EvolveCaptions: Real-Time Collaborative ASR Adaptation for DHH Speakers

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

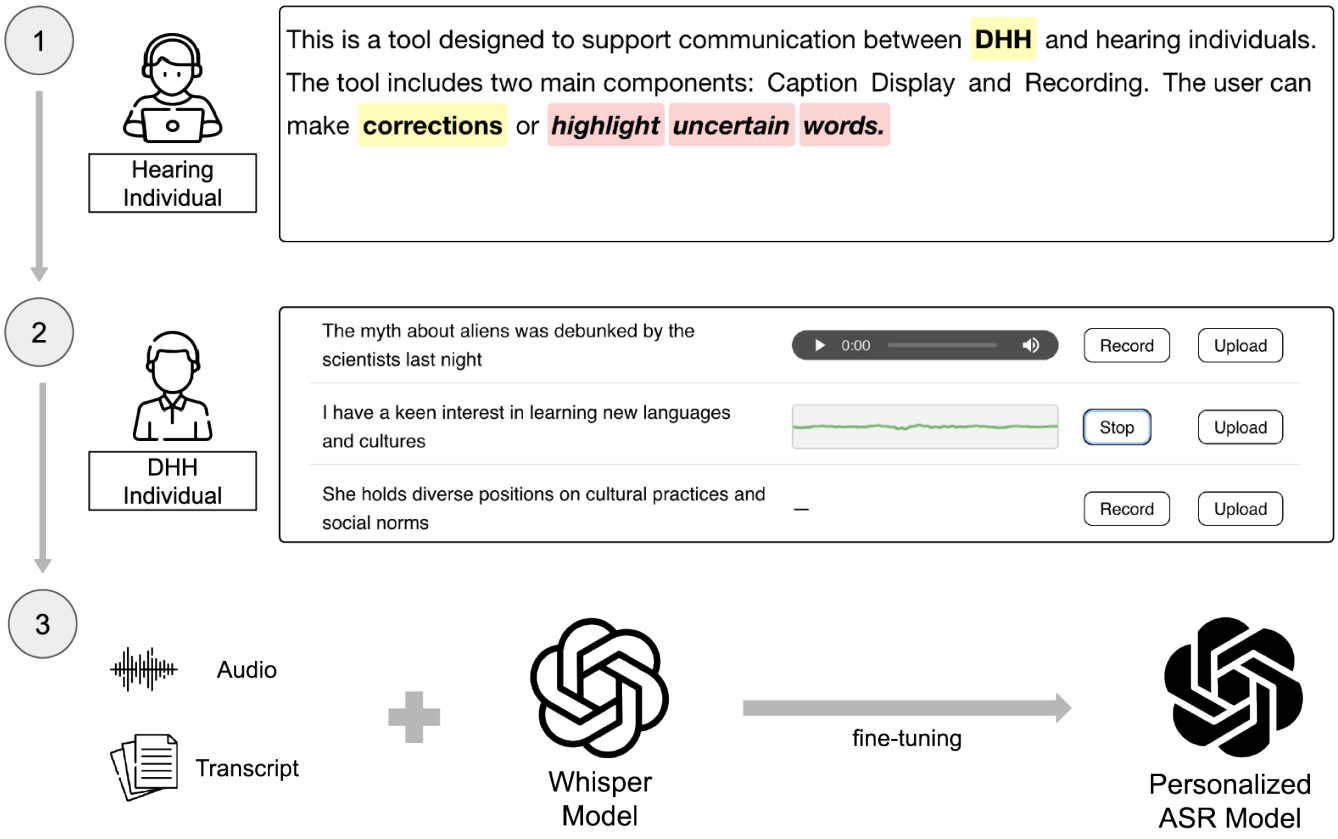

Figure 1: Overview of EvolveCaptions. (1) Hearing users correct live captions of the DHH speaker's voice. (2) The DHH speaker records targeted phrases generated from the corrected terms. (3) The Whisper ASR model is fine-tuned with the recordings and adapts to the speaker over time.

Abstract

Current ASR systems struggle to reliably recognize the speech of Deaf and Hard of Hearing (DHH) individuals, particularly in real-time communication. Existing personalization methods typically require extensive pre-recorded data and place the burden entirely on DHH users. We present EvolveCaptions, a live ASR adaptation system that supports collaborative, in-the-moment personalization. Hearing participants correct ASR errors during conversation, and the system generates short, phonetically relevant phrases for the DHH speaker to record. These recordings are then used to iteratively fine-tune the ASR model. In a preliminary evaluation, our system reduced word error rate from 0.53 to 0.27 over four adaptation rounds with minimal user effort. This work introduces a low-effort, socially collaborative method for adapting ASR to diverse DHH voices in real-world settings.

ASSETS '25, Denver, CO, USA

ACM ISBN 979-8-4007-0676-9/2025/10
https://doi.org/10.1145/3663547.3759723

CCS Concepts

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

Keywords

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

ACM Reference Format:

Liang-Yuan Wu and Dhruv Jain. 2025. EvolveCaptions: Real-Time Collaborative ASR Adaptation for DHH Speakers. In *The 27th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '25), October 26–29, 2025, Denver, CO, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3663547.3759723

1 Introduction

Automatic Speech Recognition (ASR) has found widespread application and stands as a crucial tool enhancing communication experiences for people. However, the diverse array of speaking styles presents a challenge in developing a universally accurate ASR system capable of recognizing speech from different individuals [1, 3, 13]. Despite the remarkable performance of state-of-the-art ASR models on numerous speech benchmarks [6], they still struggle to accurately transcribe the speech of deaf and hard of hearing (DHH) individuals [5, 16], contributing to inequities in access to communication technologies for people with atypical or disordered speech [4, 7]. DHH speech can vary widely in pronunciation—both across and within individuals—sometimes making it challenging to understand, even for familiar listeners [2, 10].

Prior approaches to improving ASR for DHH speakers, such as Project Euphonia [9] and Tobin et al. [15], have relied on large-scale, pre-recorded datasets collected in controlled settings. However, these methods face three key limitations. First, they impose a **substantial motivational burden:** asking DHH individuals to spend hours recording scripted speech is taxing and impractical. Second, they **lack contextual relevance:** the collected samples are disconnected from real-world usage, limiting their ability to generalize to the variability and spontaneity of live conversation. Third, they **treat ASR adaptation as a solitary, pre-emptive task,** placing the full responsibility on the DHH individual, with no mechanisms for real-time, collaborative improvement during actual communication.

In this demo paper, we present *EvolveCaptions*, an interactive ASR adaptation system that enables real-time, collaborative personalization of ASR to a DHH speaker's voice during live conversations. The system works as follows in a mixed-ability setting: when a DHH individual speaks, the ASR system transcribes their speech in real-time. Hearing participants simultaneously read the transcript and correct any errors. These error-marked segments are sent to a language model (GPT-4), which generates phonetically plausible alternatives based on the identified keywords. The DHH speaker then records only those brief corrected phrases, which are used to fine-tune the ASR model. This process is repeated iteratively, allowing the model to incrementally adapt to the speaker's unique vocal characteristics with minimal effort.

By focusing only on misrecognized segments, EvolveCaptions minimizes the recording effort required from the DHH speaker while maximizing the relevance of collected training data. This reduces the motivational barrier to participation and ensures that ASR adaptation is grounded in the actual context of communication, which static, pre-recorded methods fail to achieve. Our approach also reframes ASR training as a dynamic, in-the-moment process and redistributes the accessibility workload: hearing participants actively contribute by flagging and correcting errors as they occur. This design draws on the principle of *collective access*, which emphasizes shared responsibility in creating accessible environments [11, 12, 14].

We conducted a preliminary evaluation with one DHH-hearing participant pair, where the DHH user read a 10-minute script over five iterative trials. After each trial, the hearing participant corrected transcription errors (around 20–25 per session) and the DHH participant recorded short corrected segments (~2 minutes per trial). We observed a substantial improvement in transcription quality, with Word Error Rate (WER) dropping from 0.53 to 0.27 over five rounds of adaptation. Qualitative feedback from both participants highlighted that the system was intuitive, low-effort, and seamlessly integrated into the communication flow.

In summary, our work contributes a novel, live-collaborative method for adapting ASR to DHH speakers, leveraging both machine intelligence and social interaction to reduce effort and improve equity in communication technology.

2 The System

EvolveCaptions is an interactive ASR adaptation system designed to support mixed-hearing communication by enabling real-time caption correction and lightweight, speaker-specific model fine-tuning. The system allows hearing participants to collaboratively improve captions while automatically generating targeted prompts for the DHH speaker to record, thus refining the ASR model with minimal extra effort. We describe the design motivations, key components, and technical implementation of the system (see Figure 1).

2.1 Design Rationale

EvolveCaptions is grounded in three design goals:

(1) **Low-effort personalization**: DHH users contribute short recordings only for misrecognized words, dramatically reducing the time and effort typically required for model adaptation [15].
(2) **In-situ adaptation**: Instead of relying on pre-collected datasets, our system adapts the ASR model in the context of real communication, improving both relevance and effectiveness.
(3) **Collaborative correction**: The system enables hearing users to assist by identifying and correcting ASR errors, aligning with the principle of *collective access* [11, 14].

These design choices reflect a shift from solitary, pre-emptive adaptation to a socially co-constructed ongoing process.

2.2 Interaction Workflow

Each EvolveCaptions session includes a three-stage loop:

(1) **Live caption correction**: The system transcribes the DHH speaker's voice and displays live captions. Hearing users can highlight errors and make corrections in real-time (Figure 1.1).
(2) **Clause generation and recording**: The system uses corrected words to generate short, natural clauses (via GPT-4)

that include the corrected term. The DHH speaker reads only these targeted phrases (Figure 1.2).

(3) **ASR model update**: The system fine-tunes the ASR model in the background using the new recordings, improving its performance over time (Figure 1.3).

This loop is repeated across conversations, progressively refining the ASR model to better match the speaker's unique vocal patterns.

2.3 System Components

EvolveCaptions consists of four main components:

2.3.1 Caption Display. The system transcribes speech using a Whisper-based ASR engine and displays real-time captions to all participants. Hearing users can correct transcription errors directly by selecting individual words or short phrases. Inspired by prior crowd-correction interfaces [8], our design allows users to mark confirmed corrections (highlighted in yellow) and uncertain segments (in red), offering flexible feedback even when exact transcriptions are unclear (see Figure 1.1). All corrections are broadcast to participants in real-time.

2.3.2 Clause Generation. Corrections provided by hearing users are used to generate full clauses for the DHH speaker to record. For each corrected error, the system creates one short clause using OpenAI GPT-4. The model is prompted to produce phrases that are conversational (5–15 words), natural-sounding, and phonemically similar to the original misrecognized utterance (see full prompt in the Appendix). For example, if *"fok"* was misrecognized as *"fork,"* the system might generate: *"She picked up the fork from the table."* This contextual generation ensures that speaker recordings are both natural and effective for improving model performance.

2.3.3 Recording Collection. Generated clauses are presented to the DHH speaker in a user-friendly interface that allows them to record selectively (see Figure 1.2). Each clause includes a waveform display for visual feedback and playback functionality. Users can re-record, delete, or upload samples at will. The aim is to make recording intuitive and low-effort while ensuring high-quality, targeted data collection.

2.3.4 ASR Engine. EvolveCaptions uses OpenAI's Whisper base model for real-time speech recognition and personalized fine-tuning. During inference, the engine receives 16 kHz, 16-bit PCM audio via WebSocket and transcribes it using a low-latency setup. Whisper runs on an NVIDIA RTX 4090 GPU and streams captions to the interface with minimal delay, comparable to commercial ASR services. For fine-tuning, corrected audio-text pairs are formatted as HuggingFace datasets. Data is padded and collated for batch training, and the model is fine-tuned using Seq2SeqTrainer with lightweight settings (learning rate 1e-5, batch size 8, max steps 100). The updated model is deployed for future sessions.

2.4 System Implementation

EvolveCaptions is built as a cross-platform web application with a Python-based backend. The frontend is implemented in ReactJS using Vite for fast deployment. It includes modules for real-time caption viewing, correction, and DHH recording. Audio input is captured using React's Audio Worklet and streamed via WebSocket to the backend, which is built in FastAPI. The backend handles caption inference, correction tracking, and model training. We extended the open-source WhisperLive project for low-latency, chunked transcription. Corrections and metadata are stored in CSV/JSON formats, and a connection manager tracks sessions between users. All communication is secured over HTTPS with CORS enabled, supporting integration with external web tools or deployment on public servers.

3 Pilot User Evaluation

To evaluate EvolveCaptions, we conducted a preliminary study with one DHH–hearing participant pair. Both participants were fluent in written English and familiar with captioning tools. The DHH participant (33, male) was severe-to-profound deaf and self-identified as hard-of-hearing. The evaluation was designed to simulate a natural interaction while supporting controlled iteration. The DHH participant read aloud a 10-minute passage sourced from CNN Health news, while the hearing participant monitored and corrected the captions. After each session, the DHH participant recorded system-generated prompts. This cycle was repeated four times in one day to balance adaptation opportunities with participant effort. Each session lasted approximately 15 minutes.

Across the four iterations, the hearing participant made an average of 24 corrections per session (28, 21, 23, 24). The DHH participant contributed roughly 2 minutes of audio per round (160, 113, 125, 106 seconds). Starting with the Whisper base model, the initial Word Error Rate (WER) was 0.53. After the first adaptation round, WER dropped to 0.36, and by the fourth round declined to 0.27, demonstrating the system's ability to improve recognition accuracy with minimal, targeted user effort (see Figure 2).

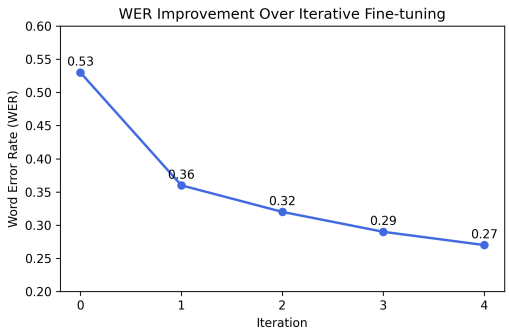

Figure 2: Word Error Rate (WER) improvement across four iterations using EvolveCaptions in our preliminary evaluation.

Qualitative feedback from both participants was generally positive. The DHH participant found the interface easy to use and appreciated that the recording prompts targeted phonemes they often pronounced differently—such as the "sh" sound—making the training feel personalized and purposeful. They also noted satisfaction in seeing ASR accuracy improve over time with little burden.

The hearing participant found the correction interface intuitive and described the editing task as lightweight. However, they noted that in fast-paced conversations, making timely corrections could

be cognitively demanding. They suggested that correction might be more sustainable in settings where the hearing person plays a more passive or secondary listening role (e.g., group meetings). Both participants appreciated that the system allowed flexible participation—they could choose to correct or record as much or as little as they had energy or time for.

4 Limitations and Future Work

While EvolveCaptions shows promising results in improving ASR performance for DHH speakers with limited user effort, this work has several limitations. First, our evaluation involved only a single DHH–hearing pair, and further studies are needed with a broader range of speakers and speech styles to generalize findings. Second, the study involved the DHH participant reading the same script multiple times—useful for measuring incremental model adaptation but not reflective of natural conversational contexts. Third, our fine-tuning pipeline, while functional, remains resource-intensive and could benefit from further exploration of lightweight, on-device, real-time adaptation strategies.

Future work will explore long-term, in-situ deployments of EvolveCaptions, including use in classroom or meeting environments where multiple hearing participants may contribute corrections collaboratively. We also aim to investigate how many corrections are sufficient for meaningful improvement, which types of recognition errors matter most for performance, and how to optimize the balance between user effort and ASR gains. Our planned larger-scale study would help us better understand how correction behaviors vary across contexts and how the system can adapt accordingly.

5 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 ASR adaptation grounded in principles of collective access. Our preliminary findings show that even a small set of targeted recordings can substantially improve model accuracy. We believe this work highlights a promising direction for equitable, human-centered ASR systems that adapt not only to speech but also recognize accessibility as a collaborative, social process.

References

[1] 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.

[2] 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.

[3] Petr Cerva, Jan Silovskỳ, Jindrich Zdánskỳ, Jan Nouza, and Jiri Malek. 2012. Real-Time Lecture Transcription using ASR for Czech Hearing Impaired or Deaf Students.. In *INTERSPEECH*. 763–766.

[4] 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.

[5] 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.

[6] 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).

[7] 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.

[8] 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.

[9] Alicia Martin, Robert MacDonald, Pan-Pan Jiang, Marilyn Ladewig, Julie Cattiau, Rus Heywood, Richard Cave, Jimmy Tobin, Philip C Nelson, and Katrin Tomanek. [n. d.]. Project Euphonia: Advancing Inclusive Speech Recognition through Expanded Data Collection and Evaluation. *Frontiers in Language Sciences* 4 ([n. d.]), 1569448.

[10] Sven Mattys, Ann Bradlow, Matthew Davis, and Sophie Scott. 2013. *Speech recognition in adverse conditions: Explorations in behaviour and neuroscience*. Psychology Press.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

A Prompt provided to GPT-4 for generating recording clauses

Prompt

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).