

EvolveCaptions: Empowering DHH Users Through Real-Time Collaborative Captioning

Liang-Yuan Wu
University of Michigan
Ann Arbor, MI, USA
lyuanwu@umich.edu

Dhruv Jain
University of Michigan
Ann Arbor, MI, USA
profdj@umich.edu

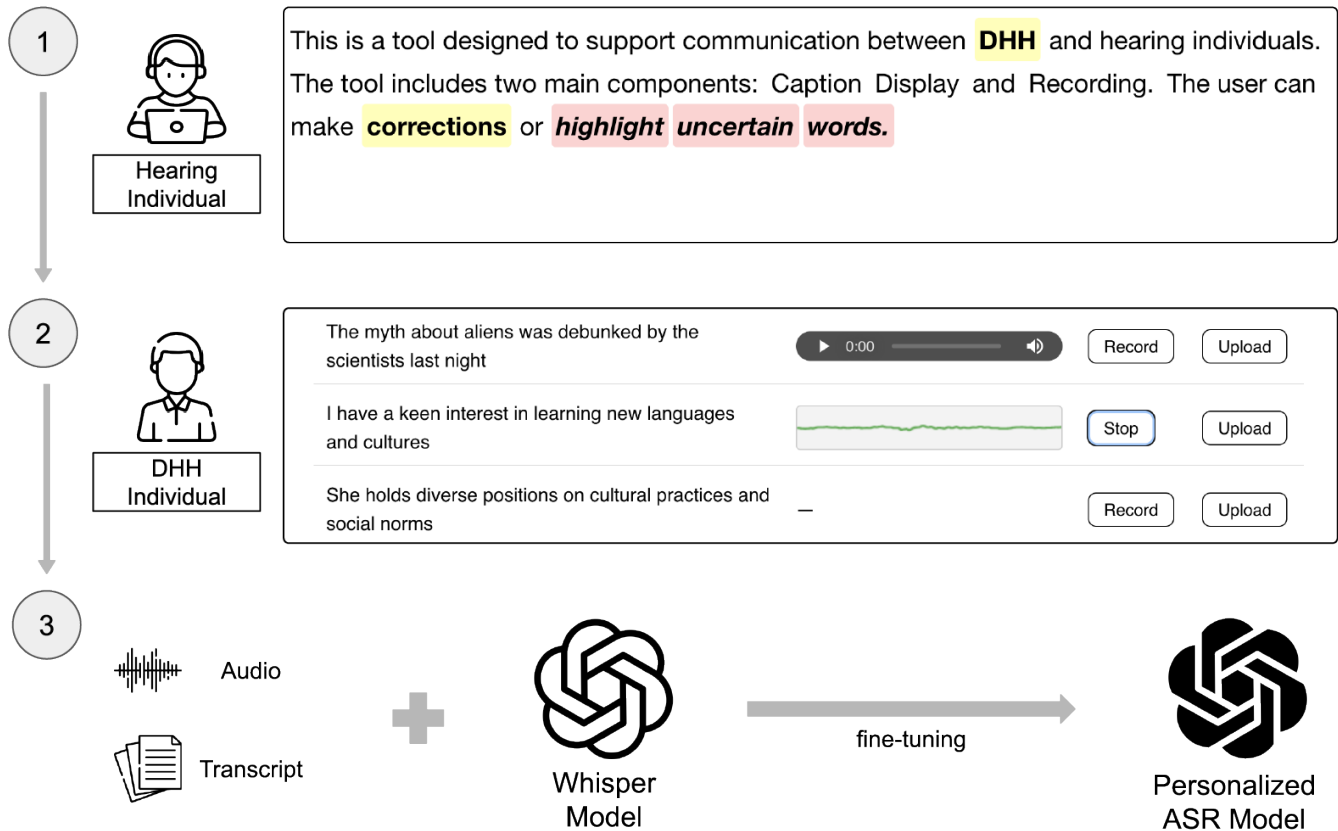


Figure 1: *EvolveCaptions* is an interactive ASR adaptation system that improves accessibility in mixed-ability conversations. It combines live caption correction with lightweight, speaker-specific fine-tuning: (1) Hearing users correct live captions of the DHH speaker’s voice; (2) the DHH speaker records targeted clauses generated from the corrected terms; (3) the Whisper ASR model is fine-tuned with these recordings and adapts to the speaker over time.

Abstract

Automatic Speech Recognition (ASR) systems often fail to accurately transcribe speech from Deaf and Hard of Hearing (DHH) individuals, especially during real-time conversations. Existing personalization approaches typically require extensive pre-recorded data and place the burden of adaptation on the DHH speaker. We present EvolveCaptions, a real-time, collaborative ASR adaptation system that supports in-situ personalization with minimal effort. Hearing participants correct ASR errors during live conversations. Based on these corrections, the system generates short, phonetically targeted prompts for the DHH speaker to record, which are

then used to fine-tune the ASR model. In a study with 12 DHH and six hearing participants, EvolveCaptions reduced Word Error Rate (WER) across all DHH users within one hour of use, using only five minutes of recording time on average. Participants described the system as intuitive, low-effort, and well-integrated into communication. These findings demonstrate the promise of collaborative, real-time ASR adaptation for more equitable communication.

CCS Concepts

• **Human-centered computing** → **Accessibility systems and tools.**

Keywords

Accessibility, Deaf and Hard of Hearing, Automatic Speech Recognition.

1 Introduction

Automatic Speech Recognition (ASR) has become a widely used technology for enhancing communication. Yet, the diversity of speaking styles presents a persistent challenge to developing universally accurate ASR systems capable of understanding all users [4, 9, 54]. While state-of-the-art models demonstrate impressive performance on benchmark datasets [18], they continue to underperform on speech produced by Deaf and Hard of Hearing (DHH) individuals [16, 67]. This disparity contributes to inequities in communication access, particularly for people with atypical or disordered speech [15, 19]. DHH speech often varies in pronunciation, both across and within individuals, sometimes posing comprehension challenges even for familiar listeners [5, 41].

Prior efforts to improve ASR for DHH users, such as Project Euphonia [40] and the work of Tobin et al. [63], have focused on personalization through large, pre-recorded datasets collected in controlled conditions. However, these approaches face three key limitations. First, they impose a substantial motivational burden: spending hours recording scripted speech is often impractical for users. Second, they lack contextual relevance: samples collected offline may not generalize to spontaneous, real-world conversation. Third, they frame adaptation as an isolated, pre-emptive task, placing the full responsibility on the DHH individual without opportunities for real-time, collaborative refinement.

In this paper, we introduce *EvolveCaptions*, an interactive ASR adaptation system that supports real-time, collaborative personalization during live conversations. In mixed-ability settings, when a DHH individual speaks, the system transcribes their speech in real time. Hearing participants read the transcript and correct any errors. These marked segments are passed to a language model (GPT-4), which generates phonetically plausible phrases containing the corrected words. The DHH speaker records clauses that include these phrases, which are then used to fine-tune the ASR model. This loop repeats iteratively, allowing the model to gradually adapt to the speaker’s voice with minimal effort.

By targeting only misrecognized segments, *EvolveCaptions* minimizes the recording burden while maximizing the utility of collected data. Unlike static pre-recorded methods, our approach grounds ASR adaptation in authentic interaction contexts. Importantly, it redistributes the labor of accessibility by involving hearing participants in correcting errors, aligning with the principle of collective access, which emphasizes shared responsibility for accessible communication environments [42, 43, 60].

We evaluated *EvolveCaptions* in a lab study with 12 DHH participants and six hearing partners. Across five progressive captioning sessions, DHH participants read randomized scripts while hearing partners corrected captioning errors in real time. These corrections generated targeted recording prompts, enabling lightweight, speaker-specific model adaptation between sessions.

Comparing the first and last sessions, *EvolveCaptions* reduced word error rate (WER) by a median of 27.2% (mean = 30.4%), a statistically significant improvement ($p < .05$, Wilcoxon signed-rank

test). To contextualize these gains, we benchmarked final models against three baselines: a static Whisper-base model, a generically fine-tuned dysarthric model, and one-time human correction without adaptation. *EvolveCaptions* consistently outperformed all three—demonstrating that real-time, speaker-specific adaptation offers clear advantages over generic fine-tuning or one-time human correction alone. Qualitative feedback echoed these results: DHH participants found the system intuitive and low-effort, while hearing partners reported that real-time correction became easier with practice—highlighting the learnability of the workflow.

In summary, our work contributes: (1) *EvolveCaptions*, an interactive captioning system that enables hearing participants to collaboratively correct ASR errors and supports lightweight, real-time adaptation to DHH speakers’ voices, (2) an empirical evaluation with 12 DHH and six hearing participants, demonstrating both quantitative improvement in recognition and qualitative insights into collaborative captioning experiences, and (3) design implications for future captioning technologies, illustrating how ASR systems can shift from requiring DHH users to adapt to the technology toward systems that adapt to users, promoting more equitable communication. We open-source the project data and artifacts through our GitHub repository: <https://github.com/binomial14/EvolveCaptions>.

2 Related Work

We examine DHH captioning needs and ASR’s role in mixed-ability communication, then situate our work within three areas: (1) ASR-based captioning for DHH users, (2) personalized ASR for atypical speech, and (3) collaborative captioning systems that share accessibility labor.

2.1 Captioning Needs of DHH People

The DHH community encompasses a diverse population with varying degrees of hearing loss, communication preferences, and cultural identities [8]. Some individuals identify as culturally Deaf, using American Sign Language (ASL) as their primary language and participating in a shared cultural-linguistic community [8, 36, 46]. Others may identify as hard of hearing or audiotically deaf, often using spoken language and relying on hearing technologies such as hearing aids or cochlear implants [8, 36]. Communication access needs across this spectrum are highly individualized, and no single approach suffices for all users.

Captions serve as a crucial accessibility tool for many DHH individuals, enabling access to spoken content across educational, professional, and social domains [26, 27, 33, 35, 38, 45]. Prior work has examined how DHH users adopt captioning technologies in classrooms [30, 34, 45], medical settings [25, 26], workplaces [14, 27], and daily communication contexts [42, 50]. These studies emphasize that captions not only improve comprehension, but also support participation, independence, and equitable engagement.

Several modalities exist for caption provision. Human-generated captions, such as those provided by Communication Access Realtime Translation (CART) professionals, offer high accuracy but are expensive and not easily scalable [13]. Crowd-sourced and peer-generated captions have been explored as lower-cost alternatives with mixed results [24, 37]. More recently, Automatic Speech Recognition (ASR) has become a popular method for generating

real-time captions due to its accessibility, scalability, and low latency [29, 60, 64]. However, ASR captions often fall short in accuracy when applied to DHH speech, leading to breakdowns in communication and increased cognitive burden on users [29, 67].

2.2 ASR for DHH–Hearing Communication

ASR is increasingly used to facilitate DHH–hearing communication in classrooms, meetings, and daily conversations [10, 15, 60, 66]. Tools like Google Live Transcribe, Otter.ai, and Zoom’s live captions, as well as open-source models like Whisper [51], offer low-cost, real-time transcription with broad accessibility.

Despite these advances, ASR systems consistently underperform on speech from DHH users. Most commercial ASR models are trained on fluent, non-disabled speech [55, 61, 63], resulting in poor performance when processing speech that diverges in rhythm, articulation, and prosody. These atypical speech characteristics—common among DHH individuals—have been shown to substantially degrade ASR accuracy [44, 49, 67]. Additionally, misalignments between how DHH speakers monitor or produce speech and how ASR systems interpret it may further reduce reliability [12, 41, 46, 53].

These transcription failures can reduce intelligibility and erode user trust, especially in mixed-ability conversations where mutual understanding is critical [4, 22, 29]. Common ASR issues such as high Word Error Rates (WER), hallucinated phrases, and missing function words can significantly distort meaning, further increasing the cognitive burden on DHH users. While Whisper and similar models perform well on standard benchmarks [51], their robustness does not consistently extend to speakers with atypical speech patterns [51, 63]. These gaps underscore the need for adaptive ASR systems that can provide equitable communication access across diverse speakers.

2.3 Personalized ASR for Atypical Speech

Personalized ASR [3] seeks to improve accuracy by adapting models to individual speakers—an approach shown to be effective for people with disordered or atypical speech [28, 39, 55, 61, 63]. In DHH contexts, where speech can vary widely in rhythm, clarity, and articulation, personalization has the potential to significantly improve transcription quality [67].

Projects like Google’s Project Euphonia [40] and Project Relate [21] have demonstrated that models fine-tuned on even a few minutes of speaker-specific data can dramatically reduce WER. Tobin et al. [63] found that most participants achieved usable accuracy levels with only 3–4 minutes of training data. Additionally, studies have explored real-time personalization in mobile ASR tools, such as Live Transcribe [38], and low-resource fine-tuning strategies for Whisper on dysarthric speech [48].

However, these systems typically require scripted, offline data collection and place the full burden of personalization on the end-user. They also fail to adapt over time during real interactions, making them less responsive to contextual variation. These limitations motivate more lightweight and dynamic personalization methods that can be integrated into everyday communication scenarios.

2.4 Collaborative Captioning

While much of the accessibility literature focuses on individual accommodations, recent HCI research has emphasized *collaborative accessibility*—designing systems where access is a shared responsibility [42, 43, 60]. This perspective, grounded in the principle of *collective access*, seeks to distribute accessibility labor across all participants in a conversation, especially in mixed-ability settings.

Several systems have explored how hearing users can assist in real-time caption correction. For example, McDonnell et al. [42] examined how hearing participants support caption repair in Zoom meetings. Others have enabled crowd-sourced or partner-based corrections of ASR transcripts to improve live communication [16, 24, 32].

These collaborative strategies can improve immediate comprehension and reduce reliance on flawed ASR output. However, most systems treat corrections as ephemeral—valuable in the moment but not retained by the system. As a result, DHH speakers must endure repeated corrections across sessions, and the ASR model does not improve over time.

Our work extends collaborative accessibility into the realm of *persistent ASR adaptation*. By turning real-time corrections into training data, we combine the strengths of collaboration and personalization to reduce long-term effort for both DHH and hearing users.

3 EvolveCaptions: An Interactive ASR Adaptation System

EvolveCaptions is an interactive ASR adaptation system designed to support mixed-ability conversations by enabling real-time caption correction and lightweight, speaker-specific model fine-tuning. The system allows hearing participants to collaboratively correct captions while automatically generating targeted prompts for DHH users to record. These brief recordings are then used to update the ASR model, enabling incremental, in-situ personalization with minimal burden on the speaker. Below, we describe the design motivations, interaction workflow, and technical implementation of the system.

3.1 Design Motivations

EvolveCaptions is grounded in three design goals, informed by prior work in ASR personalization, accessibility, and collaborative interaction:

- (1) **Low-effort personalization:** DHH speakers only record short, targeted clauses for words that were previously misrecognized, drastically reducing the time and effort required for ASR fine-tuning [61, 63].
- (2) **In-situ adaptation:** Rather than relying on large, pre-scripted datasets, our system collects training data during natural conversation, increasing contextual relevance and promoting sustained use [67].
- (3) **Collaborative correction:** Hearing participants actively assist by correcting captions in real-time, aligning with the principle of collective access and distributing accessibility labor [42, 60].

3.2 Three-Stage Interaction Loop

To meet the above design goals, EvolveCaptions follows a three-stage interactive loop (Figure 1):

3.2.1 Live Caption Correction. The system transcribes the DHH speaker’s voice using a Whisper-based ASR engine and displays real-time captions to all participants. Hearing users can collaboratively correct transcription errors by selecting and editing individual words or short phrases (Figure 1.1). Drawing on prior crowd-correction interfaces [16, 24], our interface allows users to both mark corrections (highlighted in yellow) and flag uncertain segments (in red). This dual-marking mechanism enables more consistent engagement even when users are unsure. All changes are broadcast instantly to both DHH and hearing participants, ensuring shared awareness and enhanced caption quality in the moment.

3.2.2 Clause Generation and Recording. Once a caption has been corrected, the system uses the revised word(s) to generate a natural-sounding clause for the DHH speaker to record. Rather than prompting users to repeat isolated words, EvolveCaptions embeds the corrected term into short, contextually appropriate phrases (Figure 1.2). This ensures training data is more representative of real-world utterances.

To achieve this, we prompt OpenAI GPT-4 with the following instruction:

```
You are generating short, spoken English clauses to
help improve an automatic speech recognition (ASR)
system. Based on a word that was misrecognized by
ASR, your goal is to create a new clause (5-15
words) that:
```

```
--- Sounds natural in a daily conversation
--- Contains the corrected word in a prominent,
clear context
--- Has a similar phonetic structure to the
original sentence
```

```
Original words: "{original}"
Corrected words: "{corrected}"
```

```
Generate one new clause that can be used to help
the ASR model learn this correction. Just reply
with the clause (no quotes, no explanation).
```

For example, if the ASR originally transcribed “fok” and it was corrected to “fork”, the system might generate: “*She picked up the fork from the table.*” These clauses balance phonetic similarity and natural speech structure.

DHH users are then prompted to record these generated clauses via a user-friendly interface. Each clause includes a waveform display, visual feedback, and playback controls. This supports users who rely on visual feedback and allows them to rerecord, skip, or delete prompts at will. By grounding recordings in natural clauses and offering flexible participation, the system reduces fatigue and ensures the collection of high-quality training data.

3.2.3 ASR Fine-Tuning. Corrected clause recordings are formatted into audio-text pairs using the HuggingFace dataset structure. Each

sample is padded and collated for batch training, and the ASR model is fine-tuned using Seq2SeqTrainer with lightweight hyperparameters: learning rate of 1e-5, batch size of 8, and maximum 100 steps (Figure 1.3).

We use Whisper-base (74M parameters, <1 GB memory) to balance performance and efficiency. The model is updated in the background and seamlessly replaces the previous model for subsequent captioning. During inference, the system receives 16kHz, 16-bit PCM audio via WebSocket, performs chunked transcription, and streams captions with low latency—comparable to commercial services like Google Live Transcribe [38].

3.3 User Interface

The EvolveCaptions user interface is visualized in Figure 2. It includes controls to initiate ASR and recording modes, displays live captions of the DHH speaker’s speech, and enables hearing participants to make corrections through an intuitive highlighting and editing system. During the recording phase, the interface presents the DHH user with targeted phrase prompts derived from previous corrections, along with live waveform feedback to guide clear recordings. This streamlined layout enables seamless transitions between speaking, correcting, and training, minimizing disruption to the conversation flow.

3.4 System Implementation

EvolveCaptions is implemented as a cross-platform web application. The frontend, built with ReactJS [62] and Vite [11], handles real-time caption display, collaborative correction, and DHH audio recording. Audio input is captured using Audio Worklet [47] and streamed to the backend over WebSocket.

The backend, written in FastAPI, manages transcription, correction tracking, clause generation, and model fine-tuning. We extend the open-source WhisperLive¹ project for low-latency inference. When a set of recordings is complete, the backend prepares the training data, launches fine-tuning, and deploys the updated model automatically. All data communication is secured via HTTPS with CORS enabled.

For our user study, the system was hosted on a Google Cloud VM with an NVIDIA T4 GPU. However, EvolveCaptions is lightweight enough to run on consumer-grade hardware; in local tests on a MacBook Pro (Intel i7), the system remained fully functional with modest latency, demonstrating portability across environments.

To support reproducibility and future research, the full implementation is open-sourced: <https://github.com/binomial14/EvolveCaptions>.

4 Evaluation

To evaluate the usability, effectiveness, and experiential impact of EvolveCaptions, we conducted remote, 90-minute user studies with 12 DHH (Deaf or Hard of Hearing) and six hearing participants. The study was designed to assess three key research questions:

- (1) Can EvolveCaptions reduce ASR errors over time with minimal user effort?
- (2) How do DHH and hearing participants experience the system during real-time, mixed-ability conversation?

¹<https://github.com/collabora/WhisperLive>

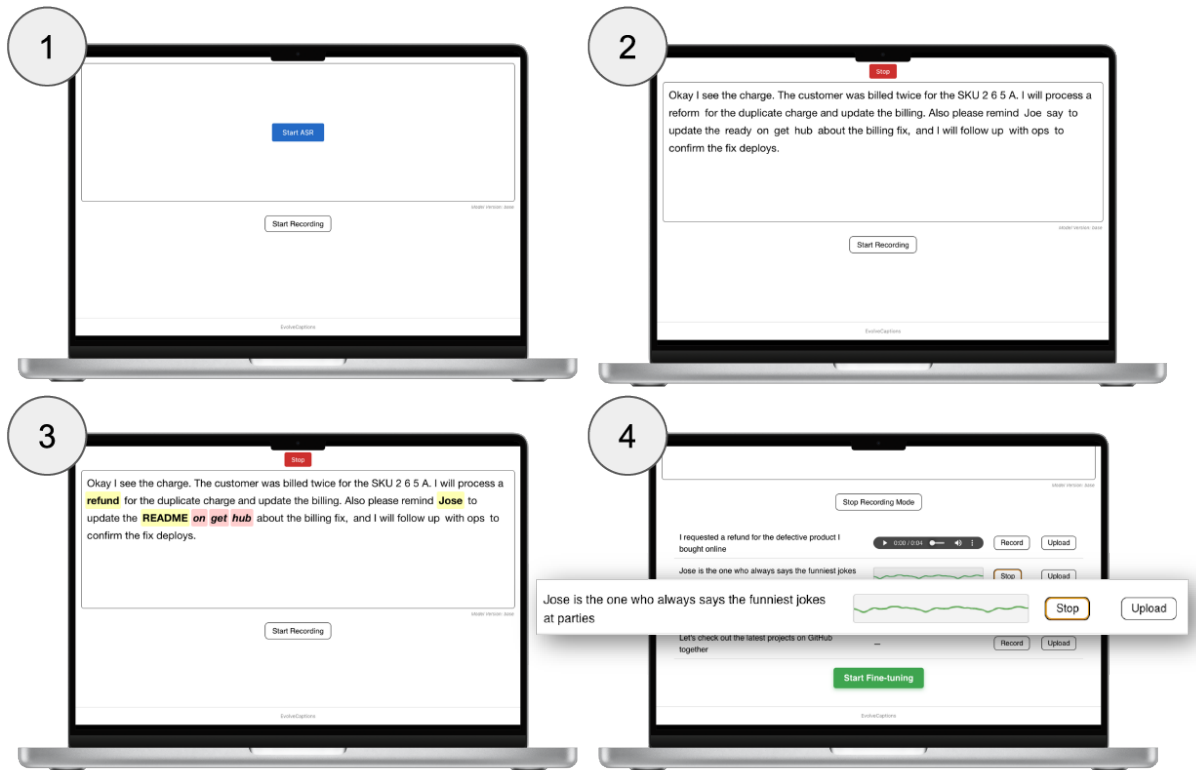


Figure 2: EvolveCaptions user interface. (1) Clicking “Start ASR” begins generating real-time captions, while “Start Recording” displays a list of targeted clauses for the DHH speaker to record; (2) real-time captions reflect the DHH speaker’s speech; (3) hearing users can refine captions by correcting errors (yellow highlights) or flagging uncertain words (red highlights); (4) during recordings, the interface shows targeted samples along with live waveforms for guidance.

- (3) What are the broader implications of collaborative ASR adaptation for accessible communication?

4.1 Participants

We recruited 12 DHH participants (5 women, 7 men; see Table 1) via email lists, social media, and snowball sampling. The average age was 37.6 years ($SD = 18.8$, $median = 30$, $range = 25 - 85$). Ten identified as hard of hearing (HoH), one as deaf, and one as Deaf. Eleven participants used hearing devices—nine used hearing aids and three used cochlear implants (some used both). Nine reported regular use of captioning tools (e.g., Apple CC [2], Google Transcribe [20], Zoom Captioning, CART [13]) to support verbal communication.

To serve as collaborative caption correctors, we also recruited six hearing participants (3 women, 3 men, Table 2) via lab mailing lists and social media. Their average age was 25.5 years ($SD = 2.9$, $median = 25.5$, $range = 21 - 29$). All were proficient in spoken English, including two native speakers, one bilingual speaker, and three fluent second-language speakers.

Each study session paired one DHH speaker with one hearing participant. Hearing participants were permitted to participate in multiple sessions with different DHH partners.

Table 1: DHH Participant demographics information for the user study.

ID	Age	Gender	Identity	Hearing loss	Onset age
P1	38	Female	HoH	Severe	18 month
P2	25	Male	Deaf	Profound	5 years
P3	68	Male	HoH	Severe	Birth
P4	30	Male	HoH	Moderately Severe	25 years
P5	85	Male	HoH	Profound	Birth
P6	32	Female	HoH	Severe	Birth
P7	29	Female	HoH	Mild	21 years
P8	30	Male	HoH	Moderately Severe	Birth
P9	28	Male	HoH	Moderate	Birth
P10	25	Male	HoH	Moderate	Birth
P11	31	Female	deaf	Profound	1 year
P12	30	Female	HoH	Moderate	Birth

4.2 Procedure

All sessions were conducted remotely over Zoom and approved by our Institutional Review Board (IRB). Each 90-minute session consisted of four phases: (1) pre-study survey, (2) system tutorial,

Table 2: Hearing Participant demographics information for the user study.

ID	Age	Gender	English Proficiency
H1	21	Female	Native Speaker
H2	25	Female	Bilingual
H3	24	Male	Native Speaker
H4	29	Male	Proficient
H5	28	Female	Proficient
H6	26	Male	Proficient

(3) interactive trial using EvolveCaptions, and (4) a semi-structured interview. Participants were provided the choice of their preferred accommodation—all participants opted for Zoom captioning support.

4.2.1 Setup and Pre-Survey. Participants began by completing a demographic and background questionnaire, including questions about prior experiences with ASR and captioning technologies. We also conducted a brief interview with DHH participants about their perceived performance of captioning tools and any frustrations or unmet needs.

4.2.2 Tutorial and System Orientation. We introduced participants to the concept of real-time collaborative ASR adaptation and demonstrated the EvolveCaptions workflow. DHH users were shown how EvolveCaptions would transcribe their speech and how the hearing user could correct errors. The hearing participants were guided through the correction interface, and the DHH user was introduced to the clause recording interface for post-session adaptation.

4.2.3 Interactive Trial. Each trial consisted of five scripted readings by the DHH participant, with live caption correction by the hearing partner. Scripts (~5 minutes each) were generated using GPT-4 (~600 words) to ensure consistency in length, linguistic complexity, and conversational tone. Topics included everyday conversational scenarios such as seasonal events, weekend activities, or local news. The order of scripts was randomized per participant to control for sequence effects. Overall, each trial lasted approximately one hour.

During reading, EvolveCaptions displayed live captions, and the hearing participant marked or corrected ASR errors. After each script, the system used these corrections to generate training clauses, which the DHH participant then recorded. Fine-tuning was performed between scripts, progressively adapting the ASR model to the speaker’s voice.

All system interactions—including transcripts, corrections, and recording metadata—were logged for later analysis.

4.2.4 Semi-Structured Interviews. Following the interactive trials, we conducted semi-structured interviews lasting ~15 minutes with each participant pair. DHH users were asked about caption quality, ease of recording, perceived improvement, and envisioned use cases. Hearing participants were asked about the effort required to correct captions, perceived impact, and usability of the interface.

All interviews were video recorded and transcribed using Zoom’s ASR, followed by manual correction. Each participant received a \$50 Amazon gift card as compensation.

4.3 Data Analysis

4.3.1 Quantitative Analysis. We collected system interaction data from all 12 sessions, including: (1) Word Error Rate (WER) for each script pre- and post-adaptation, (2) number of caption corrections made by hearing participants, and (3) number and length of clause recordings by DHH participants.

We calculated Word Error Rate (WER) after each iteration, comparing the first and final sessions using the Wilcoxon signed-rank test [52] to evaluate the statistical significance of improvement. Descriptive statistics (*mean*, *SD*) were used to report correction activity and recording load.

4.3.2 Qualitative Analysis. Interview transcripts were analyzed using applied thematic analysis [23].

One author reviewed all transcripts and created an initial codebook based on emergent themes. This was iteratively refined through discussion with the research team. The final codebook included 4 first-level, 10 second-level, and 25 third-level codes (see Supplementary Material for the full codebook). Two researchers then independently coded the full dataset using the finalized codebook.

Inter-rater reliability was calculated using Krippendorff’s alpha [31] via the ReCal2 package [17]. The average α was 0.86 ($\alpha > 0.8$ is considered a good agreement), with raw agreement of 93%. Disagreements were resolved via consensus.

Themes were then grouped under broader categories and synthesized into a qualitative narrative covering usability, effort, trust, accessibility labor, and system dynamics.

5 Findings

We present our findings from the user study, organized into three parts: (1) quantitative results evaluating system performance and usage; (2) qualitative insights from DHH and hearing participants’ experiences; and (3) implications for future deployment and co-captioning workflows.

5.1 Quantitative Evaluation of EvolveCaptions

To assess the effectiveness of EvolveCaptions, we compared its performance to three baselines:

- **Baseline 1 – Static ASR:** Unmodified Whisper-base model, the same model used in participants’ first session, representing a commercial ASR service without personalization.
- **Baseline 2 – Adapted ASR:** Whisper-base fine-tuned on approximately five minutes of atypical speech (sampled from the TORGO dataset [56]), representing general adaptation without speaker-specific training.
- **Baseline 3 – One-Round Manual Correction:** Captions manually corrected by hearing participants during the first session, serving as a reference for one-time collaborative correction without model adaptation.

Figure 3 illustrates the Word Error Rate (WER) over five sessions for all 12 DHH participants. All participants showed a clear decreasing trend in WER over time, suggesting that EvolveCaptions effectively adapts to individual speech patterns. In contrast, the static Whisper model often plateaued or worsened in comparison. The TORGO-adapted model performed slightly better than the static baseline for some participants but remained less effective than

speaker-specific fine-tuning. For instance, participant P11 began with a WER exceeding 1.0—indicating near-complete unintelligibility (i.e., the number of errors exceeded the number of reference words)—but improved to 0.63 by Session 5, representing the most significant individual gain.

To statistically assess improvement, we conducted a Wilcoxon signed-rank test comparing WER between Session 1 and Session 5. Across participants, the median reduction was 27.2%, with a mean reduction of 30.4%. This decrease was statistically significant ($W = 78.0, p < .05$), confirming the effectiveness of the adaptive workflow. Notably, participants with higher initial WERs tended to exhibit larger gains, while those who began with relatively low WERs (e.g., P3 and P5) saw minimal change. This suggests that the system is especially beneficial for users with more atypical speech.

Individual recording statistics reflect minimal speaker burden: each DHH participant recorded an average of 46.3 audio files ($SD = 14.8$), totaling approximately five minutes of speech (299.2 seconds, $SD = 100.6$, Figure 4). Meanwhile, hearing participants made an average of 72.3 caption edits or highlights ($SD = 17.3$), highlighting strong engagement with the collaborative correction workflow.

5.2 Qualitative Findings from DHH Participants

5.2.1 Captioning Experiences and Expectations. All DHH participants had prior experience using captioning tools for their own speech. While acknowledging these tools as helpful, they consistently noted limitations in accuracy, particularly with automated systems. Several participants described workarounds such as exaggerating articulation, changing platforms, or even switching modalities. Despite frustrations, most still used tools like Google Transcribe or Apple CC. As P11 shared, *“Auto captions, like Zoom, for example, are not able to capture my voice. CART captioners are more familiar with my voice so they can capture it.”*

Participants expressed optimism about the idea of a system that could adapt to their speech over time. While not all were certain whether EvolveCaptions had improved during the short study period, most felt it had. Eight participants explicitly said the system seemed to *“gradually learn their speech,”* which enhanced their sense of agency. Several noted that words previously misrecognized were accurately transcribed in later sessions. As P9 explained, *“It means it will actually recognize my speech patterns. Be also able to adapt to how I talk... so it’s more natural, and I feel like I’m teaching the application my voice.”* These participants described a sense of collaborative training, where each recording felt like a contribution toward better future captioning. Interestingly, a few participants linked the success of the system to their own recording effort, viewing the system as responsive to their participation.

5.2.2 Recording Practices and Preferences. All participants acknowledged that recording their voice required effort, but most described it as worthwhile if it led to improved captions. Eight participants said they would be willing to record several times a week, and four said they would do so daily—especially for high-stakes contexts such as work meetings or technical discussions. P9 shared, *“I’ll spend time recording in situations like work meetings... But in some situations, like, um, casual small talk. I will not be comfortable or in a noisy environment, yeah. Because it won’t be accurate anyway.”*

Two participants anticipated that recording would become less frequent over time, imagining a trajectory of improvement that would reduce the need for further training. As P9 put it, *“On a later stage, it’ll be kind of occasional, or a kind of a more targeted recording, probably during a special event.”*

Participants valued the ability to skip, re-record, or delete prompts, and appreciated the waveform display for visual feedback.

5.2.3 Correction Willingness and Concerns. All participants said they would be comfortable asking someone to help correct captions, especially familiar conversation partners. Seven said they would feel comfortable asking anyone in virtual or in-person meetings, while four preferred to rely on trusted individuals who understood their speech patterns. Some expressed concern when corrected captions did not accurately reflect what they had said, revealing a tension between correction and perceived authenticity. One participant suggested that peer DHH users might be better correction partners, envisioning a future with mutual support.

5.2.4 Perceptions of System Accuracy. A number of participants recognized patterns in how the system responded to correction and recording. They described moments when the system started to *“get it right”* following a recording, particularly with recurring misrecognized words. These perceived improvements helped reinforce the feeling that the system was *“learning,”* even if it did not reach perfection. However, there was also acknowledgment that real-time changes might not always be desirable. As P3 shared, *“The changes on the fly might not be natural and may not be accepted if it distracts from the flow of the conversation.”*

5.2.5 Privacy and Data Handling. Most participants expressed comfort with the system so long as it managed data responsibly. Several compared it to other online tools that process voice or video. Four participants raised specific privacy concerns, including the potential for misuse of voice data and the need for secure storage. Two participants explicitly requested on-device or offline processing options.

5.2.6 Feature Requests. Participants suggested a range of desired features for future iterations of EvolveCaptions. These included user accounts with privacy controls, the ability to track personal progress, and integration with platforms like Zoom or Google Meet. As P3 explained, *“I’m hoping that this speech model that you’re building be incorporated into the platform we rely on, like Zoom.”* A few also suggested customizable vocabulary support for proper nouns or specialized terminology.

5.3 Hearing Participants’ Experiences

All six hearing participants reported that EvolveCaptions was easy to use and exhibited low latency, enabling them to make caption corrections in real-time. Three had prior experience communicating with DHH individuals, while three did not. Across the board, participants noted that correcting captions introduced a modest cognitive burden but was manageable. H6 commented, *“I would be concerned that [making corrections] will be distracting me... But if this focus can help the system recognize my friend’s voice better, I am definitely willing to do it.”*

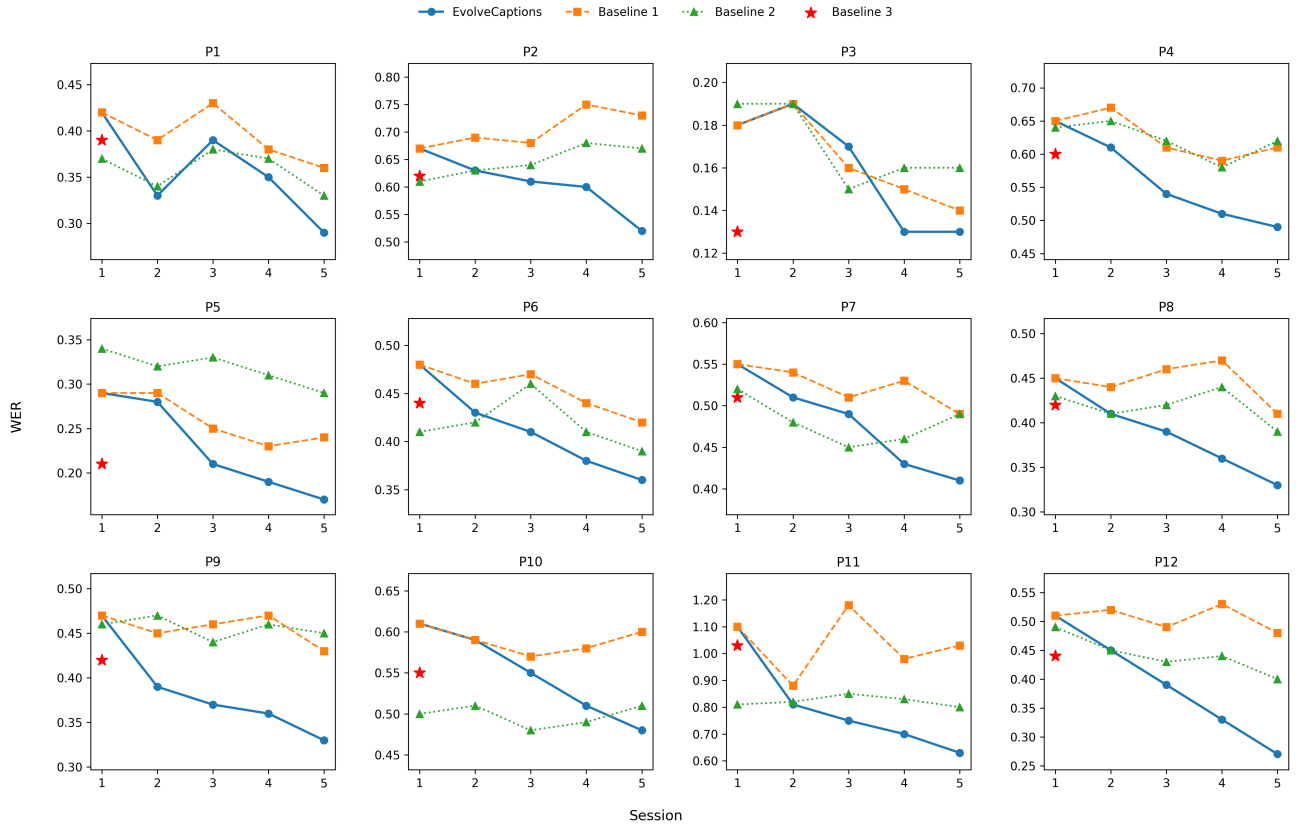


Figure 3: Word Error Rate (WER) improvement across four iterations using EvolveCaptions in our user study.

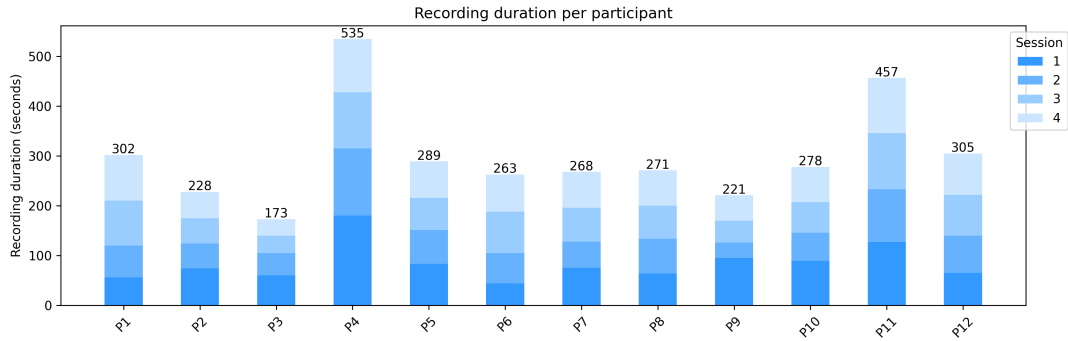


Figure 4: Participant recording durations across four sessions, with segment colors indicating session-specific durations. Totals are shown above each bar.

Participants distinguished between contexts in which they would be willing to provide corrections. In high-stakes scenarios like lectures or meetings, most said they would be more likely to assist. In casual conversations, however, they expected to prioritize the interaction over correction. Four participants said they would be willing to make corrections only for people they knew well, while two said they would help anyone if asked.

Several hearing participants also expressed appreciation for the system’s adaptive nature. H3, who participated in multiple sessions, remarked, “I was impressed by the system adapting to different speakers [who] have so different styles.” This visible sense of improvement enhanced their engagement with the task, reinforcing the feeling that their efforts contributed meaningfully to system learning.

Participants also reported that they became more effective at making corrections over time, developing personalized strategies

for managing the cognitive load. As H1 described, “*I tried to follow so I only highlighted uncertain words at first, and then went back to make corrections.*” H5 similarly explained, “*I focus on repeated words; if I find an error that happens a few times, I will correct them.*” Others prioritized content-bearing words like proper nouns and verbs while ignoring predictable or filler words. After several trials, most participants noted that the correction process became more fluid and intuitive—suggesting the workflow is not only low-friction but also learnable with modest exposure.

6 Discussion

Our findings show that real-time, adaptive captioning can reduce ASR errors for DHH speakers while distributing personalization effort across users. In this section, we situate these results in prior work, reflect on user agency and inclusivity in ASR design, and discuss study limitations and future directions.

6.1 Advancing Communication Accessibility

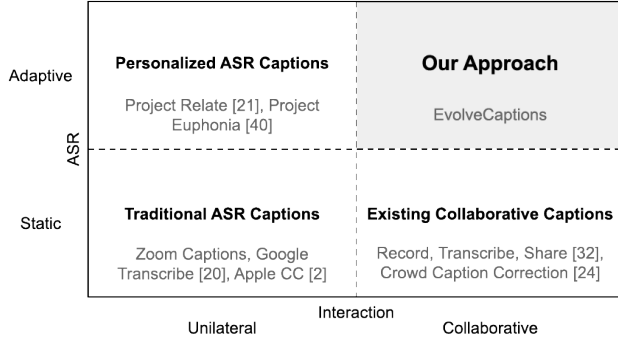


Figure 5: Comparison of captioning technologies for DHH users. The horizontal axis contrasts unilateral versus collaborative interaction, while the vertical axis contrasts static versus adaptive ASR models. EvolveCaptions is positioned in the top-right quadrant, representing a collaborative and adaptive approach.

Prior captioning systems for DHH users have primarily relied on static ASR models, which perform poorly on the diverse and often atypical speech patterns of DHH speakers. Some systems support speaker adaptation through pre-collected data [40], while others offer mechanisms for real-time correction [24, 60]. However, these approaches typically fall short in three key ways: (1) static models do not account for individual speech variability (Figure 5 bottom left); (2) adaptation often places the full burden on the DHH speaker, requiring time-consuming pre-recording or technical tuning (Figure 5 top left); and (3) corrections are transient—errors fixed in one context are not reused in the next (Figure 5 bottom right).

EvolveCaptions integrates these three strands—adaptation, correction, and real-time use—into a unified system that incrementally learns from speakers’ voices and their collaborators’ feedback. Rather than requiring extensive training data, EvolveCaptions leverages short, phonetically rich prompts that users can record during

interaction, allowing lightweight model updates without interrupting conversation. Hearing partners can correct captions in real time, and these corrections are later integrated into fine-tuning. This design distributes the work of personalization across both parties, enabling dynamic improvement without overburdening the DHH user.

By embedding collaborative correction and in-situ adaptation into a low-friction workflow, EvolveCaptions reframes captioning from a static, one-time prediction task into a socially supported, evolving dialogue. In doing so, it addresses not only technical challenges in ASR accuracy but also the lived experiences of DHH speakers navigating tools that often fail to recognize their voices.

6.2 Ability-Based Personalization

Wobbrock et al. [65] proposed ability-based design, arguing that technologies should adapt to the abilities of users, rather than requiring users to adapt to technologies. Yet, most captioning systems invert this ideal, placing the burden on DHH individuals. When ASR fails to recognize their voice, they must repeat themselves, exaggerate pronunciation, or modify the content of speech to fit the system. These practices not only increase communication effort but also change users’ natural communication.

EvolveCaptions inverts this dynamic by tuning the system to the user’s voice—not the other way around. Through targeted prompts and in-situ corrections, users retain their own speech style while benefiting from gradually improving caption quality. This reflects a broader stance on accessibility: technologies should preserve and respect how people naturally communicate, rather than imposing new constraints. Aligned with preservation-oriented approaches in assistive technology [6, 43], our goal is not to “fix” the user or their speech, but to develop systems that flex to human diversity.

6.3 Agency and Human-Centered Adaptation

EvolveCaptions makes adaptation transparent and participatory. Rather than operating as an opaque black box, the system shows users that their recordings and corrections lead to meaningful improvement. This aligns with emerging guidelines for human-centered machine learning, which call for intelligibility, controllability, and feedback [1].

Participants described the system as something they were “teaching,” reflecting a shift in the user-tool relationship. Corrections and recordings became not just usability tasks, but acts of co-training—a collaborative effort to improve accessibility. Future systems could go further by surfacing vocabularies, visualizing learning progress, or letting users modulate how aggressively the model adapts.

6.4 Social Dynamics of Co-Captioning

Our study also surfaces the interpersonal dimensions of co-captioning. Hearing participants took on the role of caption correctors—a novel interaction behavior that shaped how they attended to conversation. Over time, they became more adept at making real-time corrections, suggesting that this form of contribution is learnable and sustainable with light practice. DHH participants, in turn, reported that seeing corrections and recording responses felt validating, reinforcing their sense that the system was learning from them.

This dynamic highlights a key insight: ASR adaptation is not purely technical—it is socially mediated. Trust, familiarity, and shared effort shape not only how well the system performs, but also whether users are willing to invest in it. Future research should explore how this social co-adaptation unfolds over longer time periods and in more diverse contexts, such as classrooms or multilingual conversations.

6.5 Generalizability to Other Non-Normative Speech

While this work centers on DHH speakers, the underlying framework of EvolveCaptions—speaker-specific adaptation through minimal data, real-time corrections, and incremental fine-tuning—could generalize to other populations with non-normative speech. This includes people with dysarthria, stroke survivors, children, elderly users, or non-native speakers. Personalized ASR systems that adapt in the flow of communication, rather than relying on static models trained on normative speech corpora, could dramatically expand accessibility across a wide spectrum of users.

6.6 Ethical and Deployment Considerations

Participants expressed moderate concerns about privacy, with some requesting on-device processing or the ability to delete data. However, beyond technical safeguards, co-captioning also raises new ethical questions: Who owns the adapted model? Can others make corrections that change what a user “said”? How should consent be handled in shared or public interactions?

These questions underline the need for consent models and governance mechanisms that account for collaborative learning. While our participants viewed corrections as helpful and empowering, designers must ensure that future systems do not inadvertently override users’ voices or misrepresent intent.

6.7 Limitations and Future Work

While EvolveCaptions shows promise in improving ASR performance for DHH speakers with limited user effort, our study has several limitations and opens up important avenues for future research.

Controlled Evaluation Context. We conducted a lab-style, remote study with scripted speech and dyadic interaction. This controlled setup enabled us to isolate adaptation effects and analyze behavioral patterns systematically. However, scripted reading does not fully capture the dynamics of spontaneous conversation, such as interruptions, overlapping speech, or contextual disfluencies. Future work will involve longitudinal field deployments of EvolveCaptions in real-world communication settings, such as classrooms, workplaces, and video conferencing platforms (e.g., Zoom, Google Meet), to evaluate the system’s adaptability under more natural conditions and interactional diversity.

System Deployment Constraints. For this evaluation, EvolveCaptions was deployed on a cloud-based server to ensure consistency across participants. While the underlying pipeline can run locally, the cloud setup allowed us to control for hardware variability and focus on testing adaptation mechanisms. To support real-world deployment, future work will explore lightweight on-device fine-tuning, real-time model updates under limited compute,

and robustness in non-ideal conditions such as noisy environments or poor microphone quality.

Collaborative Correction and Cognitive Load. Although hearing participants reported that real-time correction was manageable and meaningful, the cognitive effort required—especially during fast-paced dialogue—remains a challenge. In long-term use or high-pressure scenarios, this effort may become unsustainable. We plan to explore interaction designs that reduce correction load while maintaining effectiveness, as well as evaluate hearing participants’ experience more rigorously using tools like NASA-TLX [7] or post-task comprehension assessments.

Long-Term Personalization. While EvolveCaptions supports incremental adaptation, it does not yet account for temporal variability in speech, such as changes due to fatigue, illness, stress, or post-surgical recovery. Future systems must incorporate mechanisms to detect when personalization becomes outdated, offer options for re-adaptation, and allow users to manage and sync their personalized models across devices and contexts. Supporting longitudinal flexibility will be critical for ensuring that adaptive captioning remains robust and equitable over time.

Interaction Dynamics. Previous work has explored behavioral patterns in ASR-mediated communication, such as hearing speakers slowing down or DHH participants modifying gaze behavior [57–59]. Our study adds a novel layer to this interaction: hearing participants actively monitored and corrected captions in real time, while DHH participants selectively recorded words based on those corrections. This form of mediated collaboration suggests that ASR personalization is not only technical but also social—amplified by trust, familiarity, and conversational rapport. Future work could systematically examine these co-adaptive behaviors across diverse speaker dyads to better understand how interpersonal dynamics shape the success of collaborative captioning.

7 Conclusion

EvolveCaptions demonstrates a new approach to ASR personalization for DHH speakers by combining live human correction, targeted data collection, and collaborative interaction. By involving hearing participants in the correction process and focusing training on misrecognized speech, the system enables low-effort, real-time adaptation grounded in principles of collective access. Our evaluation with 12 DHH and six hearing participants showed that even a small set of targeted recordings can substantially improve captioning quality. These findings highlight a promising path towards equitable, human-centered ASR systems that not only adapt to individual speech patterns but also reimagine accessibility as a collaborative, socially evolving process.

References

- [1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.
- [2] Apple. 2025. Get live captions of spoken audios on iPhone. <https://support.apple.com/guide/iphone/get-live-captions-of-spoken-audio-iphe0990f7bb/ios> Accessed: 2025-03-10.
- [3] Murali Karthick Baskar, Tim Herzig, Diana Nguyen, Mireia Diez, Tim Polzehl, Lukáš Burget, Jan Černocký, et al. 2022. Speaker adaptation for Wav2vec2 based dysarthric ASR. *arXiv preprint arXiv:2204.00770* (2022).

- [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*. 155–164.
- [5] Jeffrey P Bigham, Raja Kushalnagar, Ting-Hao Kenneth Huang, Juan Pablo Flores, and Saiph Savage. 2017. On how deaf people might use speech to control devices. In *Proceedings of the 19th international ACM SIGACCESS conference on computers and accessibility*. 383–384.
- [6] Jeffrey P Bigham, Craig M Prince, and Richard E Ladner. 2008. WebAnywhere: a screen reader on-the-go. In *Proceedings of the 2008 international cross-disciplinary conference on Web accessibility (W4A)*. 73–82.
- [7] Alex Cao, Keshav K Chintamani, Abhilash K Pandya, and R Darin Ellis. 2009. NASA TLX: Software for assessing subjective mental workload. *Behavior research methods* 41, 1 (2009), 113–117.
- [8] Anna Cavender and Richard E Ladner. 2008. Hearing impairments. In *Web accessibility: A foundation for research*. Springer, 25–35.
- [9] Petr Cerva, Jan Silovsky, Jindrich Zdánský, Jan Nouza, and Jiri Malek. 2012. Real-Time Lecture Transcription using ASR for Czech Hearing Impaired or Deaf Students. In *INTERSPEECH*. 763–766.
- [10] Si Chen, James Waller, Matthew Seit, Christian Vogler, Raja Kushalnagar, and Qi Wang. 2024. Towards Co-Creating Access and Inclusion: A Group Autoethnography on a Hearing Individual's Journey Towards Effective Communication in Mixed-Hearing Ability Higher Education Settings. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [11] VoidZero Inc. & Vite Contributors. 2025. Vite: Next Generation Frontend Tooling. <https://vite.dev>. Accessed: 2025-09-10.
- [12] Leo De Raeye. 2010. A longitudinal study on auditory perception and speech intelligibility in deaf children implanted younger than 18 months in comparison to those implanted at later ages. *Otology & Neurology* 31, 8 (2010), 1261–1267.
- [13] UW DO-IT. 2021. What is real-time captioning? <https://www.washington.edu/doi/what-real-time-captioning>. Accessed: 2025-03-10.
- [14] Lisa 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*. 271–272.
- [15] 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*. 405–406.
- [16] Raymond Fok, Harmanpreet Kaur, Skanda Palani, Martez E Mott, and Walter S Lasecki. 2018. Towards more robust speech interactions for deaf and hard of hearing users. In *Proceedings of the 20th international ACM SIGACCESS conference on computers and accessibility*. 57–67.
- [17] Deen Freelon. 2010. ReCal2: Reliability for 2 Coders. <http://dfreelon.org/utis/recalfront/recal2/>. Accessed: 2025-03-10.
- [18] Rupak Raj Ghimire, Bal Krishna Bal, and Prakash Poudyal. 2024. A Comprehensive Study of the Current State-of-the-Art in Nepali Automatic Speech Recognition Systems. *arXiv preprint arXiv:2402.03050* (2024).
- [19] Abraham Glasser, Kesavan Kushalnagar, and Raja Kushalnagar. 2017. Deaf, hard of hearing, and hearing perspectives on using automatic speech recognition in conversation. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. 427–432.
- [20] Google. 2025. Live Transcribe & notification. <https://play.google.com/store/apps/details?id=com.google.audio.hearing.visualization.accessibility.scribe>. Accessed: 2025-03-10.
- [21] Google Research. 2025. Project Relate. <https://sites.research.google/relate/>. Accessed: 2025-09-08.
- [22] Linda G Gottermeier, CAROL L DE FILIPPO, R AJA KUSHALNAGAR, and BONNIE L BASTIAN. 2016. User evaluation of automatic speech recognition systems for deaf-hearing interactions at school and work. *Audiology Today* 28, 2 (2016), 20–34.
- [23] Greg Guest, Kathleen M MacQueen, and Emily E Namey. 2011. *Applied thematic analysis*. sage publications.
- [24] Rebecca Perkins Harrington and Gregg C Vanderheiden. 2013. Crowd caption correction (ccc). In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–2.
- [25] Sarah E Hughes, Liang-Yuan Wu, Lindsay J Ma, Dhruv Jain, and Michael M McKee. 2025. Assessing the Role of Medical Caption Technology to Support Physician-Patient Communication for Patients with Hearing Loss: A Pilot Study (Preprint). (June 2025). doi:10.2196/preprints.79073
- [26] Lisa I Iezzoni, Bonnie L O'Day, Mary Killeen, and Heather Harker. 2004. Communicating about health care: observations from persons who are deaf or hard of hearing. *Annals of internal medicine* 140, 5 (2004), 356–362.
- [27] Dhruv Jain, Bonnie Chinh, Leah Findlater, Raja Kushalnagar, and Jon Froehlich. 2018. Exploring augmented reality approaches to real-time captioning: A preliminary autoethnographic study. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems*. 7–11.
- [28] Yicong Jiang, Tianzi Wang, Xurong Xie, Juan Liu, Wei Sun, Nan Yan, Hui Chen, Lan Wang, Xunying Liu, and Feng Tian. 2024. Perceiver-prompt: Flexible speaker adaptation in Whisper for chinese disordered speech recognition. *arXiv preprint arXiv:2406.09873* (2024).
- [29] 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*. 165–174.
- [30] Saba Kaws, 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*. 15–23.
- [31] Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.
- [32] Korbinian Kuhn, Benedikt Reuter, Niklas Egger, and Gottfried Zimmermann. 2024. Record, Transcribe, Share: An Accessible Open-Source Video Platform for Deaf and Hard of Hearing Viewers. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–6.
- [33] Raja S Kushalnagar, Walter S Lasecki, and Jeffrey P Bigham. 2013. Captions versus transcripts for online video content. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*. 1–4.
- [34] Raja S Kushalnagar, Walter S Lasecki, and Jeffrey P Bigham. 2014. Accessibility evaluation of classroom captions. *ACM Transactions on Accessible Computing (TACCESS)* 5, 3 (2014), 1–24.
- [35] Raja S Kushalnagar and Christian Vogler. 2020. Teleconference accessibility and guidelines for deaf and hard of hearing users. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–6.
- [36] Paddy Ladd. 2003. *Understanding deaf culture: In search of deafhood*. Multilingual Matters.
- [37] 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*. 23–34.
- [38] Fernando Loizides, Sara Basson, Dimitri Kanevsky, Olga Prilepova, Sagar Savla, and Susanna Zaraysky. 2020. Breaking boundaries with live transcribe: Expanding use cases beyond standard captioning scenarios. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–6.
- [39] Robert L MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn A Ladewig, Jimmy Tobin, Michael P Brenner, Philip C Nelson, et al. 2021. Disordered Speech Data Collection: Lessons Learned at 1 Million Utterances from Project Euphonia. In *Interspeech*, Vol. 2021. 4833–4837.
- [40] Alicia Martin, Robert L MacDonald, Pan-Pan Jiang, Marilyn Ladewig, Julie Cattiau, Rus Heywood, Richard Cave, Jimmy Tobin, Philip C Nelson, and Katrin Tomanek. 2025. Project Euphonia: advancing inclusive speech recognition through expanded data collection and evaluation. *Frontiers in Language Sciences* 4 (2025), 1569448.
- [41] Sven Mattys, Ann Bradlow, Matthew Davis, and Sophie Scott. 2013. *Speech recognition in adverse conditions: Explorations in behaviour and neuroscience*. Psychology Press.
- [42] Emma McDonnell. 2022. Understanding social and environmental factors to enable collective access approaches to the design of captioning technology. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–8.
- [43] Emma J McDonnell, Soo Hyun Moon, Lucy Jiang, Steven M Goodman, Raja Kushalnagar, Jon E Froehlich, and Leah Findlater. 2023. "Easier or Harder, Depending on Who the Hearing Person Is": Codesigning Videoconferencing Tools for Small Groups with Mixed Hearing Status. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [44] Kinfe Tadesse Mengistu and Frank Rudzicz. 2011. Comparing humans and automatic speech recognition systems in recognizing dysarthric speech. In *Canadian Conference on Artificial Intelligence*. Springer, 291–300.
- [45] Pam Millett. 2021. Accuracy of Speech-to-Text Captioning for Students Who are Deaf or Hard of Hearing. *Journal of Educational, Pediatric & (Re) Habilitative Audiology* 25 (2021).
- [46] Matthew S Moore and Linda Levitan. 1993. For Hearing People Only: Answers to some of the most commonly asked questions about the deaf community, its culture, and the "deaf reality". (*No Title*) (1993).
- [47] Mozilla. 2025. AudioWorklet - Web APIs. <https://developer.mozilla.org/en-US/docs/Web/API/AudioWorklet>. Accessed: 2025-09-10.
- [48] Davide Mulfari and Massimo Villari. 2024. A voice user interface on the edge for people with speech impairments. *Electronics* 13, 7 (2024), 1389.
- [49] Alisha Pradhan, Kanika Mehta, and Leah Findlater. 2018. "Accessibility Came by Accident" Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. In *Proceedings of the 2018 CHI Conference on human factors in computing systems*. 1–13.
- [50] Soraia Silva Prieth, Napoliana Silva de Souza, and Lucia Villela Leite Filgueiras. 2014. A speech-to-text system's acceptance evaluation: would deaf individuals adopt this technology in their lives?. In *International Conference on Universal*

- Access in Human-Computer Interaction*. Springer, 440–449.
- [51] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR, 28492–28518.
 - [52] Judy Robertson and Maurits Kaptein. 2016. An introduction to modern statistical methods in HCI. In *Modern Statistical Methods for HCI*. Springer, 1–14.
 - [53] Jason Rodolitz, Evan Gambill, Brittany Willis, Christian Vogler, and Raja Kushalnagar. 2019. Accessibility of voice-activated agents for people who are deaf or hard of hearing. *Journal on Technology and Persons with Disabilities* 7 (2019), 144–156.
 - [54] Hannah P Rowe, Sarah E Gutz, Marc F Maffei, Katrin Tomanek, and Jordan R Green. 2022. Characterizing dysarthria diversity for automatic speech recognition: A tutorial from the clinical perspective. *Frontiers in computer science* 4 (2022), 770210.
 - [55] Frank Rudzicz. 2010. Towards a noisy-channel model of dysarthria in speech recognition. In *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies*. 80–88.
 - [56] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. 2012. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language resources and evaluation* 46, 4 (2012), 523–541.
 - [57] Matthew Seita, Khaled Albusays, Sushant Kafle, Michael Stinson, and Matt Huenerfauth. 2018. Behavioral changes in speakers who are automatically captioned in meetings with deaf or hard-of-hearing peers. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. 68–80.
 - [58] Matthew Seita, Sarah Andrew, and Matt Huenerfauth. 2021. Deaf and hard-of-hearing users’ preferences for hearing speakers’ behavior during technology-mediated in-person and remote conversations. In *Proceedings of the 18th International Web for All Conference*. 1–12.
 - [59] Matthew Seita and Matt Huenerfauth. 2020. Deaf individuals’ views on speaking behaviors of hearing peers when using an automatic captioning app. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.
 - [60] Matthew Seita, Sooyeon Lee, Sarah Andrew, Kristen Shinohara, and Matt Huenerfauth. 2022. Remotely co-designing features for communication applications using automatic captioning with deaf and hearing pairs. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–13.
 - [61] Joel Shor, Dotan Emanuel, Oran Lang, Omry Tuval, Michael Brenner, Julie Cattiau, et al. 2019. Personalizing ASR for Dysarthric and Accented Speech with Limited Data. In *Proceedings of Interspeech 2019*. 784–788. doi:10.21437/Interspeech.2019-1427
 - [62] Meta Open Source. 2025. React:The library for web and native user interfaces. <https://react.dev> Accessed: 2025-09-10.
 - [63] Jimmy Tobin and Katrin Tomanek. 2022. Personalized automatic speech recognition trained on small disordered speech datasets. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6637–6641.
 - [64] Mike Wald and Keith Bain. 2008. Universal access to communication and learning: the role of automatic speech recognition. *Universal Access in the Information Society* 6, 4 (2008), 435–447.
 - [65] Jacob O Wobbrock, Shaun K Kane, Krzysztof Z Gajos, Susumu Harada, and Jon Froehlich. 2011. Ability-based design: Concept, principles and examples. *ACM Transactions on Accessible Computing (TACCESS)* 3, 3 (2011), 1–27.
 - [66] Kenta Yamamoto, Ippei Suzuki, Akihisa Shitara, and Yoichi Ochiai. 2021. See-through captions: real-time captioning on transparent display for deaf and hard-of-hearing people. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–4.
 - [67] Robin Zhao, Anna SG Choi, Allison Koenecke, and Anaïs Rameau. 2025. Quantification of automatic speech recognition system performance on d/deaf and hard of hearing speech. *The Laryngoscope* 135, 1 (2025), 191–197.