Bridge Communications: Kiosk Conversation Tools for d/Deaf and Hard of Hearing and Hearing

YUGO IWAMOTO, Rochester Institute of Technology, United States of America

Communication barriers between d/Deaf and Hard-of-Hearing (d/DHH) and hearing people persist in everyday contexts where interpreters or captioning services are unavailable. This study introduces a kiosk-based, multi-modal communication prototype that integrates speech-to-text (STT) and typing inputs to facilitate real-time, accessible conversations across hearing and d/DHH users. The prototype was built using Django and VOSK for offline speech recognition. The system enables instant interactions via a shared monitor and input from mobile devices. Through a mixed-methods evaluation involving 12 participants (six d/DHH, six hearing), the researcher examined their usability, workload, and input modality experiences. Findings indicate that the tool requires low mental and physical demand (NASA TLX), with high usability ratings for typing. Moderate performance for speech input, particularly challenged by environmental noise and expressiveness, was observed. Participants appreciated the system's privacy-preserving design, ease of onboarding, and flexibility in modality selection. Yet, they emphasized the need for improved emotional expressiveness and input-switching mechanisms. This work provides the design implications for inclusive, low-barrier communication tools and highlights opportunities to support d/DHH-hearing interaction in public, informal, or interpreter-limited scenarios.

Additional Key Words and Phrases: Accessibility, Human-Computer Interaction, Assistive Technology, d/Deaf and Hard of Hearing, Multi-modal Communications

ACM Reference Format:

Yugo Iwamoto. 2025. Bridge Communications: Kiosk Conversation Tools for d/Deaf and Hard of Hearing and Hearing. 1, 1 (August 2025), 19 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Communication barriers between hearing and d/Deaf and hard-of-hearing (d/DHH) individuals remain a persistent obstacle to achieving equitable participation in daily life [15]. Although American Sign Language (ASL) interpreters, captioning services, and other assistive technologies provide essential support, they are often unavailable for spontaneous, one-on-one, or informal interactions [6]. This lack of real-time assistive technology can significantly impact the independence, inclusion, and social engagement of d/DHH individuals in educational, professional, and public environments. The d/DHH community is highly diverse, with varying degrees of hearing loss, language preferences, and communication needs. Moreover, many people with hearing loss do not use ASL or any formal sign language, which underscores the pressing need for alternative, language-based communication tools that extend beyond traditional interpreter services [29].

The d/DHH community faces a variety of pervasive communication barriers that hinder their full participation in societal, educational, and professional settings [7]. Communication barriers are not just logistical obstacles—they represent systemic inequities that affect autonomy, access to information, and social inclusion for d/DHH individuals. One significant challenge is the reliance on traditional communication aids, such as interpreters, which are sometimes unavailable [19]. This shortage is particularly evident in high-demand fields such as healthcare, education, and legal services [6]. The dependency on certified ASL interpreters creates a systemic gap, leaving d/DHH individuals without access to critical conversations or services. Another issue lies in the lack of ASL proficiency among hearing individuals,

Author's Contact Information: Yugo Iwamoto, yi9686@rit.edu, Rochester Institute of Technology, Rochester, New York, United States of America.

2025. Manuscript submitted to ACM

which contributes to limited interaction and a broader lack of understanding of the needs of the d/DHH community [35]. This gap perpetuates feelings of frustration, isolation, and exclusion, particularly in educational environments where inclusive teaching practices are critical for student success. A lack of awareness and understanding of the unique needs of d/DHH individuals often leads to communication avoidance or exclusionary practices, creating additional hurdles for meaningful engagement. These barriers persist even in everyday scenarios where formal accommodations like interpreters or captioning services are not feasible. As a result, many d/DHH individuals may avoid or struggle through spontaneous conversations in public, educational, or workplace settings, limiting their full participation in society. Addressing these barriers is essential for promoting equitable communication and fostering environments where all individuals can contribute and connect.

This study investigates whether a kiosk-based, speech-to-text (STT) and typing multi-modal system can serve as an effective and user-friendly alternative communication channel for d/DHH individuals, particularly those who do not use ASL. Research by Shezi et al. [36] indicates that accessibility technologies such as STT can reduce barriers to communication by utilizing mobile devices. Hence, this study explores designing and evaluating a tool to use in real-time, particularly in impromptu conversations and settings where formal support is not readily available. The primary objectives of this research are to:

- Designing a kiosk-style real-time communication tool combining STT and typed input, and evaluating its usability and accessibility with hearing and d/DHH users.
- Examining how the system performs in supporting spontaneous, bidirectional communication between hearing and d/DHH users.
- Identifying technical and design challenges affecting speech recognition accuracy, user comprehension, and interaction flow.
- Explore user satisfaction, perceived usefulness, and situational appropriateness of the tool across user groups.

Based on these objectives, the research is guided by the following questions:

- How can kiosk-based communication tools be designed to support immediate, user-friendly conversations between hearing individuals and d/DHH users who do not use sign language?
- In what ways does the integration of STT and typing input improve or hinder communication flow, accuracy, and user comprehension?
- What usability challenges and contextual limitations emerge when deploying such tools in real-world environments with d/DHH and hearing people?
- How do d/DHH and hearing participants perceive the tool's effectiveness, accessibility, and ease of use in supporting inclusive communication?

To address these questions, this project investigates the usability and feasibility of a two-way real-time communication system powered by STT and text input. The project, through designing a tool, integrates VOSK [38], a lightweight and offline speech recognition package, to transcribe spoken English into on-screen text. Additionally, a typing interface enables d/DHH users to contribute to the conversation in written form, creating a bidirectional, accessible dialogue space between d/DHH and hearing users. To enable these functions, the backend is developed in Python Django, which handles session control and communication logic. Meanwhile, the front-end utilizes HTML, CSS, and JavaScript to create an intuitive and responsive user experience. The system architecture consists of a user's mobile phone for input (via microphone or typing) and a monitor showing live conversation. At its core, the system is optimized for spontaneous use in settings where interpreters or captionists are not immediately available. Understanding how the

system performs under different environments is essential to assess its potential role in closing accessibility gaps for the d/DHH community. Importantly, this system is not intended to replace ASL interpreters, captionists, or other existing solutions [28]. Instead, it is designed to support by offering an additional communication channel, which is beneficial for casual, unstructured, or impromptu conversations where traditional support is not feasible. For d/DHH individuals who are fluent in written English, the tool may offer a practical and empowering means to independently engage with hearing peers.

In summary, this study aims to explore how real-time STT and typing-based multi-modal communication tools can enhance inclusive interaction between d/DHH and hearing users. By identifying usability challenges, speech recognition limitations, and opportunities for design refinement, the findings contribute to the growing body of research on accessible human-computer interaction (HCI). The broader goal is to inform the design of equitable, scalable solutions that improve communication access for diverse users in everyday settings.

2 Related Work

In what follows, an analysis of prior work on communication accessibility, Automatic Speech Recognition (ASR), emotional expressiveness in speech interfaces, and multi-modal interaction systems designed for accessibility is provided. Effective communication between d/DHH and hearing individuals remains a persistent challenge [18], especially in unscripted, everyday contexts where interpreters or captioning services are unavailable. Although a range of assistive technologies, including STT applications, typing interfaces, and mobile captioning tools, have emerged to support access, few systems offer integrated real-time multi-modal solutions. Research in HCI and accessibility highlights the need for communication tools that go beyond accurate transcription to support conversational flow, emotional nuance, and user autonomy [11].

2.1 Communication Accessibility for d/Deaf and Hard-of-Hearing Users

Persistent communication barriers between d/DHH and hearing individuals continue to limit full participation in educational, professional, and public life [7]. A key barrier is the limited availability of qualified ASL interpreters, particularly for spontaneous, informal, or one-on-one conversations that occur outside of scheduled events or institutional settings [6]. Interpreter shortages are especially significant in contexts that require specific domain knowledge, such as medical or technical environments [22].

In higher education settings, d/DHH students report difficulties accessing real-time dialogue, especially in group discussions, peer collaboration, or interactions with instructors who lack ASL proficiency [31]. Even when interpreters or captionists are provided, logistical delays, turn-taking lags, or misinterpretations can disrupt the conversational flow and lead to exclusion [9]. These challenges contribute to increased cognitive load, reduced classroom engagement, and long-term barriers to academic success.

In addition, many people with d/DHH, particularly those who are late-deaf or oral communicators, may not use ASL at all [6]. This linguistic diversity within the d/DHH community further complicates the design of universal communication solutions, as assistive technologies often assume a binary choice between speech and signing. While regions like Rochester, NY, are recognized for their strong d/DHH community and accessible infrastructure [30], such examples remain exceptions rather than the norm across the United States and globally.

These persistent access limitations highlight the urgent need for inclusive, flexible, and on-demand communication tools that do not rely on shared language knowledge. Technologies that support multi-modal interaction—such as

real-time speech-to-text, typing, and visual interfaces—offer promising alternatives for bridging communication gaps in ad hoc encounters, public spaces, and multi-lingual contexts.

2.2 Speech-to-Text Systems and Their Limitations

ASR has emerged as a promising solution to enhance communication access in a variety of settings, including education, healthcare, and everyday social interactions. Cloud-based services such as Google Speech-to-Text [13] and Microsoft Azure [27] offer highly accurate transcription by leveraging deep learning models and large-scale training data. These systems are widely adopted due to their speed, language coverage, and integration into mainstream devices. However, they require consistent Internet connectivity and rely on the transmission of voice data to remote servers for processing, raising significant privacy concerns, particularly in sensitive or informal conversations [1, 4]. For d/DHH users, these limitations can restrict spontaneous usage in private or low-connectivity environments. In response, offline-capable ASR models, such as VOSK, have gained traction as lightweight, real-time transcription solutions that maintain user privacy by performing all processing locally. While beneficial for mobile and kiosk-based use cases, such models face notable limitations in transcription accuracy, especially in acoustically complex environments. The reliability of recognition declines with the presence of nonstandard speech patterns, regional dialects, background noise, or multiple overlapping speakers [12, 20]. These limitations are particularly salient in accessibility contexts, where diverse speech characteristics and environmental constraints are common. Technological advances such as beamforming microphone arrays and mmWave-enhanced speech capture have been proposed to mitigate noise and improve source separation [24]. However, these solutions often require specialized hardware and are not yet broadly accessible or integrated into low-cost communication tools. Additionally, STT output often lacks punctuation and prosodic cues, producing a monotonous stream of words that demands greater cognitive effort to interpret, especially for d/DHH readers who rely on text as their primary communication channel [21]. Finally, while ASR systems are designed to optimize word recognition accuracy, they are generally not tuned for readability or conversational flow. For real-time applications where speed and comprehension are critical, this trade-off can limit the perceived usefulness of the technology. Accessibility-focused communication tools must therefore address not only transcription accuracy but also output clarity, expressiveness, and adaptability to user preferences and environmental conditions.

2.3 Multi-modal Communication Tools and Interface Models

Recent tools such as Deaf Chat [36] and TigerChat [32] have demonstrated the potential of STT interfaces to bridge communication gaps between d/DHH individuals and hearing users. These systems primarily support mobile-based, one-to-one interactions and typically present transcriptions in a linear, chat-style format. While useful in personal or support-driven contexts, their designs are often limited to a single input modality—either speech or text—and do not provide a cohesive view of multiple participants interacting simultaneously.

In contrast, effective multi-modal interaction in d/DHH communication settings requires more than just alternating between input types. It calls for synchronous integration of diverse modalities to better reflect the fluid, turn-based nature of human conversation. The prior HCI research emphasizes the importance of modality flexibility, concurrent input handling, and interface designs that visually support turn-taking and speaker identity. For example, Lee et al. [23] introduced a multi-modal communication system for d/DHH users that incorporates gesture and speech inputs, highlighting the need for user-centered design in mixed reality environments. Similarly, Desai et al. [10] explored the role of speech reading in online communication, identifying key environmental, sociocultural, and technical factors that

influence accessibility, and proposing directions for designing future communication tools that enhance comprehension and expression.

These insights collectively underscore the value of integrated multi-modal systems that enable users to communicate through their preferred methods—whether typing or speaking—while maintaining a shared, synchronous conversational space. Such designs are especially beneficial in cross-ability contexts, where participants may rely on different modalities due to sensory differences or situational demands.

2.4 Emotion and Expressiveness in Speech Interfaces

A persistent limitation of ASR-based communication tools is their inability to convey emotional nuance. Participants in both previous research and our study consistently noted that speech recognition systems often omit essential expressive signals, such as inflection, humor, emphasis, and punctuation, that contribute to conversational tone and affective engagement [26]. This absence can hinder comprehension, particularly in accessibility contexts where emotional clarity and conversational flow are critical. While recent work has begun to explore the integration of emotion recognition into ASR pipelines, it remains limited in robustness and is not yet practical for real-time deployment in everyday settings [25]. Additionally, foundational research has highlighted the inherent trade-offs in ASR technologies, including sensitivity to noise, latency, and modeling complexity—challenges that are especially pronounced in offline or resource-constrained environments [33]. These limitations were evident in our implementation, where recognition performance was affected by background noise, speech pace, and speaker variation. To move beyond basic transcription, future communication tools must evolve to interpret and convey the emotional tone of speech. This capability is especially vital in inclusive systems designed for d/DHH users, where subtle affective signals may otherwise be lost in translation.

3 Method

Our system builds on the foundation by introducing a kiosk-based, shared-display architecture that integrates both real-time speech recognition and typed input from mobile devices. Unlike prior tools that operate independently on users' phones, our prototype enables a centralized, visible conversation stream displayed on a shared monitor. This model supports spontaneous interaction without requiring app downloads or prior setup, and enables dyadic or small group conversations where participants can independently choose their preferred input method, without sacrificing visibility or control over the interaction.

3.1 Prototype

The prototype system is designed as a kiosk-based communication tool consisting of a monitor and user-operated mobile devices Figure 1. The prototype is a web application whose front end is implemented using HTML and styled with CSS, while user interactions are handled via JavaScript. The backend is developed using Python and Django, which manages real-time speech recognition and session control. The detailed functionality of the prototype, such as STT functionality, is powered by the VOSK small English model, integrated into the backend for real-time transcription. The VOSK small English model [38] is a lightweight, offline-capable speech recognition package optimized for real-time transcription with low resource consumption. The implementation of STT enables the provision of a consistent user experience for multi-modal products, rather than relying on individual device features. The typing functionality is implemented using JavaScript, enabling quick and intuitive input. The complete system is deployed using Docker and hosted in Render [34] to ensure portability and scalability for future testing and development.

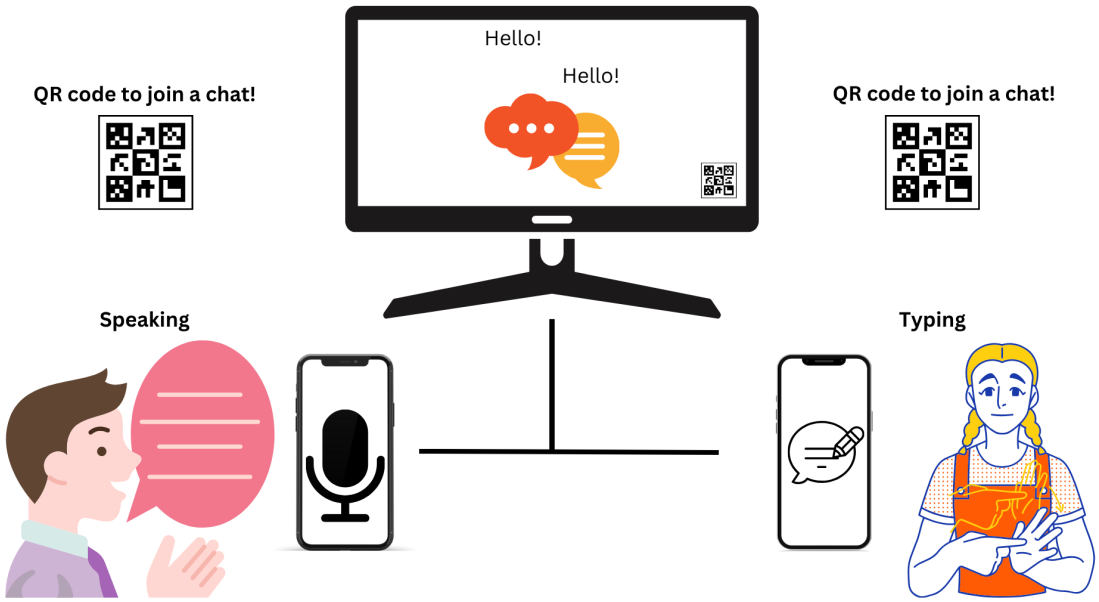

Fig. 1. Visual representation of the kiosk-based communication system. Participants join a shared conversation by scanning a QR code on their mobile device, choosing either a speech (left) or typing (right) input method. Spoken or typed messages are transcribed and displayed in real time on a central monitor, supporting inclusive communication between hearing and DHH individuals.

Upon entering the interface, users are greeted with a QR code displayed on the primary monitor (as shown in Figure 1). Scanning the QR code with a mobile device redirects the user to a login page, where they are prompted to enter their name and select an input method: speech or typing. After submission, the user interface transitions to the selected input mode, and the main display synchronously updates to a shared chatroom view. Conversations can then proceed, with participants speaking or users typing messages. In the conversation, messages are displayed with their names. In addition, messages from participants using speech recognition are displayed in blue font, and the other messages are displayed in green font.

3.2 Procedures

Recruiting participants via emails, reaching out to the d/DHH community, mentioning the purpose of the study, and the intention. The study involved six pairs of participants (12 total), each consisting of a participant in d/DHH and a hearing participant, who joined a 30-minute session expecting to collect qualitative and quantitative data [5]. The flow of the study is shown in Figure 2, and the demographics of participants are shown in the Appendix 1. Prior to the session, participants completed a pre-study survey to gather demographic information and availability. The room was set up with a monitor with a multi-modal communication prototype and two chairs facing each other. As they walked in, participants reviewed and signed a consent form. The researcher introduced the purpose of the study and the functionality of the prototype, but deliberately withheld detailed instructions for login to evaluate the intuitiveness and usability of the onboarding process. The scenario provided is that "You are in a public environment, and no ASL interpreter or captionist is available. You have to have a conversation without exchanging contact information." The

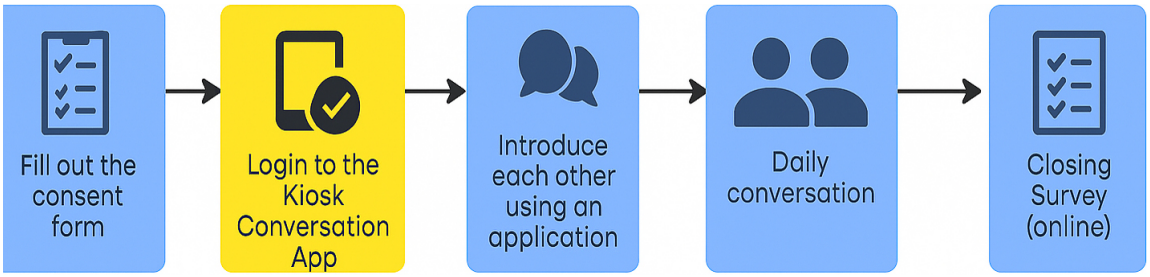

Fig. 2. Flow of study procedures for the kiosk conversation tool. Participants began by completing a consent form, followed by logging into the system. They then introduced themselves using the application, engaged in a real-time conversation using speech or typing input, and concluded the session with an online post-study survey.

input methods were opened up to decide the participants' preference; therefore, pair No.4 only used the typing feature in the session. Once participants logged into the system independently, they were prompted to introduce themselves for three minutes, followed by a 10–15 minute unstructured conversation using the prototype. Conversation topics were left open-ended to simulate natural interactions and to allow for observational analysis of communication flow and usability across different input modes. After the interaction, the participants completed a post-completion survey assessing their experience with the system, including the accuracy of speech recognition, the ease of use, and the perceived quality of the conversation.

The Rochester Institute of Technology's Institutional Review Board (IRB) reviewed and approved this study.

4 Findings

The study sessions, including interactions and post-completion conversations, were voice- and screen-recorded using Zoom. A researcher conducted a thematic analysis [2, 3, 8, 37] of survey responses, observation notes, and post-study reflections from 12 participants (six d/DHH and six hearing), shown in Appendix 1. For instance, screen recording yielded speech recognition failure codes related to the name of the city and personal information, or the post-completion survey highlighted "Slowly and talking to the microphone helps improve the recognition." From the codes of thematic analysis combining with quantitative analysis as mixed-method [14] led to six themes, "Speech Recognition Fragility in Real-World Settings," "Typing as a Reliable and Accessible Modality," "Ease of Use and Usability," "Emotional Expression and Communication Flow," and "Presentation of the Chatroom and Visual Presentation" highlight the experiential qualities of using a kiosk-based multi-modal communication prototype and uncover usability tensions, input preferences, and emotional constraints. The conversation topics the participants had were listed in the Appendix 1.

One of the strategies to evaluate the workload and physical load was using NASA TLX [16, 17], which allowed a researcher to evaluate the task load and future use in the real world. In addition, the usability rating of the overall experience is shown in Figure 3. NASA TLX results indicate lower physical and mental demands associated with using the Kiosk Conversation app, suggesting its potential for adoption in real-world settings as a multi-modal communication tool. The mean of mental demand was 3.8 out of seven, with a standard error of mean (SEM) of 0.52, the mean of physical demand was 3.1 with a SEM of 0.54, the mean of temporal demand was 3.6 with a SEM of 0.5, the mean of effort was 3.1 with a SEM of 0.57, the mean of frustration was 2.7 with SEM of 0.53, and the mean of performance was 5.2, with SEM of 0.36. In addition to the NASA TLX, the mean of usability was 3.9 out of five with a SEM of 0.29, and P12 mentioned that *"Super easy to use and pretty intuitive"*. On top of that, P7 stated that *"It can be used for group conversation*

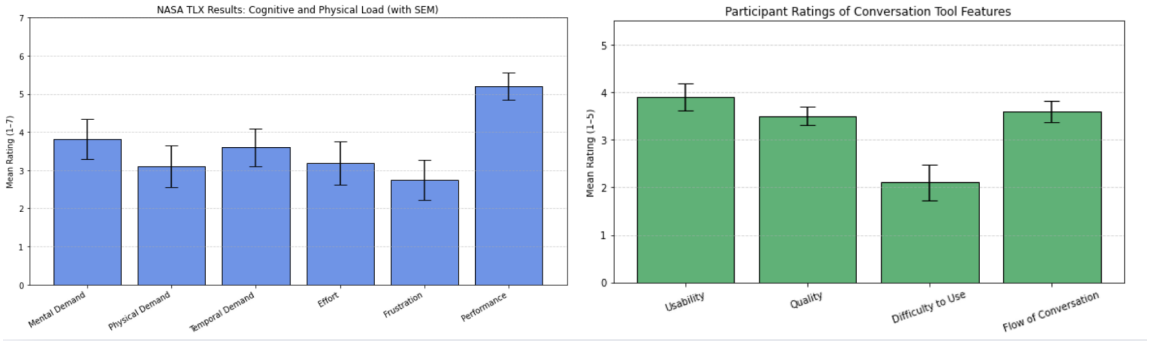

Fig. 3. Participant evaluation of the conversation tool. Left: NASA Task Load Index (TLX) ratings show moderate cognitive, physical, and temporal demands, low frustration, and high perceived performance (1 = low, 7 = high). Right: Usability ratings indicate that the tool was generally easy to use, of high quality, low in difficulty, and supported a smooth conversational flow (1 = poor, 5 = excellent). Error bars represent the standard error of the mean.

if people are loud enough to overlay each other's devices", which suggests the future expansion and possibility. On the other hand, P6 mentioned that *"Changing the input method is challenging."* as there were no buttons allowing users to change the input method on the conversation interface. As some of them attempted to switch input methods during a conversation, it was not an intuitive process. Participants reported the mean quality was 3.5 with a SEM of 0.19. P10 mentioned that *"The conversation was successful except for a minor misinterpretation of speech."* Participants reported the mean of difficulty to use was 2.1, and some reasoning as P12 stated that *"Easy to use and very simple"* and P6 reported that *"Needed to hold the button to speak. I am more used to on/off buttons."* The mean of the flow of conversation was 3.6 with a SEM of 0.22. P10 commented that *"It is pretty smooth to find the flow of communication. The flow of conversation went well, besides the natural time between reading and responding to each other."* The analysis yielded six major themes that reflected user experience with the prototype: the accuracy and fragility of speech input, the reliability of typing input, usability and learnability, emotional expression and flow, interface design, and perceived social impact. These insights are supplemented with quantitative measures of mental load, usability, and feature-specific ratings.

4.1 Speech Recognition Fragility in Real-World Settings

Participants who used the speech input option highlighted that the performance of the speech recognition system was highly context-sensitive. Specifically, the accuracy of transcription was dependent on several environmental and personal factors, including background noise, speech clarity, accent or dialect, and proximity to the microphone. While the offline small VOSK model used in this prototype was sufficient in quiet or semi-controlled settings, it exhibited limitations when faced with overlapping speech or ambient sounds, a challenge common to many STT systems, but especially pronounced in lightweight, offline models.

Quantitative data collected through participant surveys reflect these constraints. As shown in Figure 4, the mean rating for speech recognition accuracy was 3.4 out of five with a SEM of 0.24, suggesting moderate reliability. The mean processing speed was higher at 4.2 with a SEM of 0.2, indicating that participants generally found the system responsive in terms of latency. However, noise reduction also scored a moderate mean of 3.4 with a SEM of 0.24, reinforcing that environmental interference remained a significant barrier to achieving clear and consistent transcriptions.

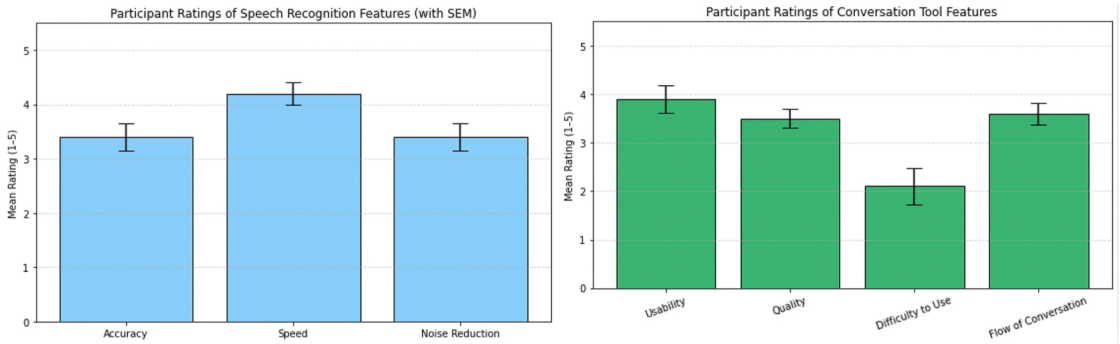

Fig. 4. Participant ratings of speech and typing input features. Left: Speech recognition was rated highly in processing speed (M = 4.2), with moderate ratings for accuracy and noise reduction (both M = 3.4), indicating responsiveness but variable recognition quality. Right: Typing input was rated highly for accuracy (M = 4.4), accessibility (M = 4.1), and feedback speed (M = 4.1), while emotional expressiveness received slightly lower scores (M = 3.8). Error bars represent the standard error of the mean.

Several participants expanded on these issues in their open-ended feedback. For example, P2 noted *"It was accurate 90% of words, [but] some specific words or ways of saying in my dialect caused the app to misinterpret, but that was not as common as I would expect."* This comment illustrates how dialectal variation, even within the same language, can affect recognition outcomes, despite generally positive impressions of the system's baseline performance. It also suggests that while users are tolerant of occasional errors, they are more likely to notice and be disrupted by misrecognition of personal or contextually important terms, such as names or local slang.

P9 echoed similar concerns, stating *"I am not sure if it was the way I spoke or how fast I was speaking that there were a handful of errors outputting from what I spoke. I enjoyed the tool, though, and would probably use it again if I had to speak with someone who was D/HOH."* This reflection highlights a key issue in STT-based communication: even minor delays or inaccuracies can lead to breakdowns in conversational flow, particularly in spontaneous or socially sensitive interactions. Despite these imperfections, P9's willingness to reuse the tool also demonstrates that users recognize the broader value of inclusive communication, even when technical flaws are present.

The researcher observed that recognition errors often disrupted the pacing of conversation, requiring participants to pause, repeat themselves, or clarify messages manually. These disruptions were especially problematic when names, acronyms, or context-specific terms were misrecognized, terms that often lack clear phonetic anchors or that differ across cultural or linguistic groups. For instance, one participant had their name transcribed incorrectly multiple times, causing confusion for the conversation partner and leading to a momentary derailment in topic flow.

Despite these shortcomings, it is important to note that speech was not the only input method available in the system. Participants who encountered issues with speech recognition were able to switch to typing, which served as a fallback modality and reduced frustration. However, as discussed in the usability and emotional expression findings, the switching mechanism was not always intuitive or seamless, which occasionally led to hesitation or dropped contributions in real-time interactions.

4.2 Typing as a Reliable and Accessible Modality

Typing was consistently regarded by participants as a more reliable and accessible input modality, especially in environments with high background noise or where verbal communication was difficult or undesirable. Across user

groups, typing was perceived to offer better control over message content, greater emotional security, and reduced risk of misinterpretation compared to speech recognition. These perceptions were supported by both quantitative ratings and qualitative feedback collected during and after the study.

As shown in Figure 4, the mean score for accuracy of the typing input was 4.4 out of five with an SEM of 0.3, the highest among all features rated, while accessibility and feedback speed were a mean of 4.1, with an accessibility SEM of 0.26 and a feedback speed SEM of 0.31. These values indicate that participants found the typing experience both intuitive and efficient. Although expression of feelings received a slightly lower mean of 3.8 with a SEM of 0.4, it was still rated positively, suggesting that while typing may limit spontaneous affective expression, it offers a stable platform for thoughtful communication.

Several participants expanded on these experiences. P9 reflected on the emotional affordances of typing: *"It was nice to use my phone; it's more accessible because I feel more comfortable using chat. To express my feelings, I had to wait to see what my partner was trying to say, and then I could express my feelings, but it was pretty too late for me to do that."* This statement reveals both the comfort and familiarity associated with typed communication and a subtle limitation in timing, that is, a delay that can occur when typing in reaction to live input. While emotional nuance can be conveyed through carefully chosen words, the asynchronous rhythm of typing makes it harder to mirror or synchronize affect in real time.

In Pair 4, both participants opted for typing as their sole input method. P7 remarked *"It did feel like we texted each other through [the] prototype, but I do see the live conversation sitting in front of us."* This observation captures an interesting duality: although the system mimicked common chat experiences, the shared monitor gave it the feeling of a live, co-present dialogue. This blending of familiar device-based messaging with real-time shared space was viewed as a strength, especially for group contexts where different individuals might choose different modalities. In this way, typing was not just a fallback; it became a primary mode of communication that enabled clarity, autonomy, and control.

Participants also noted minor usability issues that, while not critical, influenced overall comfort. The most frequent complaint was the overlap between the on-screen keyboard and the chat window on mobile devices. This occasionally made it difficult to review previous messages while composing a new one. Despite this, most users adapted quickly, and no one indicated that it significantly disrupted their conversation. As with other input modes, interface refinements such as dynamic resizing of the chat window or repositioning the text input field could resolve this issue.

For many DHH participants, typing was particularly valued for its ability to regulate pace, allowing them to read, interpret, and respond on their terms. Unlike speech, which often demands synchronous turn-taking and quick interpretation, typing allowed users to pause, compose thoughts, and edit before sending a message. This form of control is especially important in cross-ability communication, where cognitive and sensory demands vary between users.

Taken together, these insights suggest that typing, while not as fast or expressive as face-to-face conversation, offers a stable, inclusive, and empowering modality for many users, especially in settings where speech input is hindered by noise, discomfort, or technological limitations. Its integration into the multimodal system not only improves overall usability but also reflects the diversity of communication needs across hearing and d/DHH individuals.

4.3 Ease of Use and Usability

While the System Usability Scale (SUS) is a widely used tool for evaluating general usability, it is less suited for capturing the specific, feature-level feedback and contextual nuances required in accessibility-focused, multi-modal interaction studies. Therefore, we opted for custom usability questions that aligned more closely with the goals of evaluating input

modality preferences and conversational flow unique to d/DHH communication. Overall, participants reported that the prototype was intuitive and easy to use, with minimal instruction needed to begin interacting. The system was designed with a lightweight onboarding flow, beginning with a QR code scan that led users to a mobile-friendly page where they could enter their name and select their preferred input method—either typing or speaking. This process was widely described as smooth and efficient, especially for first-time users. P10 simply described the experience as a *"straightforward interaction"*, underscoring the system's accessibility through streamlined design.

Several participants appreciated that the tool required no app installation, registration, or training, which contributed to a perception of low cognitive and procedural load. This accessibility-first approach was particularly important in the context of inclusive design, where assistive tools must not only function well but also be immediately usable by individuals with diverse communication preferences and abilities. The ability to quickly get started without assistance was cited as a strength, especially in contexts where spontaneous conversations are needed, such as public kiosks, waiting rooms, or educational settings.

Beyond initial onboarding, many participants commented positively on the clarity of the interface, such as the labeled input methods, color-coded messages, and responsive feedback. These features collectively enhanced usability by making it easy for users to identify their messages, understand the flow of conversation, and remain aware of who was contributing and in what modality.

At the same time, participants offered constructive feedback on specific elements of the user interface that affected their experience. The most prominent usability concern related to the speaking button is that it requires users to hold down the button to activate the microphone. While this method was intended to prevent unintended background capture and preserve turn-taking clarity, some users found it unintuitive. For example, P8 commented *"I felt like the UI could have been better with accessibility"*, indicating that the press-and-hold mechanism might be burdensome for users unfamiliar with this control paradigm or those with motor limitations. A toggle or tap-to-talk alternative might reduce strain and better align with existing voice assistant conventions. Additionally, participants suggested that more visual cues could improve clarity, such as highlighting the active input method or indicating whether the microphone was actively listening. This feedback points to opportunities for enhancing real-time system feedback, an important principle in usability that reinforces user confidence and helps prevent errors.

Despite these minor issues, multiple participants expressed a strong desire to use the prototype in real-life scenarios, which signals both satisfaction with the current design and optimism about its broader potential. The system's simple, QR-based entry, cross-device compatibility, and multi-modal flexibility were all noted as key facilitators of successful communication. Participants appreciated that they could engage in meaningful conversations without the need to exchange contact information, download applications, or explain the tool to their partners, making the experience both efficient and socially comfortable.

In summary, while the system was generally seen as usable and accessible, especially in terms of onboarding and message flow, participant feedback highlighted the importance of refining interaction mechanisms, particularly for speech input. With minor adjustments to input controls and interface feedback, the tool could further improve its usability for diverse populations and maintain its promise as an inclusive, on-demand communication solution.

4.4 Emotional Expression and Communication Flow

Across both input modalities, speech and typing, participants consistently highlighted the lack of emotional expressiveness in the prototype's output as a significant limitation. While the system successfully facilitated functional exchanges of information, the affective dimension of conversation, including tone, intention, and subtle emotional cues, was

frequently missing or misinterpreted. This gap affected not only the perceived naturalness of the interaction but also led to occasional confusion or disconnection between participants.

For instance, P1, who identifies as hard of hearing, reflected on the difficulty of interpreting emotional tone in person when relying on assistive cues. She commented *"It's kind of hard to express feelings when we meet in person. Some voices aren't matching up on this prototype, but it's pretty well to catch up with the voices."* Although P1 could hear some speech tones, she was not always able to distinguish precise words. The researcher observed that she frequently depended on her partner's facial expressions and body language to make sense of the message's affective content. This underscores the importance of nonverbal cues in d/DHH–hearing communication and how the prototype, by focusing solely on textual output, can fall short of replicating the emotional richness of in-person conversation.

Another participant, P11, also addressed the emotional flatness of the conversation interface. As someone who could hear but relied on the prototype's live chat captions to follow the interaction, P11 stated *"Speech recognition doesn't deliver emotion or tones as a question mark is not appearing on the questions."* This comment not only points to the absence of punctuation in the transcribed output—a common limitation in many ASR systems—but also emphasizes how such omissions can lead to semantic ambiguity. Without visual markers like question marks, exclamations, or even line breaks for pacing, participants were left to infer meaning and intent without sufficient context.

The researcher further documented several moments during the study where this lack of emotional signaling led to conversational breakdowns. In one instance, a speaking participant laughed aloud during a conversation, but because the chat interface offered no indication of the laugh—such as "[laughs]" or an emoji—the conversation partner was visibly confused. The absence of cues such as tone modulation, pauses, or expressive markers made it difficult for users to know whether a message was humorous, sarcastic, or serious. This problem was compounded when multiple messages appeared in rapid succession, leading participants to miss or misinterpret subtle shifts in mood or engagement.

Even typed messages, which theoretically allow for more deliberate expression, lacked emotion due to the minimalist interface design. The chat display did not support emoji, text formatting, or indicators of intent (e.g., italics, emphasis, or typing indicators), contributing to a monotone presentation. Messages arrived as flat lines of text, devoid of paralinguistic context. Although participants occasionally clarified emotion through their words (e.g., typing "haha" or using caps), these workarounds were inconsistent and cognitively demanding.

Collectively, these observations point to a critical design gap in the prototype: while it enables efficient information exchange, it does not currently support the expressive richness necessary for human connection. Emotional nuance plays a vital role in establishing trust, regulating turn-taking, and reinforcing shared understanding. Participants' comments and behaviors suggest that enhancing the system with expressive features—such as automatic punctuation, optional sentiment tagging (e.g., [laughs], [frustrated]), or lightweight emoji support—could greatly improve the perceived authenticity and effectiveness of communication.

In addition to challenges with emotional expression, participants shared a range of experiences regarding the flow of conversation. Many expressed that once both users were engaged and familiar with their input method, the conversation felt natural and fluid. Several participants commented that the system allowed them to follow the interaction clearly, with real-time updates and turn-taking that supported a sense of connection. For instance, some noted that the shared display made it easy to know when their partner had finished responding, contributing to a smooth back-and-forth rhythm. However, there were also moments where the flow became disrupted, particularly when switching between input modes or when the speech recognition lagged. In these cases, delays in message appearance or confusion about whether a partner was typing or speaking led to brief conversational stalls. Still, participants generally felt that the prototype allowed for a relatively smooth and responsive exchange, especially considering the novelty of the interaction

setting. These observations suggest that while the baseline conversation flow was strong, it could be further enhanced with more visible input indicators, feedback animations, or system cues to help regulate pacing and support real-time responsiveness.

4.5 Presentation of the Chatroom and Visual Presentation

The prototype was designed with a clean, minimalist interface using a black and white color scheme to ensure high contrast and readability. In the chatroom, typing participants' messages were displayed in green font, while speech-generated messages were shown in blue font, alongside the participant's name. This color-coded distinction aimed to help users quickly identify the origin and modality of each message in real time. While this approach was largely functional, participant feedback revealed several opportunities for improvement in both the visual layout and the mobile interface experience.

Participants emphasized the importance of familiarity and visual structure in enhancing readability and engagement during real-time conversations. For instance, P2 suggested that the chat display adopt a design more similar to iMessage or SMS interfaces, where messages are shown in bubble formats aligned to the left or right of the screen, depending on the sender. This format would not only improve message tracking but also mimic a familiar model that many users are accustomed to, reducing cognitive load and making it easier to follow turn-taking, particularly in fast-paced or multi-user conversations. The current center alignment, while symmetric, made it difficult to distinguish between conversation threads, especially when speech messages arrived in delayed bursts.

In terms of mobile usability, P4 expressed a preference for larger microphone buttons when using the speech interface, noting that a bigger target area would reduce the effort required to engage the feature and allow her to focus more on speaking rather than precision tapping. This highlights a broader need for touch-friendly design, especially in accessibility tools where the users may need to alternate between input modes quickly or operate the system without full visual attention.

Another recurring concern is the keyboard overlay issue during typing interactions. Participants P3 and P8 observed that while the chatbox was appropriately sized for reading when the screen was static, it became partially obstructed once the virtual keyboard was activated. This made it difficult to monitor both one's messages and the partner's real-time responses—an issue particularly problematic for users relying on the system for nuanced conversation or emotional expression. These findings suggest the need for responsive design elements that dynamically resize or reposition content when input fields are active, such as shifting the chat display upward or temporarily collapsing nonessential UI elements.

Additionally, Participant P3 noted that the input field on the login screen—used to type in participant names—was relatively small and could benefit from increased padding and font size. Although this was a minor complaint, it reflects the broader principle that first impressions and onboarding clarity matter, especially when the system is intended for impromptu or unsupervised use in public settings.

Overall, participants' feedback on visual presentation underscores the value of familiar design metaphors, modality distinction, and responsive interface behaviors. While the current version achieved functional clarity through color and layout, users desired more personalization, visual separation between speakers, and accessible controls that adapt fluidly to context. These insights point toward future interface iterations that balance simplicity with adaptive responsiveness and user-centered visual structure, making the tool more intuitive and usable in varied real-world environments.

4.6 Perceived Impact and Future Use

Participants expressed a strong interest in the real-world applicability of the kiosk-based conversation tool, frequently highlighting its flexibility, low barrier to entry, and potential to support spontaneous, inclusive interactions in a variety of settings. These reflections indicate that, beyond initial usability, the tool has meaningful potential to reshape communication experiences between d/DHH and hearing individuals in both private and public spaces.

One notable theme that emerged was the system's potential for multi-user and group communication. P11 remarked on the feasibility of using the system in conversations involving more than two people, emphasizing that each participant could use their mobile device and select their preferred input modality, either typing or speaking. This design supports individual autonomy, allowing users to participate in a collective conversation without being forced into a one-size-fits-all modality. P11 also noted that if multiple users choose to speak, the use of personal microphones could assist in differentiating voices, especially if future versions of the system incorporate speaker diarization or voice separation features.

Beyond conversational flexibility, participants and the researcher identified the tool's utility in specialized environments where traditional communication methods are either impractical or exclusionary. For instance, the researcher shared a compelling use case from a test proctoring setting, where maintaining a quiet atmosphere is essential. In such a context, whispering is often used to avoid disrupting others. However, students who are d/DHH may struggle to detect whispered speech due to its low volume and lack of articulation clarity. The tool allowed the proctor to communicate silently and precisely through typing, eliminating the need to write on paper or rely on potentially unclear gestures. This anecdote illustrates how the system can bridge situational access gaps without requiring systemic infrastructure changes or specialized training.

Participants also envisioned the tool being used in loud environments—such as concerts, symposiums, or busy public venues—where spoken communication may be difficult for anyone, not just d/DHH individuals. In these cases, the shared screen offers a centralized, visual representation of conversation, enhancing inclusivity for both d/DHH and hearing participants navigating noise-intensive spaces. Museums, which often encourage silent engagement, were also suggested as promising deployment sites, where typing inputs could quietly facilitate dialogue between visitors and staff.

The privacy-preserving nature of the system was another highly appreciated feature. Several participants noted that they felt more comfortable using the kiosk tool precisely because it did not require them to exchange phone numbers, social media handles, or other personal identifiers. This design choice was interpreted as reducing social pressure and increasing willingness to engage in short-term or spontaneous conversations with strangers, whether at a public kiosk, in a waiting area, or during collaborative activities like group assignments or networking events.

Importantly, two d/DHH participants explicitly requested a link to the prototype following the session, indicating their desire to use the system in their daily lives. This unsolicited feedback points to a high level of perceived usefulness and confirms that the tool filled a meaningful gap in their current communication strategies. Their enthusiasm further reinforces the notion that the system's low-friction design, QR code access, multimodal input, and centralized display have real potential for adoption beyond the research setting. Together, these insights suggest that the tool is not only usable but desirable in a range of contexts where communication access is typically limited or uneven. The design's ability to adapt to user preferences and environmental demands without requiring prior installation or shared language knowledge makes it a scalable and inclusive solution. Future deployments could explore integration with public services,

institutional spaces, and social platforms, ultimately broadening its reach and enhancing everyday accessibility for d/DHH users and their hearing peers.

5 Discussion

This study explored the usability, accessibility, and experiential impact of a kiosk-based, multi-modal communication prototype designed to support real-time dialogue between d/DHH and hearing individuals. While text messaging provides a familiar and accessible mode of communication, it is not optimized for spontaneous, in-person conversations between d/DHH and hearing individuals, especially in public or semi-public settings. Text messaging typically requires the exchange of personal contact information, introduces latency in back-and-forth interactions, and lacks shared context, which can disrupt the natural flow of communication. In contrast, this system is designed for co-located, real-time communication, eliminating the need for prior contact, and offers a shared display and multiple input modalities that better support accessibility, immediacy, and situational inclusiveness. Our findings demonstrate both the promise and the complexity of inclusive interaction design. Participants responded positively to the dual-modality system, especially the flexibility of choosing between typing and speech, but also revealed critical design tensions related to speech recognition fragility, lack of emotional expressiveness, and public-facing interface constraints. While the prototype successfully supported spontaneous interaction and imposed minimal physical or cognitive load, participants' feedback highlighted clear areas for refinement to ensure sustained usability, affective clarity, and privacy in real-world use.

5.1 The Trade-Off Between Accuracy and Privacy

Participants' experiences with the VOSK-powered speech recognition system reinforced a central challenge in the design of accessible ASR tools: the trade-off between local privacy and transcription accuracy. Cloud-based systems such as Google Speech-to-Text and Microsoft Azure typically offer high performance by leveraging deep learning models at scale, but they require a persistent internet connection and transmit audio data to external servers, raising serious concerns over data privacy and surveillance [1, 4]. In contrast, our offline implementation preserved data locality, allowing for real-time transcription without cloud dependencies. However, participants noted performance degradation in noisy environments, with increased misrecognition and delays.

These issues were not merely technical. As some participants pointed out, speaking aloud, especially in shared spaces, posed social and informational privacy risks, particularly when confidential or sensitive topics were discussed. Moreover, people with mobility impairments or visual disabilities may prefer voice interaction, but this demands highly accurate recognition at low volumes, a level not yet reliably achieved by many offline STT models. Thus, the broader implication is that future multi-modal systems must delicately balance privacy, accessibility, and reliability, allowing users to choose interaction modes based on both environmental context and personal comfort.

5.2 Typing as a Reliable Anchor in Multi-modal Interaction

Typing emerged as a consistently preferred input method, particularly in noisy environments or when privacy was a concern. Participants described typing as accurate, emotionally safer, and more familiar, especially among d/DHH users who valued the ability to control message timing, content, and pacing. These findings echo prior research emphasizing the importance of modality flexibility in inclusive system design [36].

Despite this strong preference, usability frictions were noted, particularly with keyboard overlap on mobile devices, which occasionally obstructed the chat view. These concerns highlight the need for adaptive user interface behaviors,

such as dynamic resizing or modal input windows. Additionally, participants appreciated being able to use their phones, indicating that bring-your-own-device (BYOD) compatibility significantly reduced learning time and increased willingness to engage.

The successful use of the tool in a test proctoring environment further demonstrates real-world value: d/DHH students were able to silently communicate with hearing staff, resolving accessibility issues inherent in whispered or written communication. This suggests a broader application for asynchronous or discreet conversations in environments that typically exclude d/Deaf users due to auditory reliance.

5.3 Expressiveness Is Central, Not Peripheral

One of the most consistent and impactful findings was the lack of affective expressiveness in the system, particularly for speech input. Participants reported that flat, unpunctuated text made it difficult to discern tone, humor, or emotional cues. As one user noted, even the absence of question marks changed the perceived intention of a message. These limitations affected conversational flow and interpersonal connection, which are crucial in cross-modal communication.

While it may be tempting to consider expressiveness as secondary to basic functionality, HCI research argues otherwise. Emotional cues, delivered through prosody, punctuation, and nonverbal signals, are essential to social cohesion, trust, and conversational nuance [26]. Our findings confirm that without tone indicators or expressive feedback (e.g., emojis, [laughs], or sentiment tags), even high-accuracy systems may feel cold or ambiguous.

Future iterations of this tool should prioritize effective augmentation of both speech and typed inputs. Simple additions like punctuation prediction, visual tone tags, or live emotion indicators could dramatically enhance interpersonal understanding, especially between users from different communication cultures.

5.4 Shared Screens and Social Context

The use of a shared monitor to display the ongoing conversation was met with mixed reactions. On one hand, participants appreciated the mutual visibility and synchronization it enabled, especially when conversation partners were sitting side by side. On the other hand, some voiced concern about having their messages publicly visible, particularly in group or sensitive contexts. This raises important questions about visibility, discretion, and agency in shared interface environments.

Participants suggested introducing options for display toggling or privacy modes, where users could temporarily hide the shared screen or direct specific messages to individual devices. These features would enhance control and encourage use in public kiosks, clinics, classrooms, and service desks, where users may vary in their comfort with public display. This aligns with broader HCI findings that public-facing technologies require context-aware interface behavior, adaptable to both physical and social environments.

5.5 Low Frustration and Workload Support Real-World Use

Usability metrics from NASA TLX reveal that the prototype imposed low to moderate levels of workload, with a particularly low score for mental demand (M = 3.8) and frustration (M = 2.7). Importantly, all participants were able to access the system, log in, and begin a conversation without assistance. This indicates that the tool supports spontaneous and informal use, even among first-time users with varying technical backgrounds.

Participants also noted that QR code onboarding was fast and non-intrusive, and they appreciated that no contact information needed to be shared to begin communicating. These features make the tool especially suitable for interpersonal interactions among strangers, such as those that occur in public or institutional settings where trust has not

yet been established. The system's minimal cognitive overhead and strong sense of user control were central to its perceived accessibility.

6 Conclusion

This study introduced and evaluated a kiosk-based multi-modal conversation tool designed to support inclusive communication between d/DHH and hearing individuals. By combining real-time speech recognition and typing input displayed on a shared monitor, the system aimed to offer accessible, on-demand communication in public and semi-public environments. Through a mixed-methods evaluation with 12 participants, including both d/DHH and hearing users, we identified key strengths and challenges of the conversation tool. The tool was rated positively for usability, with participants highlighting its intuitive design, accessibility, and minimal setup burden. The typing modality was consistently rated as more accurate and expressive, especially in noisy or socially sensitive environments. In contrast, while speech recognition was praised for its speed, participants reported variability in accuracy and tone recognition, particularly under challenging acoustic conditions. Despite these limitations, participants saw strong potential in the tool's real-world applicability, particularly in scenarios such as test proctoring, museum visits, and spontaneous interactions where interpreters or captionists may not be available. The ability to communicate anonymously without exchanging contact information also emerged as a valued feature, reinforcing the importance of privacy and ease in inclusive technology design.

6.1 Limitation & Future Works

Future work will focus on enhancing the system's emotional expressiveness, interaction flexibility, and adaptability for broader use cases. One of the most pressing areas for improvement is the lack of effective cues in both speech and typing outputs. Integrating features such as automatic punctuation, sentiment tagging (e.g., [laughs], [question]), or emoji support could help convey tone and emotion more effectively. In addition, providing a variety of emojis as an option allows users to personalize the message. Besides, Participants expressed difficulty when switching between input methods during a conversation. This suggests the need for smoother transitions, such as gesture-based toggles, persistent mode indicators, or real-time system prompts to guide users through input changes without confusion.

Improvements in visual design are also necessary. Participants requested interface refinements that resemble familiar messaging platforms like iMessage, including bubble-based message alignment and clearer distinctions between speaker turns. These design choices would not only increase usability but also enhance readability, especially in fast-paced or group conversations. Speaking of groups, another important direction is to expand the system for multi-user interactions. Allowing several users to engage in a shared conversation, each with their preferred input mode, would extend the applicability of the tool in social and public settings.

Lastly, future studies should include longer-term, ecologically valid deployments across various environments such as classrooms, museums, test proctoring rooms, or service desks. These studies will help examine how the tool integrates into daily communication practices, sustains user satisfaction, and supports ongoing accessibility needs. Collectively, these future developments aim to further close the communication gap between d/DHH and hearing individuals while maintaining flexibility, privacy, and ease of use.

References

[1] Shimaa Ahmed, Abhishek Roy Chowdhury, Amrita, Kassem Fawaz, and Prameshr Ramanathan. 2019. Preech: A System for Privacy-Preserving Speech Transcription. *arXiv preprint arXiv:1909.04198* (2019). https://arxiv.org/abs/1909.04198

[2] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. *Qualitative HCI research: Going behind the scenes.* Morgan & Claypool Publishers.

[3] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa

[4] Tom Bäckström. 2023. Privacy in Speech Technology. *arXiv preprint arXiv:2305.05227* (2023). https://arxiv.org/abs/2305.05227

[5] K. Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI' 16)*. ACM, New York, NY, USA, 981–992. doi:10.1145/2858036.2858498

[6] National Deaf Center. 2022. The ASL Interpreter Shortage and Its Impact on Accessibility in College Settings. https://nationaldeafcenter.org/news-items/the-asl-interpreter-shortage-and-its-impact-on-accessibility-in-college-settings/ Retrieved November 16, 2024.

[7] Shi Chen, Xiaodong Wang, Weijun Li, Jingao Zhang, Yuge Qi, Jiaqi Teng, and Zhihan Zeng. 2024. Silent Delivery: Practices and Challenges of Delivering Among Deaf or Hard of Hearing Couriers. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 386, 17 pages. doi:10.1145/3613904.3642801

[8] Victoria Clarke and Virginia Braun. 2017. Thematic analysis. *The journal of positive psychology* 12, 3 (2017), 297–298. doi:10.1080/17439760.2016.1262613

[9] Dennis Cokely. 1986. The effects of lag time on interpreter errors. *Sign Language Studies* 53, 1 (1986), 341–375.

[10] Aashaka Desai, Jennifer Mankoff, and Richard E. Ladner. 2023. Understanding and Enhancing The Role of Speechreading in Online d/DHH Communication Accessibility. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 608, 17 pages. doi:10.1145/3544548.3580810

[11] Elias Dritsas, Maria Trigka, Christos Troussas, and Phivos Mylonas. 2025. Multimodal interaction, interfaces, and communication: a survey. *Multimodal Technologies and Interaction* 9, 1 (2025), 6.

[12] Shahram Ghorbani and John H.L. Hansen. 2022. Domain Expansion for End-to-End Speech Recognition: Applications for Accent/Dialect Speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2022), 762–774. doi:10.1109/TASLP.2022.3233238

[13] Google Cloud. 2025. Speech-to-Text Documentation. https://cloud.google.com/speech-to-text. Accessed: 2025-07-14.

[14] Elizabeth Halcomb and Louise Hickman. 2015. Mixed methods research. *Nursing Standard (2014+)* 29, 32 (Apr 08 2015), 41. https://ezproxy.rit.edu/login?url=https://www.proquest.com/scholarly-journals/mixed-methods-research/docview/1785222694/se-2 Copyright - Copyright: 2012 (c)2012 RCN Publishing Company Ltd. All rights reserved. Not to be copied, transmitted or recorded in any way, in whole or part, without prior permission of the publishers; Last updated - 2023-11-30; CODEN - NSTAEU.

[15] Muslimah Nathifa Harris. 2012. *How do deaf and hard of hearing freshmen students persist until they graduate.* Ph. D. Dissertation. California State University, Sacramento.

[16] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage Publications, Sage CA: Los Angeles, CA, 904–908.

[17] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183. doi:10.1016/S0166-4115(08)62386-9

[18] Scott Haynes. 2014. Effectiveness of Communication Strategies for Deaf or Hard of Hearing Workers in Group Settings. *Work* 48, 2 (2014), 193–202. doi:10.3233/WOR-131612

[19] Celia Hulme, Alys Young, Katherine Rogers, and Kevin J Munro. 2023. Cultural competence in NHS hearing aid clinics: a mixed-methods case study of services for Deaf British sign language users in the UK. *BMC Health Services Research* 23, 1 (2023), 1440.

[20] K. Kuhn, V. Kersken, B. Reuter, N. Egger, and G. Zimmermann. 2024. Measuring the Accuracy of Automatic Speech Recognition Solutions. *ACM Transactions on Accessible Computing (TACCESS)* 16, 4 (2024), Article 25, 23 pages. doi:10.1145/3636513

[21] R. S. Kushalnagar, G. W. Behm, A. W. Kelstone, and S. Ali. 2015. Tracked Speech-To-Text Display: Enhancing Accessibility and Readability of Real-Time Speech-To-Text. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS' 15)*. ACM, 223–230. doi:10.1145/2700648.2809843

[22] Language Services Associates. 2023. Why is There a Shortage of Certified American Sign Language Interpreters? https://lsa.inc/why-is-there-a-shortage-of-certified-american-sign-language-interpreters/ Accessed: 2025-07-17.

[23] Gi-Bbeum Lee, Hyuckjin Jang, Hyundeok Jeong, and Woontack Woo. 2021. Designing a Multi-Modal Communication System for the Deaf and Hard-of-Hearing Users. In *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. 429–434. doi:10.1109/ISMAR-Adjunct54149.2021.00097

[24] Tiantian Liu, Chao Wang, Zhengxiong Li, Ming-Chun Huang, Wenyao Xu, and Feng Lin. 2024. Wavoice: An mmWave-Assisted Noise-Resistant Speech Recognition System. *ACM Transactions on Sensor Networks* 20, 4 (2024), Article 86, 29 pages. doi:10.1145/3597457

[25] Yi Liu, Yuekang Li, Gelei Deng, Felix Juefei-Xu, Yao Du, Cen Zhang, Chengwei Liu, Yeting Li, Lei Ma, and Yang Liu. 2024. Aster: Automatic Speech Recognition System Accessibility Testing for Stutterers. In *Proceedings of the 38th IEEE/ACM International Conference on Automated Software Engineering (ASE '23)*. IEEE Press, Echternach, Luxembourg, 510–521. doi:10.1109/ASE56229.2023.00107

[26] Sandra Luo. 2024. Navigating the Diverse Challenges of Speech Emotion Recognition: A Deep Learning Perspective. In *Proceedings of the 27th International Academic Mindtrek Conference (Mindtrek '24)*. Association for Computing Machinery, Tampere, Finland, 133–146. doi:10.1145/3681716.3681725

[27] Microsoft Azure. 2025. Azure AI Speech Services. https://azure.microsoft.com/en-us/products/ai-services/ai-speech. Accessed: 2025-07-14.

[28] National Institute on Deafness and Other Communication Disorders. 2019. Assistive Devices for People with Hearing, Voice, Speech, or Language Disorders. https://www.nidcd.nih.gov/health/assistive-devices-people-hearing-voice-speech-or-language-disorders Accessed: 2025-07-17.

[29] National Institute on Deafness and Other Communication Disorders. 2023. Quick Statistics About Hearing. https://www.nidcd.nih.gov/health/statistics/quick-statistics-hearing. Accessed: 2025-07-14.

[30] WXXI News. 2022. Resources and culture make Rochester a draw for many in the Deaf community. https://www.wxxinews.org/inclusion-desk/2022-01-24/resources-culture-make-rochester-a-draw-for-many-in-deaf-community Inclusion Desk, WXXI News, January 24, 2022. Accessed: November 16, 2024.

[31] M. Nikolaraizi and V. Argyropoulos. 2015. The Learning and Communication Barriers of Deaf and Hard of Hearing Students in Higher Education. In *7th International Conference on Education and New Learning Technologies (EDULEARN15 Proceedings)*. IATED, Barcelona, Spain, 4130–4134. https://library.iated.org/view/NIKOLARAIZI2015LEA Accessed: November 16, 2024.

[32] Rochester Institute of Technology. 2024. TigerChat: Real-Time Speech-to-Text Communication Tool. https://www.rit.edu/accesstechnology/tigerchat Retrieved November 29, 2024.

[33] Douglas O'Shaughnessy. 2008. Automatic speech recognition: History, methods and challenges. *Pattern Recognition* 41, 10 (2008), 2965–2979.

[34] Render. 2024. *Render: The Modern Cloud for Application Developers*. https://render.com/ Accessed: July 19, 2025.

[35] Rhode Island Commission on the Deaf and Hard of Hearing. 1992. American Sign Language (ASL). https://cdhh.ri.gov/information-referral/american-sign-language.php Accessed: 2025-07-17.

[36] Mandlenkosi Shezi and Abejide Ade-Ibijola. 2020. Deaf chat: A speech-to-text communication aid for hearing deficiency. *Advances in Science, Technology and Engineering Systems Journal* 5, 5 (2020), 826–833.

[37] Robert Soden, Austin Toombs, and Michaelanne Thomas. 2024. Evaluating interpretive research in HCI. *Interactions* 31, 1 (2024), 38–42. doi:10.1145/3633200

[38] VOSK. [n. d.]. VOSK Speech Recognition Models. https://alphacephei.com/vosk/models. Accessed: 2025-06-24.

A Participant Demographics

Table 1. Participant Demographics

Pair ID	PID	DHH	Primarly Language	Gender	Input Methods	Topics
1	P1	Yes	American Sign Language	Female	Typing	internship, majors, summer plan
1	P2	No	English	Male	Speaking	internship, Majors, summer plan
2	P3	Yes	English	Female	Typing	capstone projects, research area
2	P4	No	English	Female	Speaking	capstone projects, research area
3	P5	Yes	Others	Male	Typing	robots, past experience, majors
3	P6	No	English	Female	Speaking	robots, past experience, majors
4	P7	Yes	English	Female	Typing	summer plan
4	P8	No	English	Male	Typing	summer plan
5	P9	Yes	American Sign Language	Female	Typing	about jobs, pets, post graduation plan
5	P10	No	English	Female	Speaking	about jobs, pets, post graduation plan
6	P11	Yes	English	Female	Typing	double major, about Rochester, internship
6	P12	No	English	Female	Speaking	double major, about Rochester, internship