The Laryngoscope

Quantification of Automatic Speech Recognition System Performance on d/Deaf and Hard of Hearing Speech

Robin Zhao, BS; Anna S.G. Choi, MS; Allison Koenecke, PhD; Anaïs Rameau, MD, MPhil, MS

Objective: To evaluate the performance of commercial automatic speech recognition (ASR) systems on d/Deaf and hard-of-hearing (d/Dhh) speech.

Methods: A corpus containing 850 audio files of d/Dhh and normal hearing (NH) speech from the University of Memphis Speech Perception Assessment Laboratory was tested on four speech-to-text application program interfaces (APIs): Amazon Web Services, Microsoft Azure, Google Chirp, and OpenAI Whisper. We quantified the Word Error Rate (WER) of API transcriptions for 24 d/Dhh and nine NH participants and performed subgroup analysis by speech intelligibility classification (SIC), hearing loss (HL) onset, and primary communication mode.

Results: Mean WER averaged across APIs was 10 times higher for the d/Dhh group (52.6%) than the NH group (5.0%). APIs performed significantly worse for "low" and "medium" SIC (85.9% and 46.6% WER, respectively) as compared to "high" SIC group (9.5% WER, comparable to NH group). APIs performed significantly worse for speakers with prelingual HL relative to postlingual HL (80.5% and 37.1% WER, respectively). APIs performed significantly worse for speakers primarily communicating with sign language (70.2% WER) relative to speakers with both oral and sign language communication (51.5%) or oral communication only (19.7%).

Conclusion: Commercial ASR systems underperform for d/Dhh individuals, especially those with "low" and "medium" SIC, prelingual onset of HL, and sign language as primary communication mode. This contrasts with Big Tech companies' promises of accessibility, indicating the need for ASR systems ethically trained on heterogeneous d/Dhh speech data.

Key Words: artificial intelligence, voice.

Level of Evidence: 3

***Laryngoscope*, 135:191–197, 2025**

INTRODUCTION

Artificial intelligence (AI) applied to speech data has accelerated significantly in the past two decades and has fueled the development of high performance automatic speech recognition (ASR) services that convert spoken language into text using machine learning (ML). Mostly developed by large technology companies, ASR systems are already integrated into everyday life for end users via speech-to-text messaging, "smart home" devices, the Internet of Things, closed captioning, voicemail transcription, and more. ASR technology holds the promise to increase productivity and digital accessibility and is marketed as such. However, ASR systems continue to be plagued by disparate performance for speakers of different languages, dialects, and accents.[1] There are also concerning findings of racial disparities with state-of-the-art commercial ASR systems, with significantly higher word error rates for Black speakers.[2] Of relevance to otolaryngology—head and neck surgery, commercial ASR systems have been found to perform less well on speech produced with dysphonic voices.[3,4]

Research on bias in Voice AI/ML models regarding the d/Deaf and hard-of-hearing (d/Dhh) community has not yet been thoroughly investigated. No study to date in otolaryngology—head and neck surgery has evaluated ASR performance on speech from individuals with hearing loss (HL). The interplay between speech production and hearing function is, however, well studied in our literature.[5–7] To follow the cultural norms of the Deaf community, we use the terminology Deaf, deaf, or hard of hearing (d/Dhh) to encompass Deaf (capital D) people who prefer to communicate with sign language and identify themselves as culturally Deaf, and deaf (lower case d) and hard of hearing individuals who have HL and may not identify as culturally Deaf.[8,9] Due to lack of or less feedback on their produced speech, d/Dhh people may produce speech of variable intelligibility.[10–12] Furthermore, there are notable features

From the Sean Parker Institute for the Voice, Weill Cornell Medical College (R.Z., A.R.), New York, New York, U.S.A.; Department of Information Science (A.S.G.C., A.K.), Cornell University, Ithaca, New York, U.S.A.

Additional supporting information may be found in the online version of this article.

Editor's Note: This Manuscript was accepted for publication on July 15, 2024.

Robin Zhao and Anna Seo Gyeong Choi are sharing first authorship.

Allison Koenecke and Anaïs Rameau are sharing last authorship.

Anaïs Rameau was supported by a Paul B. Beeson Emerging Leaders Career Development Award in Aging (K76 AG079040) from the National Institute on Aging and by the Bridge2AI award (OT2 OD032720) from the NIH Common Fund. Anaïs Rameau owns equity of Perceptron Health, Inc. Anaïs Rameau is an advisor for Savorease, Inc., and Sound Health Systems, Inc. No other disclosures were reported.

Meeting Information: 145th Annual Meeting of the American Laryngological Association in Chicago, IL, USA, May 16–18, 2024 [Oral Presentation].

Send correspondence to Anaïs Rameau, Sean Parker Institute for the Voice, 240 E 59th St, New York, NY 10022.

Email: anr2783@med.cornell.edu

Allison Koenecke, 107 Hoy Rd. Office #227, Ithaca, NY 14853.

Email: koenecke@cornell.edu

DOI: 10.1002/lary.31713

distinct to the d/Dhh population, including omission, substitution, place of articulation errors, and other voice characteristics such as harshness, breathiness, and hyper-and hypo-nasality.[13,14] ASR systems are generally trained on speech data from hearing people,[15] and a handful of studies from a single institution auditing outdated ASR systems' performance on speech from individuals from d/Dhh communities found they performed poorly. These studies were published over 5 years ago and focused on three ASR outdated systems (Microsoft Translator Speech API,[16] Presentation Translator for Microsoft PowerPoint,[17] and IBM Watson Speech to Text[18]), finding high error rates for speech from d/Dhh people, and unpredictable performance, even when the d/Dhh individual was categorized as having "good" speech intelligibility by a speech-language pathologist.[12,19]

Here, we conduct an algorithmic audit and evaluate the performance of four state-of-the-art commercial ASR systems' APIs created by Amazon, Google, Microsoft, and OpenAI on transcribed speech uttered by individuals in the d/Dhh community. We do so by quantifying the Word Error Rate (WER) and comparing it to transcribed speech from individuals with normal hearing (NH) with a pure-tone average better than 20 dB hearing level, using an open-access dataset from the Speech Perception Assessment Laboratory (SPAL) at the University of Memphis.[15] We emphasize that the d/Dhh community is not a monolith, and study API performance on d/Dhh speakers heterogeneous by speech intelligibility classification (SIC), onset of HL, and primary mode of communication. Our primary hypothesis is that state-of-the-art commercial ASR systems perform significantly worse on speech from d/Dhh speakers with lower SIC, relative to high SIC and NH speakers. Our secondary hypotheses are that ASR systems perform worse on speech from d/Dhh speakers with prelingual onset of HL (relative to postlingual onset of HL) due to lack of previous exposure to oral speech and speech development, and that ASR systems will perform worse on speech from d/Dhh speakers whose primary mode of communication includes sign language, relative to oral communication. Furthermore, we comment on the concerning performance of even the highest-performing ASR API systems.

MATERIALS AND METHODS

Data Acquisition

We used the corpus of d/Dhh speech from SPAL at the University of Memphis,[15] which provides audio recordings of US-based d/Dhh and NH speakers reading comparable passages, along with corresponding speaker demographic information (including age, gender, age of HL onset, start age of amplification use, type and model of amplification, SIC, and communication mode). While the corpus contains 850 unique audio files, we restricted files to a size of less than 25 megabyte (MB) to comply with maximum ASR input size limits for OpenAI Whisper (Table S1). The resulting 484 audio files were used in our ensuing analyses (comprising 291 d/Dhh and 193 NH audio files). On average, the d/Dhh participants each produced 12.1 audio files (one file per read passage), and the NH participants each produced 21.4 audio files. Additional details on the unrestricted corpus are detailed in the Tables S1 and S2 including transcription file size and subjects' demographic details. Table S3 shows comparable results to our main findings with and without 25 MB size limit on the transcription files. All code and transcription data are open access via our public repository: https://github.com/koenecke/ASR_dDhh_performance.

TABLE I.
Demographics ($n = 31$).

Characteristics	Participants, No. (%) ($N = 31$)	
	d/Deaf and Hard of Hearing ($N = 24$)	Normal Hearing ($N = 9$)
Age, median (IQR)	54 (47.0–62.0)	24 (23.0–25.0)
Sex assigned at birth		
Female	18 (75.0)	5 (55.6)
Male	6 (25.0)	4 (44.4)
Speech intelligibility classification		
High	4 (16.7)	
Medium	10 (41.7)	
Low	10 (41.7)	
Onset of hearing loss		
Post	14 (58.3)	
Pre	10 (41.7)	
Communication mode		
Oral	3 (12.5)	
Oral and sign	10 (41.7)	
Sign only	11 (45.8)	

Table I depicts the characteristics of the 31 General American English-speaking participants in our analysis, comprising 24 d/Dhh participants and nine NH participants. The d/Dhh group had overall poor speech production capabilities based on the Computerized Articulation and Phonology Evaluation System (CAPES), while NH participants received a normal range on their CAPES test. SIC was performed by two experienced listeners through the SPAL who rated the d/Dhh speech from 0 to 7, with 0 indicating completely unintelligible and 7 indicating extremely intelligible. The SIC scores were then classified into three categories: high (rating of 6 or 7), medium (rating of 4 or 5), and low (rating of 1, 2, or 3). In our included data, 4 d/Dhh speakers were classified as having high intelligibility, 10 with medium intelligibility, and 10 with low intelligibility. The onset of HL was categorized into two groups, prelingual and postlingual. Ten d/Dhh speakers were classified as having prelingual HL onset and 14 d/Dhh speakers with postlingual HL onset. Communication mode was categorized into three groups: oral only, oral and sign, and sign only; d/Dhh speakers predominantly consisted of individuals who sign ($n = 3$, 10, and 11, respectively). The d/Dhh participants' ages ranged from 30 to 75 years, with a median age of 54 (IQR = 47.0–62.0), and comprised 18 female and six male participants. The NH participants' ages ranged from 15 to 51 years, with a median age of 24 (IQR = 23.0–25.0) and comprised five female and four male participants. We adjusted for age and gender differences between groups in our analyses. Both groups reported good physical health with no physical, mental, cognitive, or emotional limitations. See Table S2 for demographic analysis before the exclusion of participants for d/Dhh and NH groups.

Automatic Speech Recognition Collection

This study used Python (version 3.10.9) scripts to input audio files and generate text transcriptions from four ASR APIs:

Amazon Web Services (AWS), Google Chirp, Microsoft Azure, and OpenAI Whisper. Data collection opt out was confirmed for all APIs to ensure the SPAL audio files were not used by the ASR services. All APIs were run during the same week in October 2023, and with comparable settings, that is, specifying English as the language of interest. Some ASR services failed to generate a transcription for certain audio files, which we treated as a transcription with no text, and included in our analysis. Further details are available in the Data S1.

Transcription Processing

To determine the quality of ASR-generated transcriptions, we compared them to an established ground truth transcription of what is truly being uttered in each audio file. The SPAL dataset provided such ground truth data with the text passages read by each speaker, and we secondarily reviewed these ground truth transcriptions for faithfulness to the audio recordings.

To ensure that we did not inappropriately penalize any ASRs for different treatment of similar phrases, we performed industry-standard text cleaning using a Python script applied to both the ground truth transcriptions and the ASR-generated transcription. For example, we removed filler words (such as "um" or "uh") and word fragments (such as "h-" from "hello"), and standardized spellings of comparable words (e.g., "okay" and "ok" are considered the same). Additional details of text cleaning are available in the Data S1 and Table S4.

Word Error Rate Calculation

We computed ASR performance using the WER metric, using the Python *jiwer* package (version 3.0.3). This metric compares the ASR transcription output with the ground truth transcription in terms of the number of words being inserted (I), deleted (D), and substituted (S), normalized by the total number of words in the ground truth transcript (N).

$$WER = \frac{S + D + I}{N}$$

Comparable WER values of read speech for NH populations range from 2% to 20%,[20] and prior work on d/Dhh WERs ranged from 51% to 97%.[19]

Data Analysis

We performed descriptive analysis with Welch 2-sample *t* tests, analysis of variance (ANOVA), and regression analysis. The *t* tests were used to calculate inter-platform WER differences between the d/Dhh group and NH group; ANOVA was used to compare WER differences among more granular speaker subgroups. We performed regressions and Mahalanobis distance matching[21] to calculate the effect of being in the d/Dhh subgroup on WER, accounting for speaker demographic characteristics and passage characteristics.

RESULTS

Average Word Error Rate

The results of the average WER calculation for the four APIs are shown in Table II. The average WER from the four APIs on the d/Dhh group was 52.6%. WERs by API were 45.1%, 52.3%, 55.7%, and 57.3% for OpenAI Whisper, Amazon AWS, Google Chirp, and Microsoft Azure, respectively. The average WER from the four APIs on the NH group was 5.0%, more than 10 times lower than the d/Dhh group WER (WERs by API were 3.8%, 4.3%, 5.9%, and 5.9% for Whisper, AWS, Chirp, and Azure, respectively). Each of the four ASRs yielded statistically significantly worse performance for the d/Dhh group as compared to the NH group (*t*-tests: all $p < 0.001$). Furthermore, while no statistically significant inter-platform differences were identified in pairwise comparisons of d/Dhh WER across Amazon AWS, Google Chirp, and Microsoft Azure ($0.12 < p < 0.68$), the OpenAI Whisper d/Dhh WER was found to perform statistically significantly better relative to the other three API services ($p = 0.061$, $p = 0.004$, and $p = 0.001$, respectively; Figure S1 shows WER distributions by ASR service comparing d/Dhh and NH groups). In addition, Table S3 provides a WER comparison of the four APIs with and without the 25 MB size limit. When separated by gender, the average WER for the female d/Dhh versus NH group were 55.2% versus 2.8%, 61.6% versus 3.0%, 62.4% versus 4.4%, and 66.9% versus 4.2%; and male d/Dhh versus NH group were 13.4% versus 5.2%, 23.1% versus 6.2%, 34.5% versus 7.9%, and 26.9% versus 8.3% for Whisper, AWS, Chirp, and Azure, respectively (see Table S5). These results, however, are confounded by the fact that the average age of the female d/Dhh population was older than the male d/Dhh population (53.7 years vs. 52.0 years), while the average age of the female NH population was younger than the male NH population (24.5 years vs. 29.9 years).

TABLE II.
Average Word Error Rate for Four Automatic Speech Recognition Systems.

Automatic Speech Recognition Models	d/Deaf and Hard of Hearing Group	Normal Hearing Group
OpenAI Whisper	45.1%	3.8%
Google Chirp	55.7%	5.9%
Microsoft Azure	57.3%	5.9%
Amazon AWS	52.3%	4.3%
Average	52.6%	5.0%

It remained the case that d/Dhh groups yielded higher WER across APIs when controlling for speaker characteristics. Our linear regression model (adjusting for participant age, gender, number of words in read passage, and API) estimates that being in the d/Dhh group relative to the NH group, all else equal, increases expected WER by 21.7, 26.2, 28.9, and 29.8 percentage points for OpenAI Whisper, Amazon AWS, Google Chirp, and Microsoft Azure, respectively (see Table S6). On a Mahalanobis matched subsample of 54 audio files (27 d/Dhh and 27 NH) spoken by speakers of the same gender and reading the same passage, and close in age, we observe an even larger difference in WER between d/Dhh versus NH groups (59.9% vs. 4.0%, 67.3% vs. 3.9%, 59.8% vs. 5.6%, and 71.4% vs. 5.2% for Whisper, AWS, Chirp, and Azure, respectively; see Table S7). Notably, the matched subsample of d/Dhh individuals consists of three high, 13 medium, and 11 low SIC audio files.

Subgroup Analysis

The results comparing the three subgroups are depicted in Table III, Figures 1–3. There were significant differences (ANOVA $p < 0.001$) among the speech intelligibility classification groups for all four APIs (Fig. 1). However, while the pairwise difference in WER was found to be statistically significant between low SIC and NH groups, and between medium SIC and NH groups, this was not consistently true between high SIC and NH groups. In particular, no significant difference was detected for AWS or Azure ($p = 0.986$ and 0.127, respectively); Whisper had a significantly lower WER ($p = 0.001$) while Chirp had a significantly higher WER ($p < 0.001$) for the high SIC group versus the NH group. Differences in WER distributions across SIC groups by ASR are shown in Figure S2. As per Table IV, our regression analysis (with standard errors clustered by participant) shows a statistically significant increase in WER for speakers in the low SIC ($p < 0.001$) and medium SIC group ($p = 0.002$) relative to the NH group, but a nonsignificant change ($p = 0.830$) in WER for speakers in the high SIC group relative to the NH group. Adjusting for participant age, gender, number of words in read passage and API, this linear model estimates that, all else equal, being in the low SIC group relative to the NH group increases expected WER by 64.33, 68.86, 71.50, and 72.48 percentage points for OpenAI Whisper, Amazon AWS, Google Chirp, and Microsoft Azure, respectively. For the medium SIC group, the expected increase in WER relative to the NH group is 26.50, 31.03, 33.67, and 34.65 percentage points, for the same four ASRs, respectively. For the high SIC group, there is nearly no difference in expected WER relative to the NH group, with −1.56, 2.97, 5.61, and 6.59 change in WER for the same four ASRs, respectively.

We further provide evidence that low ASR performance on the d/Dhh population stems from API inadequacy in parsing lower SIC speech; we measured this using two other subgroup variables: onset of HL and communication mode (Figs 2 and 3). In regression analysis (see Tables S8 and S9), we found significantly higher WER for speakers with prelingual HL onset relative to

TABLE III.
Average Word Error Rate by Speech Intelligibility Classification, Onset of Hearing Loss, and Communication Mode.

	Average WER Across APIs
Speech intelligibility classification	
High	9.5%
Medium	46.6%
Low	85.9%
Communication mode	
Oral	19.7%
Oral and sign	51.5%
Sign only	70.2%
Onset of hearing loss	
Postlingual	37.1%
Prelingual	80.5%

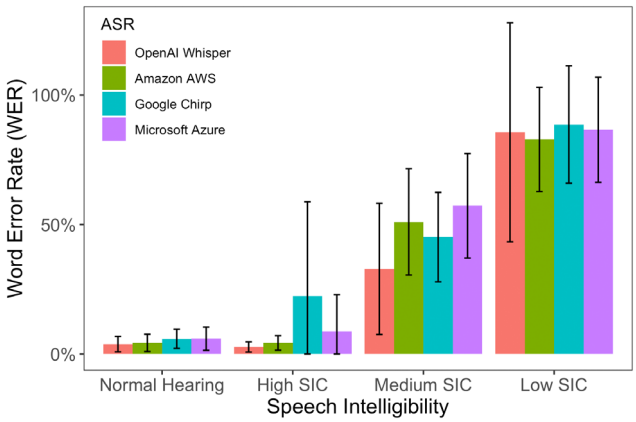

Fig. 1. Mean Word Error Rate by speech intelligibility classification. [Color figure can be viewed in the online issue, which is available at www.laryngoscope.com.]

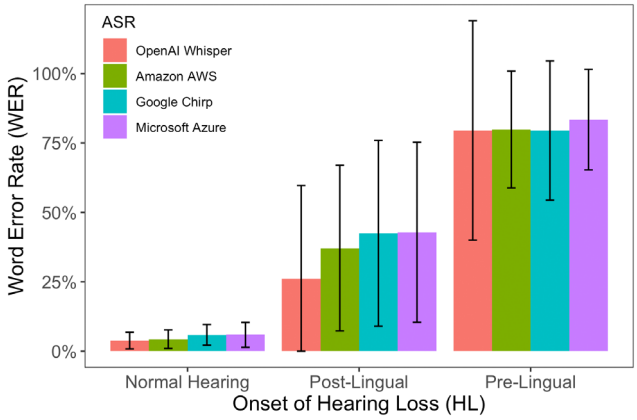

Fig. 2. Mean Word Error Rate by onset of hearing loss. [Color figure can be viewed in the online issue, which is available at www.laryngoscope.com.]

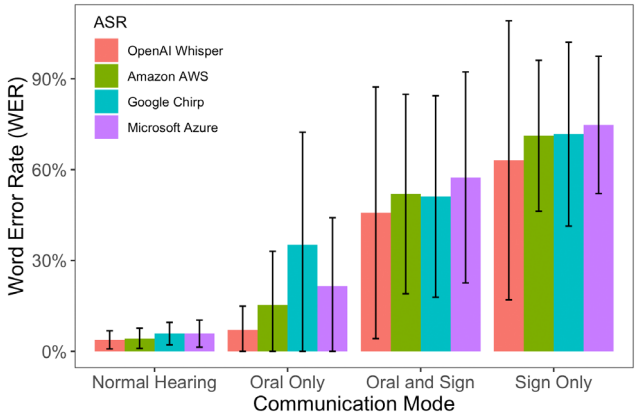

Fig. 3. Mean Word Error Rate by communication mode. [Color figure can be viewed in the online issue, which is available at www.laryngoscope.com.]

TABLE IV.
Regression Analysis of Word Error Rate by Participant Demographic and Automatic Speech Recognition With Clustered SE on Participant (Reference Levels are the Normal Hearing Group and OpenAI Whisper ASR).

Covariate	Estimate	SE	*p*-Value
Intercept	−0.1661*	0.0827	0.0447
High SIC	−0.0156	0.0724	0.8298
Medium SIC	0.265**	0.0868	0.0023
Low SIC	0.6433***	0.1485	<0.0001
Age	0.0062*	0.003	0.0387
Male	0.1841*	0.0929	0.0477
Word count in ground truth	0.0000	0.0000	0.3883
Amazon AWS ASR	0.0453*	0.0223	0.0421
Google Chirp ASR	0.0717*	0.0295	0.0152
Microsoft Azure ASR	0.0815**	0.0252	0.0012
Age * Male interaction	−0.0052	0.0028	0.0639

*significant ($p < 0.05$).
**very significant ($p < 0.01$).
***extremely significant ($p < 0.001$).

the NH group ($\beta = 48.7$, $p = 0.002$), but no statistically significant coefficient for speakers with postlingual HL onset relative to the NH group ($\beta = 15.2$, $p = 0.14$). Similarly, we found statistically significantly higher WER for speakers with sign language only as communication mode relative to the NH group ($\beta = 40.8$, $p = 0.002$), but close to significant coefficients for speakers communicating with oral and sign ($\beta = 21.8$, $p = 0.066$), and nonsignificant coefficients for speakers communicating only orally ($\beta = -3.0$, $p = 0.801$). For WER distributions by ASR service by onset of HL or communication mode, see Figures S3 and S4, respectively.

DISCUSSION

In this study, we investigated state-of-the-art commercial ASR performance on speech produced by d/Dhh individuals, revealing an average WER that was roughly 10 times higher compared to speech produced by NH individuals. This means that while roughly one out of every 20 words from NH was mistranscribed, roughly every other word spoken by d/Dhh speakers was transcribed incorrectly. In further subgroup analysis, we identified that ASR systems specifically underperformed for d/Dhh "low" and "medium" SIC groups, whereas the "high" SIC group had comparable performance to the NH for all API systems except for Google Chirp due to ASR-specific transcription errors (see Figure S2).

Among the three SIC subgroups, all four ASR systems had the lowest average WER (i.e., highest performance) for speakers in the high SIC subgroup. High speech intelligibility has been associated with cochlear implantation, especially bilateral implantation,[22–25] access to speech-language pathology services,[26,27] and family socioeconomic status.[28] There is great variability in speech intelligibility between d/Dhh individuals. Even trained professionals may have difficulty understanding d/Dhh speech,[29] and this is especially true for individuals in the medium and low speech intelligibility subgroups.

The most comparable study to ours is one conducted 5 years ago: researchers at the Rochester Institute of Technology (RIT) and the National Technical Institute for the Deaf evaluated a limited number of outdated ASR systems' performance on speech from d/Dhh individuals.[19] Specifically, Microsoft Translator Speech API, Presentation Translator for Microsoft PowerPoint, and IBM Watson Speech to Text ASRs were audited on speech data from 650 d/Dhh students reading standardized passages.[19] Similar to our results, the study found that the mean WER for the d/Dhh population was above 45% across the tested ASRs, and that d/Dhh individuals with less than the highest speech intelligibility (classified as "good" in the RIT study) had lower ASR performance compared to the NH group. In contrast, our work finds that ASR performance can be comparable between NH and high SIC speakers, which is an improvement from the previous study (wherein the "good" SIC subgroup was found to perform significantly worse than the NH group).

Notably, the aforementioned study was conducted prior to the development of OpenAI Whisper,[30] a state-of-the-art ASR system using generative AI modeling. Of the ASR systems we audited, OpenAI Whisper exhibited the lowest WER for both the d/Dhh group and the NH group, outperforming Amazon AWS, Google Chirp, and Microsoft Azure. Despite this high overall performance, there remained statistically significant differences in Whisper's WERs between d/Dhh and NH groups with WER approximately 10 times higher for the d/Dhh group in our main analysis, and 15 times higher in our demographic-matched analysis. Furthermore, the Whisper transcriptions included hallucinated text not uttered in the audio,[31] a serious text-level concern that is masked by only reporting averaged WERs, which warrants further research in improving ASR performance on d/Dhh specific speech.[32]

In the additional subgroup analyses, the group with postlingual HL onset exhibited lower WER (i.e., better performance) than the group with prelingual HL onset. The latter also exhibited statistically significantly higher WER than the NH group, consistent across all ASRs studied. This is consistent with previous research on age of HL onset, with findings that individuals with postlingual HL onset who have some prior exposure to oral speech and associated speech development tend to exhibit fewer omission, substitution, and place of articulation errors.[13,14] By modes of communication, the oral communication group exhibited the lowest average WER (i.e., best performance), followed by the oral and sign communication group, and the sign-only communication group. All three subgroups had higher average WERs than the NH group, with a strongly significant difference between the sign-only group and NH group, and a close to significant difference between the oral and sign group and NH group. These findings highlight the need to improve ASR performance on d/Dhh speech, as individuals relying on some oral communication could benefit greatly from voice-based functions when interacting with technology platforms, which are considered critical

accessibility features for those with limited literacy or with a disability limiting the ability to interact with a traditional keyboard.

Our study highlights the urgent need for technology companies—that often promote their focus on accessibility features—to include more diverse speech data in the training datasets for ASR systems.[1,33,34] ASR in the form of real-time captioning is in high demand within the d/Dhh community. For instance, one university noted a 68% increase in classroom instructions captioned from 15,440 h in 2007 to 25,978 h in 2019.[35,36] Given the growing reliance of d/Dhh individuals on ASR for receptive communication, it is critical to promote effective ASR to support expressive communication for those with oral skills as well as truly foster technological equity in this diverse population.

There are several limitations to our study. First, we were limited by the small sample size and nature of the SPAL database, potentially affecting the external validity of the findings. If anything, our WER estimates are conservative due to the read speech provided by SPAL, as spontaneous speech (not read off a script) tends to yield worse WERs.[37] Second, the Whisper API has a 25 MB size limit, so our analysis is focused on relatively smaller file sizes. Again, this means that our WER estimates are conservative since larger file sizes tended to have worse WERs (see Table S3 for Google, Amazon, and Microsoft WERs on the non-truncated 850 audio files). Future studies may focus on training ASR models using additional or upsampled d/Dhh speech data, and evaluating ASR performance on d/Dhh groups with greater socioeconomic diversity and more granular information on intelligibility levels.

CONCLUSION

This is the first study to investigate the performance of contemporary ASR systems on d/Dhh speech. Our results suggest that the ASR system APIs perform worse with d/Dhh speech than speech from the NH group, particularly when speech intelligibility is considered medium or low by human experts. Our findings suggest that while Big Tech companies advertise accessibility as a marketing strategy, their ASR systems perform poorly with a large portion of the d/Dhh population. This demonstrates the need for more advanced machine learning models trained ethically on d/Dhh audio data—in particular, on a heterogeneous set of d/Dhh speech—to uphold their promise of accessibility.

BIBLIOGRAPHY

1. Papakyriakopoulos O, Choi ASG, Thong W, Zhao D, Andrews J, Bourke R, Xiang A, Koenecke A. Augmented Datasheets for Speech Datasets and Ethical Decision-Making. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT'23). 2023;881–904.
2. Koenecke A, Nam A, Lake E, et al. Racial disparities in automated speech recognition. *Proc Natl Acad Sci*. 2020;117(14):7684-7689.
3. Rohlfing ML, Buckley DP, Piraquive J, Stepp CE, Tracy LF. Hey siri: how effective are common voice recognition systems at recognizing dysphonic voices? *Laryngoscope*. 2021;131(7):1599-1607. https://doi.org/10.1002/lary.29082.
4. Hidalgo Lopez JC, Sandeep S, Wright M, Wandell GM, Law AB. Quantifying and improving the performance of speech recognition systems on dysphonic speech. *Otolaryngol Head Neck Surg*. 2023;168(5):1130-1138.
5. Tye-Murray N, Spencer L, Gilbert-Bedia E. Relationships between speech production and speech perception skills in young cochlear-implant users. *J Acoust Soc Am*. 1995;98(5 Pt 1):2454-2460. https://doi.org/10.1121/1.413278.
6. Higgins MB, Carney AE, Schulte L. Physiological assessment of speech and voice production of adults with hearing loss. *J Speech Hear Res*. 1994;37(3):510-521. https://doi.org/10.1044/jshr.3703.510.
7. Waldstein RS. Effects of postlingual deafness on speech production: implications for the role of auditory feedback. *J Acoust Soc Am*. 1990;88(5):2099-2114. https://doi.org/10.1121/1.400107.
8. Murray JB, Klinger L, McKinnon CC. The deaf: an exploration of their participation in community life. OTJR. *Occupat Ther J Res*. 2007;27(3):113-120. https://doi.org/10.1177/153944920702700305.
9. Stebnicki JA, Coeling HV. The culture of the deaf. *J Transcult Nurs*. 1999;10(4):350-357. https://doi.org/10.1177/104365969901000413.
10. Thirumalai MS, Gayathri SG. *Speech of the Hearing Impaired*. Central Institute of Indian Languages; 2004.
11. Osberger MJ, McGarr NS. Speech production characteristics of the hearing impaired. *Speech Langu*. 1982;8:221-283. https://doi.org/10.1016/b978-0-12-608608-9.50013-9.
12. Glasser AT, Kushalnagar KR, Kushalnagar RS. Feasibility of using automatic speech recognition with voices of deaf and hard-of-hearing individuals. In Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility. 2017: 373–374.
13. Hedrick M, Bahng J, von Hapsburg D, Younger MS. Weighting of cues for fricative place of articulation perception by children wearing cochlear implants. *Int J Audiol*. 2011;50(8):540-547. https://doi.org/10.3109/14992027.2010.549515.
14. Banerjee B, Kapourchali MH, Najnin S, Mendel LL, Lee S, Patro C, and Pousson M. Inferring hearing loss from learned speech kernels. Proceedings of IEEE International Conference on Machine Learning and Applications 2016: 26–31.
15. Mendel LL, Lee S, Pousson M, et al. Corpus of deaf speech for acoustic and speech production research. *J Acoust Soc Am*. 2017;142(1):EL102-EL107. https://doi.org/10.1121/1.4994288.
16. Accessed April 14, 2024. https://translator.microsoft.com/
17. Accessed April 14, 2024. https://www.microsoft.com/en-us/translator/APPS/PRESENTATION-TRANSLATOR/
18. Accessed April 14, 2024. https://www.ibm.com/products/speech-to-text
19. Glasser, A. Automatic speech recognition services: Deaf and hard-of-hearing usability. In Extended abstracts of the 2019 CHI conference on human factors in computing systems. 2019: 1–6.
20. Chang X, Yan B, Choi K, et al. Exploring speech recognition, translation, and understanding with discrete speech units: A comparative study. ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2024 https://doi.org/10.1109/icassp48485.2024.10447929.
21. Rubin DB. Bias reduction using Mahalanobis-metric matching. *Biometrics*. 1980;36(2):293-298. https://doi.org/10.2307/2529981.
22. Chin SB, Tsai PL, Gao S. Connected speech intelligibility of children with cochlear implants and children with normal hearing. *Am J Speech Lang Pathol*. 2003;12(4):440-451. https://doi.org/10.1044/1058-0360(2003/090).
23. Calmels MN, Saliba I, Wanna G, et al. Speech perception and speech intelligibility in children after cochlear implantation. *Int J Pediatr Otorhinolaryngol*. 2004;68(3):347-351. https://doi.org/10.1016/j.ijporl.2003.11.006.
24. Litovsky RY, Parkinson A, Arcaroli J. Spatial hearing and speech intelligibility in bilateral cochlear implant users. *Ear Hear*. 2009;30(4):419-431. https://doi.org/10.1097/AUD.0b013e3181a165be.
25. Culling JF, Jelfs S, Talbert A, Grange JA, Backhouse SS. The benefit of bilateral versus unilateral cochlear implantation to speech intelligibility in noise. *Ear Hear*. 2012;33(6):673-682. https://doi.org/10.1097/AUD.0b013e3182587356.
26. Hassan SM, Hegazi M, Al-Kassaby R. The effect of intensive auditory training on auditory skills and on speech intelligibility of prelingual cochlear implanted adolescents and adults. *Egypt J Ear Nose Throat Allied Sci*. 2013;14(3):201-206.
27. Ashori M, Yazdanipour M, Pahlavani M. The effectiveness of cognitive rehabilitation program on auditory perception and verbal intelligibility of deaf children. *Am J Otolaryngol*. 2019;40(5):724-728.
28. Geers AE, Nicholas JG, Sedey AL. Language skills of children with early cochlear implantation. *Ear Hear*. 2003;24(1 Suppl):46S-58S. https://doi.org/10.1097/01.AUD.0000051689.57380.1B.
29. McGarr NS. The intelligibility of deaf speech to experienced and inexperienced listeners. *J Speech Lang Hear Res*. 1983;26(3):451-458.
30. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning. 2023: 28492–28518. https://doi.org/10.48550/arXiv.2212.04356
31. Koenecke A, Choi ASG, Mei K, Schellmann H, Sloane M. Careless Whisper: Speech-to-Text Hallucination Harms. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT'24). 2024;1672–1681.
32. Tomanek K, Tobin J, Venugopalan S, Cave R, Seaver K, Green JR, Heywood R. Large Language Models As A Proxy For Human Evaluation In Assessing The Comprehensibility Of Disordered Speech Transcription.

In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2024; 10846–10850. https://doi.org/10.1109/ICASSP48485.2024.10447177
33. Project Euphoria. Accessed April 14, 2024. https://sites.research.google/euphonia/about/
34. Cattiau J. A communication tool for people with non-standard speech. 2022. https://blog.google/outreach-initiatives/accessibility/project-relate/.
35. 2019 NTID Annual Report final. Accessed April 14, 2024. https://www.rit.edu/ntid/sites/rit.edu.ntid/files/aboutntid/annual_report_2019.pdf
36. 2007 NTID Annual Report. Accessed April 14, 2024. https://www.rit.edu/ntid/sites/rit.edu.ntid/files/aboutntid/annual_report2007.pdf
37. Furui S, Nakamura M, Ichiba T, Iwano K. Why is the recognition of spontaneous speech so hard? Text, speech and dialogue. 2005:9–22. https://doi.org/10.1007/11551874_3