair of researchers produced a transcript of every word spoken. They then met afterward to discuss their annotation and prepare a consensus annotation. At a later time, a third researcher examined all of these transcripts while listening to the original video recordings, to correct any remaining transcription or timing errors that were noticed.

The annotation included filler words such as "um," "uh," "ah," "like," "okay," or "right." In addition, false-starts and repetitions of words were labeled in the annotation. During the subsequent ASR accuracy analysis, when calculating word error rate, it is customary to filter out these disfluencies in the transcript prior to calculating Word Error Rate (WER). We followed this custom during our analysis, but we wanted to first encode this faithful transcription.

ANALYSIS AND RESULTS

We pre-processed the audio recordings prior to using the Praat speech-audio analysis tool [6]. Specifically, we omitted any filled-pause sounds (e.g. "um"), non-speech sounds (e.g. laughing), or periods of time when two people were speaking simultaneously (e.g. when participants spoke over or interrupted one another) – since these audio events would have confounded our later analysis. Our annotators had marked these events on the ELAN timeline, which we used to omit these segments from each audio file. If the entirety of a word was uttered while someone else was also speaking, the whole word was eliminated. If, however, only a portion of a word was spoken over, the start and/or end times were adjusted to remove the overlap – essentially this truncated some words that were spoken. While partial-word audio was used for audio analysis (H1-H7), it was excluded from ASR accuracy analysis (H8). While the focus of this study is on the speech of the hearing participants, we had annotated any speech produced by the DHH participant. This information was used to exclude these segments of time from the audio, including times when the DHH participant's speech overlapped others.

Periods of silence (outside of any spoken utterance time) were cropped out of each recording, thereby producing an audio file appropriate for the analysis for H1-H7 below. Voice data that occurred while participants were practicing using ASR before the conditions, or if they were otherwise talking outside of the actual experiment, was excluded. The result of this pre-processing was a set of three audio files per hearing participant, which contained continuous audio for each of the three conditions. We used these audio files as input to Praat for analysis.[3] A script was used to extract values for hypotheses H1–H5. The other hypotheses could not be analyzed using Praat; so, different methods were used for these, as will be explained in later sections.

In addition to producing audio files for Praat analysis for H1-H7 below), we also produced .csv files containing the transcript of each word spoken by each hearing participant, along with start- and end-times for each. In addition, we used the log-files from our ASR mobile application to record the text output that had been displayed for each ASR-dictated message produced by hearing participants. These text files were used as a basis for the ASR WER analysis to evaluate hypothesis H8 below.

For hypotheses H1–H7, we calculated the relevant speech feature for each participant, for each condition. Then, we compared a participant's results across the three conditions using paired t-tests and applying a Bonferroni correction. We selected this form of analysis (rather than a repeated measures ANOVA) because three hearing participants in the "No ASR" condition only used pencil and paper to communicate and never used their voice. Thus, we did not have any speech audio signal for those three individuals for that one condition. (We had speech audio for all other conditions and for all other hearing participants.) Rather than omitting those three individuals from the analysis (since we did not have data from them for the "No ASR"

[3] Pratt input parameters were as follows: [To Intensity: 25, 0, "yes"; intensity_median = Get quantile: 0, 0, 0.5]; [To Pitch (ac): 0, 25, 15, "yes", 0.03, 0.45, 0.01, 0.35, 0.14, 1000; pitch_median = Get quantile: 0, 0, 0.5, "Hertz"]; [To Harmonicity (cc): 0.01, 25, 0.1, 4.5; harm_mean = Get mean: 0, 0]; [To Formant (burg): 0, 5, 8000, 0.025, 50; for formnum from 1 to 2: f_mean = Get mean: formnum, 0, 0, "Hertz"]. These are generally the default values, which is standard practice for these features.

condition), we instead chose to conduct the analysis with three paired t-tests, with Bonferroni correction on the p-values due to repeated measures, to analyze these results.

Results for H1: Intensity

We analyzed intensity of each hearing participant's voice (median voice volume), across all three conditions, as described above. Significant differences were found between ASR versus No ASR [t(8)=3.91, p=0.00447], and Markup versus No ASR [t(8)=3.05, p=0.0157]. However, no significant difference was observed when comparing Markup and ASR [t(11)=0.251, p=0.806]. These results are shown in Figure 4, with significant pairwise differences indicated with "**" in the image. The "Markup" and "ASR" plots on the left side of the figure display results for the subset (9 of 12) hearing participants who spoke in the "No ASR" condition. Data for all 12 hearing participants appear on the right side of the figure. A similar method of presenting the results from Markup and ASR is used throughout the other Figures in this section.

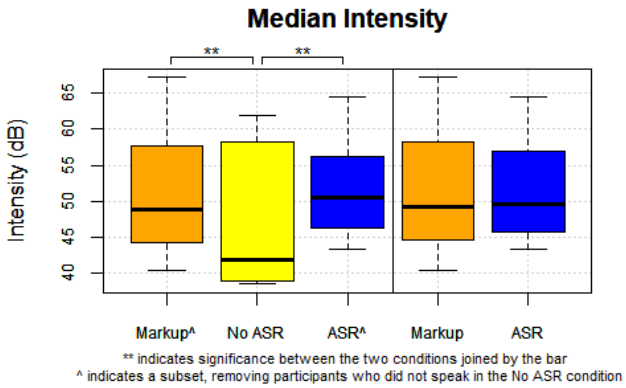

Figure 4: Box plots displaying the median intensity for each hearing participant's voice, across all three conditions

These results indicate that intensity was higher for the ASR and Markup conditions when compared to the No ASR condition: Participants were speaking more loudly in the presence of ASR software, both with and without the *underlining and italics* markup for low-confidence words. Since users were aware that they were speaking into a mobile device, it is reasonable that they may have increased their voice volume as they dictated text into the system.

Results for H2: Harmonics-to-Noise Ratio (HNR)

We analyzed each participant's HNR across the three conditions using paired t-tests and applying a Bonferroni correction (similar to how the analysis had been structured for H1 above). A significant difference was found between Markup versus No ASR [t(8)=3.47, p=0.00845] only. No significant differences were observed between other pairs: ASR vs. No ASR [t(8)=1.59, p=0.152] or Markup vs. ASR [t(11)=1.49, p=0.164]. These results are shown in Figure 5.

Since we observed greater voice intensity (in our results for H1 above), we had actually expected HNR to decline in the Markup and ASR conditions (since people spoke more loudly). Instead, we observed that hearing participants produced speech with higher HNR in the Markup condition. Since higher HNR is often associated with younger adult voices or more positive subjective judgements from listeners [24, 25], observing this effect may indicate that speakers are changing their voice quality when using an ASR system that is presenting them with visual feedback about when it is successfully understanding their voice.

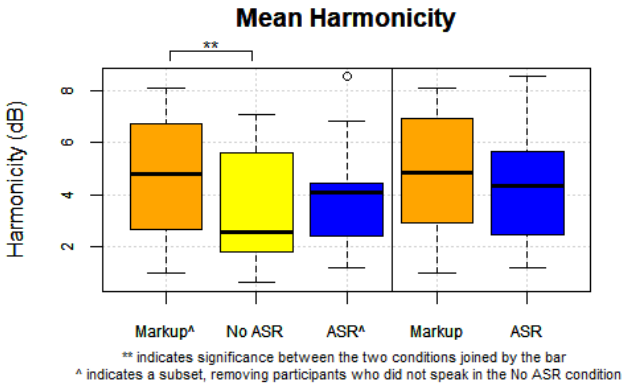

Figure 5: Box plots for harmonicity (mean harmonics-to-noise ratio), for each hearing participant, across the three conditions

Results for H3: Pitch

We analyzed the median pitch across the three conditions (with paired t-tests and a Bonferroni correction, as above). No significant differences were observed across any pair of conditions: ASR vs. No ASR [t(8)=0.74, p=0.480], Markup vs. No ASR [t(8)=0.90, p=0.392], and Markup vs. ASR [t(11)=0.60, p=0.560]. This suggests that using ASR software in this meeting context did not influence speakers' pitch; this agrees with prior research results for speech directed at non-native speakers [33] or ASR systems [7].

Results for H4: F_1 Formant

We analyzed participants' mean F_1 formant frequency across conditions (with paired t-tests and a Bonferroni correction, as above). These results are shown in Figure 6.

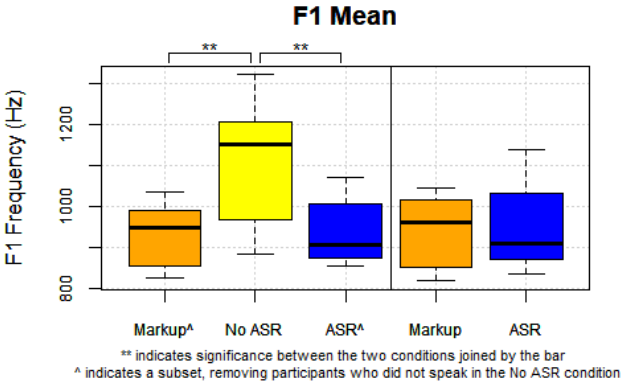

Figure 6: Box plots for F_1 formant means across conditions

We observed significant differences between: ASR vs. No ASR [t(8)=3.93, p=0.00437], and Markup vs. No ASR [t(8)=3.14, p=0.0138]. No difference was observed between Markup and ASR [t(11)=0.63, p=0.539]. This result

suggests that, in meetings with DHH participants, speakers are modifying their articulation when addressing their speech to the ASR system, as compared to when they are speaking without an ASR system.

Results for H5: F_2 Formant

We analyzed the mean F_2 formant frequency across the three conditions (with paired t-tests and a Bonferroni correction, as above). Significant differences were observed between two pairs of conditions: ASR vs. No ASR [t(8)=4.44, p=0.00217], and Markup vs. No ASR [t(8)=3.43, p=0.00893]. No difference was observed between Markup and ASR [t(11)=1.14, p=0.277]. These results are shown in Figure 7. As discussed above (for H4), changes in formant frequencies can be indicative of changes in a speaker's articulation, and this finding aligns with prior work that has observed shifts in F_2 values when speakers are addressing people who are DHH or ASR systems.

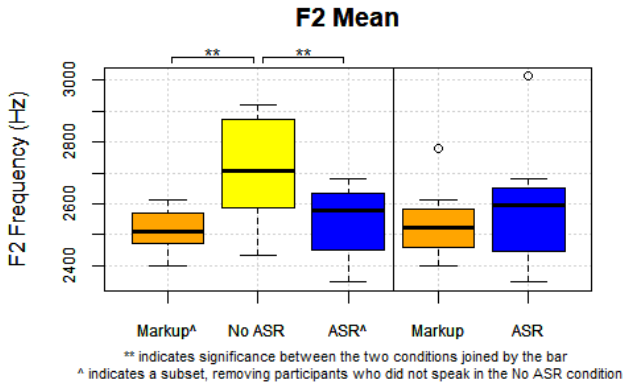

Figure 7: Box plots for F_2 formant means across conditions

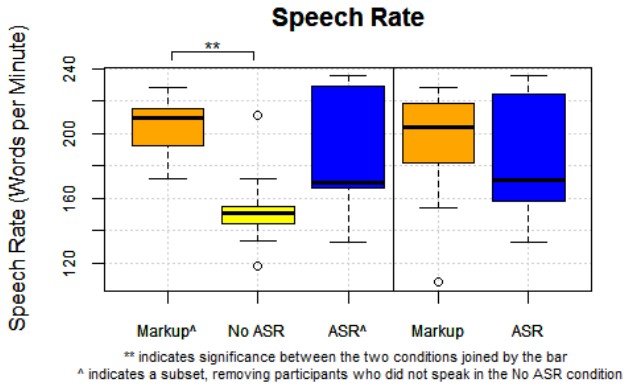

Figure 8: Box plot of results for speech rate, in words per minute, across conditions

Results for H6: Speech Rate

We analyzed the speech rate (the number of words produced per minute) for each participant across the three conditions using paired t-tests and applying a Bonferroni correction, as above. A threshold of 2.0 seconds of silence was used to identify utterance boundaries. Speech rate was calculated by dividing number of words spoken by the total length of all utterances. A significant difference was found between Markup and No ASR [t(8)=4.48, p=0.00206]. However, no difference was observed for other pairs: ASR vs. No ASR [t(8)=2.59, p=0.0319] and Markup vs. ASR [t(11)=0.76, p=0.465]. These results are shown in Figure 8.

Our hearing participants spoke more slowly when directly addressing a DHH participant, rather than when speaking into an ASR-based communication tool with markup. We had originally hypothesized that the use of the ASR technology in this context may lead participants to speak more slowly, but in fact, participants tended to speak more slowly when addressing a DHH participant, without the use of the assistive communication application. Typical rates of conversational speech rate in English vary based on several factors, e.g. [41], but typical conversational speech rate for utterances over 6 words in length has been measured around 240 words per minute in some studies [41]. Thus, the use of ASR was able to bring the speaking rate closer to this natural conversational pace. Of course, for DHH users who are attempting to speech-read while using an ASR conversational app, it may be desirable for their conversational partner to speak more slowly. However, we speculate that in larger group settings, hearing participants may be less likely to modulate their speaking rate to accommodate a DHH conversational partner.

Results for H7: Speech Intelligibility Index (SII)

As discussed previously, the Speech Intelligibility Index (SII) is an automatic metric for predicting how easy it is for a listener to understand the speech contained within an audio file, considering the acoustic energy at various frequency bands; SII is defined by an ANSI standard [1]. We had predicted that participants may change their speech when using the mobile applications, leading to SII changes.

For each of the participant audio files for each condition, we calculated the Speech Intelligibility Index using the SII library for R [39]. After converting audio files to WAV, we read the acoustic data using the tuneR library for R [23]. After applying a Fast Fourier transform (FFT) to the signal, we sampled the acoustic energy at each of the frequency-band centers defined in the "sic.critical" table of the ANSI/ASA S3.5-1997 standard for SII [1], which provides importance scores for 21 critical frequency bands. Based on these values, we calculated SII for each audio file (for each condition for each participant), following the procedure defined in the ANSI/ASA S3.5-1997 standard [1].

We analyzed the Speech Intelligibility Index (SII) for each participant across the three conditions using paired t-tests and applying a Bonferroni correction, as above. No significant differences were observed across any pair of conditions: ASR vs. No ASR [t(8)=0.52, p=0.619], Markup vs. No ASR [t(8)=0.34, p=0.745], and Markup vs. ASR [t(11)=0.18, p=0.861]. While we had previously measured changes in intensity and other speech features across conditions, which would have led us to expect to see SII changes, we did not observe significant SII differences. This suggests that using ASR software in this meeting context did not influence the hearing participants' speech in

a manner that led to changes in SII. It may be the case that the differences in acoustic energy caused by intensity and formant changes (e.g., H1, H4, H5), did not affect acoustic energy sufficiently in SII-relevant frequency bands.

Results for H8: Word Error Rate

To investigate whether displaying markup to indicate words with low ASR confidence influenced participants' speech to cause differences in ASR accuracy, we compared word error rate (WER) between the Markup and ASR conditions. For this analysis, we compared: (a) the text transcript of the ASR-generated text displayed on the mobile application, with (b) the accurate transcriptions of what the person actually dictated into the application (as identified by our annotators who transcribed the recording of each session). We conducted standard text pre-processing prior to calculating WER, which included time-aligning utterance boundaries and editing the transcript to remove filled-pauses (e.g. "um") and transcriptions of other disfluencies, which are not included in the text output of ASR. Finally, we used sclite [27] to align these texts and calculate WER for each participant, for each of the two conditions that involved using ASR technology (ASR and Markup). We used the standard uniform weighting of substitutions, insertions, and deletions for our WER calculations. A paired t-test did not reveal any significant difference between Markup and ASR [$t(11)=1.07$, $p=0.309$]. This was expected, since we had not observed any differences between this pair of conditions for H1–H7.

CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

Our experimental study, with recordings of collaborative discussions between DHH and hearing participants, enabled us to analyze the speech characteristics of the hearing participants: without any special communication tool, with an ASR-based chat application, and with a version of this application that also indicated low-confidence words. Our analysis revealed that when using the application, hearing users spoke more loudly, at a faster rate, with higher HNR, and with non-standard articulation. While some prior work had examined speech directed to ASR or to a person who indicated difficulty understanding, this was the first study to examine multiple conditions in a context with both.

Our identification of non-standard speech behaviors among hearing individuals in this setting has important implications for the future development of ASR-based communication tools for DHH users. Specifically, if hearing individuals speak differently when using such systems, it would be important for researchers to include such speech in datasets used for training and testing ASR systems, to ensure that they are well-suited to this task. In addition, an advantage of applying ASR to meeting contexts rather than lecture settings (discussed by [4]), is that participants may dynamically adapt their behavior in this setting, if they notice the ASR tool is not recognizing their speech. To enable future designers of such systems to leverage this interaction, fundamental research is necessary for understanding how use of this technology may lead to these behavioral effects – as presented in this paper.

There are several limitations of this study, which the reader should consider when determining how these findings may generalize to other contexts. Given the resource-intensive nature of this study (e.g. conducting simulated collaborative meetings, recording and transcribing speech and other key events), the study included relatively few participants (12 hearing, 9 DHH), all of whom were young adults pursuing university education. In addition, the "survival scenario" task, while effective at prompting collaborative discussion, may not have been sufficiently typical of a small-group meeting in a workplace or education settings, in which participants may be familiar with the topics of discussion and may know their conversational partners in advance. While our study included a training and familiarization procedure, this study was not able to examine the speech characteristics of long-term or habitual users of an ASR communication application in this setting. In future work, we intend to conduct deployments of this application in actual workplace settings with DHH individuals, who would use the tool during impromptu small-group meetings, of longer duration than the brief meetings in this study.

We also plan to consider additional baseline conditions: (1) a hearing person speaking to another hearing individual or (2) a hearing person speaking into ASR with the screen seen only by the DHH individual. Such a comparison would determine whether hearing users' speech in small-group meetings when they are using ASR to speak to DHH peers is fundamentally different than their day-to-day speech behavior when speaking with other hearing people – and whether seeing ASR output influences speech behavior. We would also like to explore whether hearing users influence each other's speech patterns during the meetings.

In addition, our study included only one design variation for this application (with or without one type of confidence markup). In future work, we intend to investigate alternative designs of such a tool, which might differ in their effect on participants' speech features. Our ultimate goal is to better understand the interaction between various design parameters and users' communication behaviors and success, to support DHH users in a variety of contexts.

ACKNOWLEGMENTS

This material is based upon work supported by the National Science Foundation under awards 1400802, 1462280, 1460894, 1746056, as well as the NSF Graduate Research Fellowship Program and National Technical Institute for the Deaf. We are grateful for the contributions of Tousif Ahmed, Ben Kassman, Catherine Seita, Jason Antal, Jaron Rekhop, Abraham Glasser, and Larwan Berke.

REFERENCES

[1] ANSI/ASA S3.5-1997. 1997. *American National Standard Methods for Calculation of the Speech Intelligibility Index*. Acoustical Society of America

(ASA) and American National Standards Institute (ANSI), New York, NY, USA.

[2] Keith Bain, Sara H. Basson, and Mike Wald. 2002. Speech recognition in university classrooms: liberated learning project. In *Proceedings of the fifth international ACM conference on Assistive technologies (Assets '02).* ACM, New York, NY, USA, 192-196. http://dx.doi.org/10.1145/638249.638284

[3] Jon P. Barker, Ricard Marxer, Emmanuel Vincent, Shinji Watanabe. 2017. The CHiME challenges: Robust speech recognition in everyday environments. In: Watanabe S., Delcroix M., Metze F., Hershey J. (eds.), *New Era for Robust Speech Recognition.* Springer, Cham, 327-344. https://doi.org/10.1007/978-3-319-64680-0_14

[4] Larwan Berke, Christopher Caulfield, and Matt Huenerfauth. 2017. Deaf and Hard-of-Hearing Perspectives on Imperfect Automatic Speech Recognition for Captioning One-on-One Meetings. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '17).* ACM, NY, NY, USA, 155–164. https://doi.org/10.1145/3132525.3132541

[5] Larwan Berke, Sushant Kafle, Matt Huenerfauth. 2018. Methods for Evaluation of Imperfect Captioning Tools by Deaf or Hard-of-Hearing Users at Different Reading Literacy Levels. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI'18)*. ACM, New York, NY, USA. https://doi.org/10.1145/3173574.3173665

[6] Paul Boersma and David Weenink. 2018. *Praat: doing phonetics by computer [Computer program].* Version 6.0.39. Retrieved April 3, 2018, from http://www.praat.org/

[7] Denis Burnham, Sebastian Joeffry, Lauren Rice. 2010. Computer- and Human-Directed Speech Before and After Correction. In *Proceedings of the 9th Speech Science and Technology Conference 2010*, Melbourne, Australia. Australian Speech Science and Technology Association.

[8] Esteban Buz, Michael K. Tanenhaus, and T. Florian Jaeger. 2016. Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations. *Journal of Memory and Language* 89, Supplement C (2016), 68 – 86. https://doi.org/10.1016/j.jml.2015.12.009

[9] ELAN (Version 5.0.0-beta) [Computer software]. 2017. Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from https://tla.mpi.nl/tools/tla-tools/elan/

[10] Lisa B. Elliot, Michael Stinson, James Mallory, Donna Easton, and Matt Huenerfauth. 2016. Deaf and Hard of Hearing Individuals' Perceptions of Communication with Hearing Colleagues in Small Groups. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '16).* ACM, NY, NY, USA, 271–272. https://doi.org/10.1145/2982142.2982198

[11] Lisa B. Elliot, Michael Stinson, Syed Ahmed, and Donna Easton. 2017. User Experiences When Testing a Messaging App for Communication Between Individuals Who Are Hearing and Deaf or Hard of Hearing. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '17).* ACM, NY, NY, USA, 405–406. https://doi.org/10.1145/3132525.3134798

[12] Maria Federico and Marco Furini. 2012. Enhancing Learning Accessibility Through Fully Automatic Captioning. In *Proceedings of the International CrossDisciplinary Conference on Web Accessibility (W4A '12).* ACM, New York, NY, USA, Article 40, 4 pages. https://doi.org/10.1145/2207016.2207053

[13] Ira R. Forman, Ben Fletcher, John Hartley, Bill Rippon, and Allen Wilson. 2012. Blue Herd: Automated Captioning for Videoconferences. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '12).* ACM, NY, NY, USA, 227–228. https://doi.org/10.1145/2384916.2384966

[14] Carrie Lou Garberoglio, Stephanie Cawthon and Mark Bond. 2016. *Deaf People and Employment in the United States: 2016.* National Deaf Center on Postsecondary Outcomes. Retrieved on April 15, 2018, from https://www.nationaldeafcenter.org/sites/default/files/Deaf%20Employment%20Report_final.pdf

[15] Jay Hall. 1983. The rejection of deviates as a function of threat. Doctoral Dissertation, University of Texas.

[16] Hearing Loss Association of America. 2017. Basic Facts About Hearing Loss. Retrieved December 17, 2017 from http://www.hearingloss.org/content/basic-facts-about-hearing-loss

[17] Sushant Kafle and Matt Huenerfauth. 2017. Evaluating the Usability of Automatically Generated Captions for People Who Are Deaf or Hard of Hearing. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '17).* ACM, NY, NY, USA, 165–174. https://doi.org/10.1145/3132525.3132542

[18] Saba Kawas, George Karalis, Tzu Wen, and Richard E. Ladner. 2016. Improving Real-Time Captioning Experiences for Deaf and Hard of Hearing Students. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '16).* ACM, NY, NY, USA, 15–23. https://doi.org/10.1145/2982142.2982164

[19] Ronald R. Kelly. 2015. The Employment and Career Growth of Deaf and Hard of Hearing Individuals. *Raising and Educating Deaf Children: Foundations for Policy, Practice, and Outcomes.* Retrieved from http://www.raisingandeducatingdeafchildren.org/2015/01/12/the-employment-and-career-growth-of-deaf-and-hard-of-hearing-individuals/

[20] S. Koster. 2001. Acoustic-phonetic characteristics of hyperarticulated speech for different speaking styles. In *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing.* (Cat. No.01CH37221). IEEE. https://doi.org/10.1109/icassp.2001.941054.

[21] Raja Kushalnagar, Walter Lasecki, Jeffrey Bigham. 2014. Accessibility evaluation of classroom captions. *ACM Trans. Access. Comput.* 5, 3, Article 7 (Jan. 2014), 24 pp. http://dx.doi.org/10.1145/2543578

[22] Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. 2012. Real-time captioning by groups of non-experts. In *Proceedings of the 25th annual ACM symposium on User interface software and technology (UIST '12).* ACM, New York, NY, USA, 23-34. https://doi.org/10.1145/2380116.2380122

[23] Uwe Ligges. 2017. *Package 'tuneR' (version 1.3.2): Analysis of Music and Speech.* Retrieved on April 15, 2018, from https://cran.r-project.org/web/packages/tuneR/tuneR.pdf

[24] Steven R. Livingstone, Deanna H. Choi, and Frank A. Russo. 2014. The influence of vocal training and acting experience on measures of voice quality and emotional genuineness. *Frontiers in Psychology* 5 (Mar 2014). https://doi.org/10.3389/fpsyg.2014.00156

[25] Catherine L. Lortie, Mélanie Thibeault, Matthieu J. Guitton, and Pascale Tremblay. 2015. Effects of age on the amplitude, frequency and perceived quality of voice. *AGE* 37, 6 (Nov 2015). https://doi.org/10.1007/s11357-015-9854-1

[26] James R. Mallory, Michael Stinson, Lisa Elliot, and Donna Easton. 2017. Personal Perspectives on Using Automatic Speech Recognition to Facilitate Communication Between Deaf Students and Hearing Customers. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '17).* ACM, NY, NY, USA, 419–421. https://doi.org/10.1145/3132525.3134779

[27] National Institute of Standards and Technology. 2015. *Sclite (version 2.4.10).* Retrieved on April 10, 2018, from ftp://jaguar.ncsl.nist.gov/pub/sctk-2.4.10-20151007-1312Z.tar.bz2

[28] Nuance. 2017. *Documentation for the SpeechKit 2 SDK for Android.* Retrieved on June 1, 2018, from https://developer.nuance.com/public/Help/DragonMobileSDKReference_Android/index.html

[29] Sharon Oviatt, Gina-Anne Levow, Elliott Moreton, and Margaret MacEachern. 1998. Modeling Global and Focal Hyperarticulation during Human–Computer Error Resolution. *J. Acoust. Soc. Amer.* 104, 3080–3098. https://doi.org/10.1121/1.423888

[30] Benjamin Picart, Thomas Drugman, and Thierry Dutoit. 2010. Analysis and synthesis of hypo-and hyperarticulated speech. In *Proceedings of the Seventh ISCA Workshop on Speech Synthesis.*

[31] Koenraad S. Rhebergena, Johannes Lyzenga, Wouter A. Dreschler, Joost M. Festen. 2010. Modeling speech intelligibility in quiet and noise in listeners with normal and impaired hearing. The Journal of the Acoustical Society of America 127(3), 1570-1583. Acoustical Society of America. https://doi.org/10.1121/1.3291000

[32] RIT News. 2018. RIT/NTID and Microsoft launch partnership for AI driven accessibility. Retrieved on April 15, 2018, from http://www.ntid.rit.edu/news/ritntidand-microsoft-launch-partnership-ai-driven-accessibility

[33] Rein Ove Sikveland. 2006. How do We Speak to Foreigners? — Phonetic Analyses of Speech Communication between L1 and L2 Speakers of Norwegian. *Working Papers* 52, 109–112. Centre for Language and Literature, Lund University, Sweden.

[34] Hagen Soltau and Alex Waibel. 1998. On the Influence of Hyperarticulated Speech on Recognition Performance. In *Proceedings of the 5th International Conference on Spoken Language Processing, ICSLP 1998,* Sydney, Australia.

[35] Amanda J. Stent, Marie K. Huffman, and Susan E. Brennan. 2008. Adapting Speaking After Evidence of Misrecognition: Local and Global Hyperarticulation. *Speech Commun*. 50, 3 (March 2008), 163–178. https://doi.org/10.1016/j.specom.2007.07.005

[36] Michael Stinson, Carly Linneah, Jonathan MacDonald, and Chelsea Powers. 2014. Using technology to improve communication in small groups with deaf and hearing students. Presentation at *the 2nd Annual Effective Access Technology Conference*. Rochester Institute of Technology, Rochester, NY, USA.

[37] Hironobu Takagi, Takashi Itoh, and Kaoru Shinkawa. 2015. Evaluation of Realtime Captioning by Machine Recognition with Human Support. In *Proceedings of the 12th Web for All Conference (W4A '15).* ACM, New York, NY, USA, Article 5, 4 pages. https://doi.org/10.1145/2745555.2746648

[38] M. Wald. 2011. Crowdsourcing Correction of Speech Recognition Captioning Errors. In *Proceedings of the*

International Cross-Disciplinary Conference on Web Accessibility (W4A '11). ACM, New York, NY, USA, Article 22, 2 pages. https://doi.org/10.1145/1969289.1969318

[39] Gregory R. Warnes. 2013. *Calculating Speech Intelligibility Index (SII) using R.* Retrieved on April 12, 2018, from https://cran.r-project.org/web/packages/SII/vignettes/SII.pdf

[40] W. Xiong, J. Droppo, X. Huang, F. Seide, . Seltzer, A. Stolcke, D. Yu, G. Zweig. 2016. Achieving human parity in conversational speech recognition. *Computing Research Repository (CoRR),* http://arxiv.org/abs/1610.05256

[41] Jiahong Yuan, Mark Liberman, Christopher Cieri. 2007. Towards an Integrated Understanding of Speaking Rate in Conversation. In *Proceedings of the 16th International Congress of Phonetic Sciences ICPhS XVI, 6-10 August 2007, Saarbrücken, Germany,* 1337-1340, University of Saarbrüken.