Reviewing Speech Input with Audio: Differences between Blind and Sighted Users

JONGGI HONG, CHRISTINE VAING, and HERNISA KACORRI, University of Maryland, College Park
LEAH FINDLATER, University of Washington

Speech input is a primary method of interaction for blind mobile device users, yet the process of dictating and reviewing recognized text through audio only (i.e., without access to visual feedback) has received little attention. A recent study found that sighted users could identify only about half of automatic speech recognition (ASR) errors when listening to text-to-speech output of the ASR results. Blind screen reader users, in contrast, may be better able to identify ASR errors through audio due to their greater use of speech interaction and increased ability to comprehend synthesized speech. To compare the experiences of blind and sighted users with speech input and ASR errors, as well as to compare their ability to identify ASR errors through audio-only interaction, we conducted a lab study with 12 blind and 12 sighted participants. The study included a semi-structured interview portion to qualitatively understand experiences with ASR, followed by a controlled speech input task to quantitatively compare participants' ability to identify ASR errors in their dictated text. Findings revealed differences between blind and sighted participants in terms of how they use speech input and their level of concern for ASR errors (e.g., blind users were more highly concerned). In the speech input task, blind participants were able to identify only 40% of ASR errors, which, counter to our hypothesis, was not significantly different from sighted participants' performance. In depth analysis of speech input, ASR errors, and strategy of identifying ASR errors scrutinized how participants entered a text with speech input and reviewed it. Our findings indicate the need for future work on how to support blind users in confidently using speech input to generate accurate, error-free text.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**; **Empirical studies in accessibility**; **Text input**; • **Computing methodologies → Speech recognition**;

Additional Key Words and Phrases: Speech input, dictation, ASR errors, synthesized speech, text entry, blind, visual impairment

ACM Reference format:
Jonggi Hong, Christine Vaing, Hernisa Kacorri, and Leah Findlater. 2020. Reviewing Speech Input with Audio: Differences between Blind and Sighted Users. *ACM Trans. Access. Comput.* 13, 1, Article 2 (April 2020), 28 pages.
https://doi.org/10.1145/3382039

Author's addresses: J. Hong, Department of Computer Science, University of Maryland, 3173 AV Williams, College Park, MD 20742; email: jhong12@umd.edu; C. Vaing and H. Kacorri, College of Information Studies, University of Maryland, 4105 Hornbake South, College Park, MD 20742; emails: christine.vaing@gmail.com, hernisa@umd.edu; L. Findlater, University of Washington, 428 Sieg Hall, 3960 Benton Lane NE, Seattle, WA 98195; email: leahkf@uw.edu.

1936-7228/2020/04-ART2 $15.00
https://doi.org/10.1145/3382039

1 INTRODUCTION

Due to the inefficiency of using touchscreen keyboards with screen readers, people with visual impairments are using wireless keyboards or speech input as their primary means of text entry for mobile devices [3, 67]; this combination is reported as being 10 times faster compared to touchsceen text entry [6]. Speech recognition has improved substantially with advances in machine learning, yet the output from automatic speech recognizers (ASR) can still contain errors due to speaker variation, disfluency (e.g., filled pauses and revisions), ambiguity of words (e.g., homophones and proper nouns), background noise, and other mistakes from users [26, 35]. While sighted users can visually review the recognized version of their input to identify ASR errors, blind users must catch errors while listening to the synthesized speech. Correcting errors by redictation or other text entry methods (e.g., soft keyboard with a screen reader) is also a time-consuming task for blind users. For example, when blind people enter a text on a mobile device, they typically spent most of the text entry time correcting ASR errors with a soft keyboard to edit characters or move a cursor [3]. Among the two steps of ASR error correction (i.e., identification and correction), this article focuses on the first step, identifying ASR errors.

Prior work has focused on interfaces for editing text using speech input [21–23], but the challenge of detecting errors through audio-only interaction has received relatively little attention. In a study of speech-input use by blind users, Azenkot et al. [3] noted that some ASR errors were difficult to detect with a screen reader, because the errors sounded like the user's intended words (e.g., "lost my sight" vs. "lost my site"). Studying sighted users, Hong et al. [33] showed that participants missed about 50% of ASR errors when reviewing dictated text with no visual output and standard text-to-speech synthesis (at 200 words per minute). Despite the importance of accurate audio-only speech input for blind users—for example, blind users make use of speech input at higher rates than sighted users [3]—the ability of screen reader users to identify ASR errors has not been evaluated with blind participants. Though prior studies have investigated the performance of text entry methods with speech recognition with both blind and sighted participants [5, 6, 34, 41, 46, 55, 65], the ability of simply identifying ASR errors has not been researched. Previous studies evaluated the accuracy of *correcting* ASR errors, that is, the process of both identifying and correcting ASR errors. Given that individuals with visual impairments who use screen readers are known to comprehend synthesized speech better than sighted people [48], this leads to the following research questions: *How does blind and sighted individuals' experience with speech input and concerns for ASR errors differ? How well can blind screen reader users identify ASR errors when using speech input?*

In this article, we report on an exploratory user study that compares blind and sighted users' experience with speech input and interactions with ASR errors to better understand challenges and strategies in reviewing ASR errors through audio only. Study sessions included a semi-structured interview on experiences with speech dictation and synthesized text-to-speech output on different types of devices (e.g., desktop, laptop, mobile device, smart device with conversational interface), followed by a speech dictation task where participants were asked to identify ASR errors in their dictated text on a mobile device. We found that while both user groups confirmed the importance of speech input, blind participants used speech input more frequently than sighted participants (confirming results from Reference [3]). Other differences between the two groups included the most common uses of synthesized speech (reading text on a screen for blind participants *vs.* conversational interfaces such as Siri for sighted participants) and methods to review the inputted text (visual magnifier[1] or audio for blind participants vs. visual review for sighted participants).

[1]Our blind participants included two legally blind individuals who used a magnifier to read text.

During the initial interview, most participants reported that identifying ASR errors is *not* challenging, but the performance data in our study suggests otherwise. In the speech dictation task, participants in both groups were only able to identify around 40% of ASR errors in the speech dictation task, and, counter to our hypothesis, there were no significant performance differences between the two user groups. While the challenge of identifying ASR errors through audio only has been identified for sighted users [33], sighted users can choose to review important text visually when needed. *That audio-only identification of ASR errors is equally challenging for blind users with substantial synthesized speech experience—and who do not have the option for visual review—emphasizes the importance of developing speech input techniques that more accurately allow blind users to review and edit dictated text.*

We show that identifying ASR errors with audio is even more difficult for longer text, indicating that length may need to be considered when designing interfaces for reviewing dictated text. Based on the analysis of the audio recordings, we found that blind participants dictated their messages slower than sighted participants, perhaps compensating for system limitations, though this difference was not reflected in the corresponding ASR errors. Similarly, we observed that shorter words were used on average for longer messages yet more ASR errors were observed. Most importantly, we identified three distinct strategies that participants used to indicate ASR errors in the played back messages that could lead to novel interactions for reviewing ASR errors: pointing to a specific word(s), indicating the location of the errors in the message, and counting overall errors that they spotted.

The contributions of this article include: (i) comparison of blind and sighted users' experience with speech input and synthesized speech output; (ii) characterizing the task of entering a text through audio-only interactions with the analysis of speech input, ASR errors, and strategy of pointing ASR errors; and (iii) empirical evidence that blind users identify only about 40% of ASR errors and identifying ASR errors in long text is significantly more difficult than short text.

2 RELATED WORK

Speech input, automatic speech recognition, and synthesized text-to-speech have been employed in a variety of accessibility scenarios. Here, we review state-of-the-art ASR systems, applications to accessibility with a focus on people with visual impairments, studies related to the comprehension of synthesized speech, and ASR error recognition through audio only.

2.1 Automatic Speech Recognition and Error Detection

The performance of ASR systems have been improved with various techniques. The techinques can be categorized into three approaches: acoustic-phonetic approach, pattern recognition approach, and artificial intelligence approach [37, 61]. Early ASR systems were built with the acoustic-phonetic approach [64]. This approach has been particularly useful for various applications using speech, such as multilingual speech recognition, accent classification, and speech activity detection systems [49]. The pattern recognition approach had been the dominant method to build an ASR system until the artifical intelligence approach emerged with the advance of deep learning techniques. State-of-the art ASR systems employ an artifical intelligence approach using a deep neural network, reaching only a 5% word error rate (WER) recently [13]. Though prior studies achieved low error rates in restricted environments (e.g., noise-free sound, limited vocabulary, articulate speech), many factors such as speaker variation and background noise may cause ASR errors in practice [26, 35]. Therefore, researchers (e.g., Reference [19]) have been exploring techniques for automatically detecting ASR errors to supplement inherently error-prone ASR systems.

Studies on ASR systems have typically used WER as a metric to evaluate their performance [1, 61]. Accordingly, we measured the performance of the ASR in our user study with WER. In prior

studies [18, 20, 25, 59], the accuracy of automatic error detection techniques has been evaluated with precision (number of correctly detected errors divided by number of errors labeled as errors) and recall (number of correctly detected errors divided by number of actual errors). Therefore, in this study, we evaluated a participant's ability of identifying ASR errors using these same metrics.

2.2 Automatic Speech Recognition for Accessibility

People with disabilities have been early adopters of user interfaces with speech input. Speech input can allow for efficient control of a computer, home-based IPAs (e.g., Amazon Echo, Google Home), or mobile device for people with visual (e.g., References [2, 8, 52, 68]) or motor impairments (e.g., References [7, 16, 30, 42, 44]). ASR can also provide access to spoken information for people who are Deaf/deaf or hard-of-hearing (e.g., Reference [17]). For people with speech impairments, speech input has been used to support self-assessment of pronunciation (e.g., Reference [50, 56]) and to recognize a user's dictation and reproduce it through a synthesized voice (e.g., Reference [31]).

In this article, we characterize the strategies and challenges in detecting ASR errors using synthesized speech (i.e., text-to-speech) among blind and sighted users. When comparing these two user groups, prior work has shown that blind users make use of speech dictation on mobile devices more often than sighted users [2], likely due to the inefficiency of using touchscreen keyboards with a screen reader [67]. Blind users also make use of speech input to access smartphone apps [68] and to browse the web [2, 8]. In the latter cases, users can infer errors based on system response (e.g., which app opens), but, for dictation tasks, ASR errors need to be identified by listening to the text-to-speech output from the screen reader. This ASR identification task is the focus of our study.

2.3 Comprehension of Synthesized Speech

Several studies have concluded that blind people comprehend synthesized speech better than sighted people. For example, Papadopoulos and Koustriava [48] found that the comprehensibility of synthesized speech was higher for blind users, probably due to greater experience with screen readers, while natural speech was easier to understand than synthesized speech for both blind and sighted users. Similarly, Stent et al. [57] showed that participants' experience with synthesized speech positively impacts the accuracy of transcribing fast synthesized speech; they tested 300 to 500 words per minute (WPM) speech rates in users with early-onset blindness. A recent study by Bragg et al. [9] measured the accuracy of answering questions based on synthesized speech ranging from 100 to 800 WPM and they found that the maximum intelligible speech rate was higher for blind users than sighted users. Blind users have also rated the degree of understanding ultra-fast synthesized speech, at a rate of 17–22 syllables per second (680–880 WPM), higher than sighted users [45]. However, the differences between these two groups of users may disappear when there are multiple streams of speech, called the *Cocktail Party* environment [14]. In support of this, Guerreiro and Gonçalves [29] found no differences between blind and sighted users in being able to focus on a specific source when exposed to 2–4 synthesized concurrent speech sources.

While the above studies evaluated the intelligibility and comprehensibility of synthesized speech and compared performance between blind and sighted people, they focused on speech output *without* errors, which contrasts our focus on identifying ASR errors through synthesized speech.

2.4 Identifying Speech Recognition Errors

Editing text is a known bottleneck in the speech input process [38] and some studies have investigated this challenge specifically in a non-visual context. Azenkot et al. [3] found that blind users spend 80% of their time reviewing and editing text when doing a speech dictation task.

Moreover, while their study did not focus on users' ability to identify ASR errors, some examples were reported, such as where ASR errors were difficult to detect, because the recognized word was a homonym for the dictated word (e.g., "sight" and "site"). Focusing more specifically on ASR error identification, Hong et al. [33] conducted a study with sighted participants, measuring the accuracy with which users could identify ASR errors when reviewing dictated text using synthesized speech output. Participants were able to identify only around 50% of ASR errors at a standard speech rate of 200 WPM, confirming the difficulty of identifying ASR errors through synthesized speech. Hong and Findlater's study [33] also revealed that inserting pauses and slowing down the speed of synthesized speech improved the accuracy of identifying ASR errors. In comparison to these studies, we investigated the ability of blind users with screen reader experience to identify ASR errors through synthesized speech and compare their performance to that of sighted users.

Prior work has attempted to detect ASR errors automatically or to help users in identifying ASR errors. A simple approach is to visually highlight words that are grammatically incorrect, which is common in mainstream mobile devices (e.g., Reference [28]), or words that have low ASR confidence [4]. Researchers have also attempted to automatically detect ASR errors for enhancing speech-based interfaces (e.g., confirming a voice request for clarification when the system detects a potential recognition error [59]). While recent studies have developed methods to predict ASR errors using neural networks [24, 25, 59], the predictions reach 70% precision and 60% recall at best, suggesting that this is an open area of research.

The focus of our study is on *identifying* ASR errors by users in non-visual context, but a follow-on step is to correct those errors by *editing* the dictated text. With the exception of Azenkot et al. [3], already discussed, work on editing ASR results has assumed that users will visually review and edit the text. These visual editing approaches can be defined as unimodal (speech used to edit) and multi-modal (other input modalities used to edit) [58]. Multimodal solutions have combined speech with modalities such as pen, touchscreen, and keyboard input [5, 6, 34, 41, 46, 55, 65]. As an example of unimodal (speech only) correction, Choi et al. [15] developed a prediction model for distinguishing whether a user's utterance is intended to be a dictation input or a correction command, achieving 84% accuracy in offline experiments. However, unimodal interfaces suffer from cascading side effects where speech input commands for correcting errors cause further ASR errors [38].

3 METHOD

To compare blind and sighted users' experiences with speech input and their ability to identify ASR errors with only audio output, we recruited 24 participants and conducted a two-part study that included a semi-structured interview followed by a speech dictation task.

We recruited 12 blind participants (6 male, 6 female) who were screen reader users and 12 sighted participants (5 male, 7 female) from campus email lists and local organizations. Sample size was in line with typical sample sizes in this community and designed to balance research goals with practical issues of recruitment and burden on the participant community [12]. Blind participants ranged in age from 23 to 67 ($M = 49.9$, $SD = 15.1$) and sighted participants were 19 to 31 years old ($M = 22.1$, $SD = 3.9$). Blind participants reported being totally blind ($N = 6$), having some light perception ($N = 2$), or being legally blind ($N = 4$). All but two participants (one blind and one sighted) were native English speakers.[2] Background information for all participants is shown in Table 1; blind participants are denoted "B#" and sighted participants are denoted "S#."

[2]However, the two non-native English speakers (B10, S12) were not found to be outliers in terms of message length, ASR errors, or missed errors on the speech dictation task, with outliers at 1.5 times the interquartile range [60]. Thus, their data are included in the analysis.

Table 1. Participant Characteristics, with "B" Denoting Blind and "S" Sighted Participants

ID	Age	Gender	Visual impairment	Age of onset	ID	Age	Gender
B1	33	F	Total blindness	27	S1	26	F
B2	40	M	Light perception	35	S2	22	M
B3	30	F	Legally blind	23	S3	22	M
B4	65	M	Total blindness	Birth	S4	20	M
B5	52	F	Total blindness	15	S5	19	F
B6	59	F	Total blindness	1	S6	27	M
B7	63	M	Light perception	40	S7	19	F
B8	23	F	Total blindness	13	S8	19	F
B9	49	F	Total blindness	34	S9	21	M
B10	54	M	Legally blind	6 months	S10	20	F
B11	67	M	Legally blind	Birth	S11	19	F
B12	64	M	Legally blind	50	S12	31	F

All but B10 and S12 were Native English speakers; B10 and S12 had lived in the US for 30 and 27 years, respectively.

Our blind participants were all familiar with synthesized speech, since it serves as speech output for their screen readers; participants used a screen reader several times a day ($N = 11$) or several times a week (only B11). Only one participant across both groups (B12) reported some hearing loss.[3]

3.1 Procedure

Study sessions took up to 1.5 hours and were conducted in a quiet room. The whole procedure was video recorded for later analysis of participants' input in the interview and speech dictation task. The session started with a questionnaire to collect demographic information and experience with a screen reader.

Semi-structured Interview. We then conducted a semi-structured interview (~30 minutes) on prior experience with synthesized speech, speech input, and ASR errors. For the questions about ASR errors, we defined the speech recognition errors as texts recorded incorrectly by the device, because it misunderstands a word or words that the user said. Specifically, participants responded to questions about:

—frequency of use, usefulness, devices, and applications for synthesized speech output
—frequency of speech rate adjustment and reasoning behind these adjustments
—frequency of use, usefulness, devices, and applications for speech input
—maximum length for previously dictated text and reviewing practices for dictated text
—frequency of encountering and fixing ASR errors
—ASR error importance and how that relate to specific situations
—strategies for identifying and fixing ASR errors

For the two questions regarding the frequencies of using speech input or synthesized speech, frequencies were measured in an *absolute* 7-point scale adopted from Rosen et al. [54] (Never, Once a month, Several times a month, Once a week, Several times a week, Once a day, Several times a

[3]However, we did not find B12 to be an outlier in terms of message length, ASR errors, or missed errors on the speech dictation task, with outliers at 1.5 times the interquartile range. Thus, B12 were also included in the analysis.

Fig. 1. Study setup for the speech dictation task, showing researcher (left) and participant (right) perspectives. The screen was blank across all participants to control for access to visual information.

day). For example, the absolute scale was used when asking "How often do you use speech input to dictate text?" (full list in Appendix C).

Another four questions, which were relative to the frequency of using the speech input or synthesized speech, employed a *relative* 6-point scale (never, very rarely, rarely, occasionally, very frequently, always) [11]. For example, a question with the relative scale asked "How often do you encounter speech recognition errors when you dictate text?"

Speech Dictation Task. Participants then completed a speech dictation task using our custom experimental testbed built for the Apple iPhone 8 and using iOS's built-in ASR[4] and synthesized speech[5] features (included in Appendix D). A female voice with 175 word per minute (WPM) speech rate was used for the synthesized speech. The study setup is shown in Figure 1. We employed a free-form text entry task (i.e., composing the text for speech input by a participant) instead of asking participants to read reference text. The free-form text entry task is more realistic than reading reference phrases for the speech input, because people usually compose a text rather than reading a reference text when they use speech input. Moreover, the free-form text entry task allows us to recruit blind participants from general population without restrictions on Braille literacy.

The task consisted of four practice trials followed by 30 test trials. For each trial the participant composing short text or email messages in response to a series of prompted scenarios, then reviewing the recognized text to identify any ASR errors. The overall task description was as follows:

> In this task, you will be given a series of situations in which you need to compose a text message or email. For each situation, you will listen to a description with a chime sound at the beginning, then dictate a short text message or email with 1–2 sentences in response.

The 30 different prompts for the test trials were presented in random order. The test prompts (in Appendix A) were selected from a list of short scenarios ("situations") studied by Vertanen and Kristensson [62] for a freeform text composition task, such as: *"Your housemate has been sick for the last week. You are currently shopping downtown. See if he requires anything."* We asked participants to limit the dictated messages to one to two sentences so that they would remember

[4] https://developer.apple.com/documentation/speech.
[5] https://developer.apple.com/documentation/avfoundation/speech_synthesis.

their original input easily when it came time to review for ASR errors. Participants were allowed to make up names for message recipients when desired. As shown in Figure 1, the screen was blank throughout the task so that neither blind nor sighted participants received visual feedback. After completing the 30 trials with short scenarios, the testbed presented three additional trials (prompts for narrative writing from New York Times [27], available in Appendix B) with open question prompts that were intended to elicit longer descriptive answers, such as: *"You are filling out an online questionnaire about customer reviews of products. Describe how much you trust online reviews and why."* In these three trials, participants were given no length limit for their dictated messages.

At the start of each trial, the testbed played a chime sound followed by an audio recording of the prompt description. We chose to use pre-recorded audio spoken by a native English speaker for all the prompts to control for any potential effect of synthesized speech for this description on participants' ability to later identify ASR errors through synthesized speech. Participants were allowed to repeat the prompt multiple times to ensure that they understood it and were ready to dictate a response. Participants then double tapped on the iPhone screen, dictated their message, and double tapped again to end the dictation. Sound effects played to provide feedback when the system started and stopped recording (the on/off sounds used for Siri on iOS), to help participants speak only while the ASR was activated. Immediately following dictation, the text recognized by the ASR system was played using synthesized speech. After listening the synthesized speech output, participants were asked to verbally report any difference(s) between the original speech they had dictated and the text they heard via the synthesized speech output. Participants could listen to the synthesized speech of the ASR result only once in a trial to exclude the impact of listening to the ASR result multiple times on the accuracy of identifying errors Participants were not asked to correct the errors (i.e., the difference). We did not put a time limit on dictation and finding the difference to allow participants to identify errors as many as possible. Participants also reported how certain they were that they had identified all errors in the message by using a 4-point scale (very certain, certain, uncertain, very uncertain). Participants were allowed to redo the dictation for a trial once and only once if they felt they had made a mistake while speaking (e.g., stumbling over words). Within all participants and 720 trials in total, S5, B1, and B12 opted to re-dictate their input in 1, 1, and 6 trials with short scenarios, respectively. Only one of these instances (a trial of B12) occurred after the synthesized speech output had played. An additional three trials with short scenarios for B5 were redone, because the participant's accidental input caused the system to prematurely end the trial. B12 also re-dictated the input in one trial with open questions while speaking in the first attempt.[6]

Post-Study Questions. At the end of the study, we asked questions about the overall experience of reviewing the dictated message during the task. Specifically, participants reported their agreement to the following statements by using a 5-point scale (strongly agree, agree, neither agree nor disagree, disagree, strongly disagree) from Rosen et al. [54]:

- "The system correctly recognized almost everything I said."
- "It was difficult to identify errors made by the speech recognition system."

Open-ended questions were used to obtain rationale for their responses as well as feedback on strategies and challenges in identifying errors.

[6]The trials where participants re-dictated their input were not found to be outliers, with outliers at 1.5 times the interquartile range. Thus, these trials were included in the statistical analysis.

3.2 Measures and Data Analysis

The responses from the participants in the semi-structured interview and speech dictation task were transcribed from the videos of the user study and used to analyze the results. We logged the timing of speech input and the ASR results from the experimental testbed.

Semi-structured Interview. We qualitatively coded the responses of open questions using a thematic coding method to identify the major themes in the participants' responses [10]. Two researchers collaborated to code the interviews. The first researcher transcribed all of the interview data. The second researcher prepared the initial codebook based on transcription and coded the answers. The first researcher then conducted a peer review of the codebook and of randomly selected transcripts from two blind and two sighted participants. There were 10 disagreements out of 72 coded answers. The two researchers then resolved the disagreements through consensus and updated the codebook with 132 codes for 16 open questions to include two new codes about why participants were using synthesized speech and the method of reviewing text from ASR. Answers for all Likert-scale questions were analyzed with Mann-Whitney U tests, a non-parametric test that compares ordinal data from the two participant groups.

Speech Dictation Task. The speech dictation task used a mixed factorial design with a within-subjects factor of *Prompt* (short scenarios vs. open questions) and a between-subjects factor of *Vision* (blind vs. sighted). To analyze ASR errors from the speech dictation task, we manually transcribed the participants' original speech input and the verbal report of the ASR errors from the video recordings.

The differences between the manually transcribed speech input and the ASR results recorded by the experimental system were considered to be ASR errors. We defined an *error instance* as an ASR error with a word or a group of consecutive words. Error instances were coded based on their identification by participants as one of three levels of correctness:

- ***Identified***: A participant identified the specific incorrect word(s). For example, if the original input is *"Can I have the vendor's price lists?"*, the ASR result is *"Can I have the vendor's price list?"* (i.e., missing an 's'), and the participant says, *"I said lists instead of list."*, then an identified error instance is "lists." Error instances with multiple consecutive words were considered to be identified if at least *one* of those words was identified exactly, based on the assumption that users would be able to locate that error instance if they wanted to edit it.
- ***Noticed***: An error instance was noticed by the participant but was described with some ambiguity. If the participant says, "I think there was an error in there" in the above example, then "lists" is a noticed error instance.
- ***Missed***: A participant did not notice any of the misrecognized words or error instances.

Based on the coded errors, we computed precision (when a participant thinks they identified an error vs. how often it is actually an error) and recall (the proportion of error instances that participants were able to identify). We measured WER $((S + D + I)/N$ where S is the number of substituted words, D is the number of deleted words, I is the number of inserted words, N is the number of all words in the reference) of the ASR results, consistent with other studies evaluating the performance of ASR systems (e.g., Reference [13]) to see how frequently errors occurred. The length of messages was measured as the number of characters and the number of words in the original speech input. WER, recall, and precision did not violate the normality assumption (Shapiro Wilk test, $p > .05$) and were analyzed using Welch's t-tests ($\alpha = 0.05$). Message length violated the normality assumption for sighted participants (Shapiro Wilk test, $p = .023$), so we used a Mann–Whitney U test, a non-parametric alternative to the t-test, for this measure.

We examined how participants reported ASR errors during the speech dictation task from the transcribed data. The strategies of reporting errors would be potentially related to how people identify and remember the ASR errors while they are reviewing an ASR result. We found three distinct strategies that participants employed to report the ASR errors on the short scenario trials:

- ***Finding a specific word(s)***: An error instance was pointed out with the specific incorrect word(s). For example, a participant reported error by saying *"I think there was one error where it missed the word 'the'", "last word it said 'think' instead of 'thinking'."*
- ***Indicating the location***: An error instance was indicated by its location in the text. For example, a participant said *"I think the last part is messed up [...]"* in this case.
- ***Counting***: A participant counted the errors in ASR result (e.g., *"I heard two errors."*).

The strategies were not used exclusively; participants used one or more than one method in a trial. A total of 274 error instances from 2 blind and 2 sighted participants (randomly selected) were independently coded by two researchers for interrater validation. There was a substantial agreement in the level of correctness (Cohen's *kappa*[7] = 0.75) and almost perfect agreement in the strategy of reporting errors (Cohen's *kappa* = 0.83) [40]. After the validation process, one of the two researchers coded the error instances from all participants.

4 INSIGHTS FROM SEMI-STRUCTURED INTERVIEW

The main themes from the interview included experience with synthesized speech and speech input as well as strategies for detecting ASR errors on any types of devices (e.g., desktop, laptop, mobile device, smart device with conversational interface). We provided the definition of the screen reader, *"text-to-speech output, which is also called 'voice output,' 'synthesized speech,' 'Siri,' or 'Alexa,'"* to the participants at the beginning of the interview.

4.1 Experience with Synthesized Speech

While 11 of 12 blind participants reported using their screen readers several times a day, when asked about frequency of use for synthesized speech only 9 participants reported several times a day. We suspect that the other 3 participants might have not associated the term "synthesized speech" with their screen reader voice when answering this question. Still, blind participants reported using synthesized speech more frequently than sighted participants ($U = 17.5$, $p < .001$; $r = 0.60$); only two of the 12 sighted participants used synthesized speech on a daily basis (Figure 2). Participants in both groups reported using synthesized speech with a range of devices, such as a computer, smartphone, tablet, watch, TV, or smart speaker (e.g., Amazon Echo, Google Home). However, while smartphones were the most popular device for both groups, only one sighted participant used synthesized speech on a computer versus 9 of the 12 blind participants. Blind participants also primarily used synthesized speech when using screen readers ($N = 12$), whereas sighted participants used it mostly with conversational interfaces such as asking Siri a question ($N = 6$) and calling ($N = 3$).

Unsurprisingly, as indicated by prior studies (e.g., References [9, 45]), blind participants preferred faster speech rates compared to sighted participants. More than half of the blind participants ($N = 7$) preferred a speech rate setting of 51–100 (around 250–780 WPM [9]) on iOS, which is faster than the default speech rate of 50; the rest preferred the default ($N = 4$) or a slightly slow speech rate ($N = 1$). Blind participants who used faster speech than the default rate were used to listening to fast synthesized speech. Some of the participants mentioned the balance of the comprehensibility and speed. When asked about the speech rate, B3 said "*[...] I think mine is set to*

[7]Using cohen.kappa from R package "psych" [53].

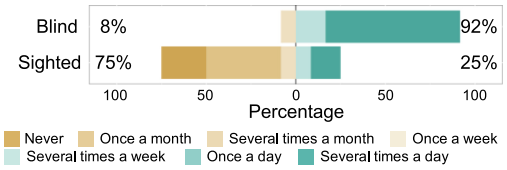

Fig. 2. Reported frequency of using synthesized speech ($N = 24$).

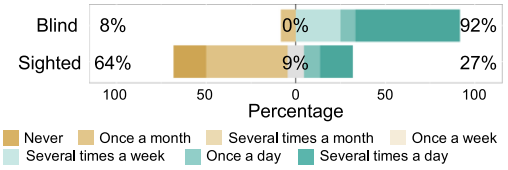

Fig. 3. Reported frequency of using speech input for dictation and voice commands ($N = 24$).

something like 57% and basically I can understand everything. If it's faster than that, I may miss some things that it says because it may sound jumbled. If it's slower than that, it may be aggravating [...]" However, sighted participants were not concerned by the speech rate, saying they did not have any preferred speech rate ($N = 7$) or that they preferred the default ($N = 5$).

Nine of 12 blind participants had experience adjusting the speed of synthesized speech, while none of the sighted participants did. Only one of the blind participants, B7, reported doing so frequently, using a fast speech rate for standard listening, but slowing it down for books or articles. Other blind participants adjusted the speech rate occasionally ($N = 4$) and very rarely ($N = 4$) for various reasons: reading certain words or content carefully (e.g., email, books, address), when letting other people use their device to get help or share contents, when getting used to a new device, and just for variety's sake. B2 said, *"If I'm working on someone else's device I would have to adjust their rate to match what my rate is [...] If I'm teaching, I would have to adjust it, so another person could understand because it may be too fast for them."*

4.2 Experience with Speech Input

Blind participants also used speech input more frequently than sighted participants ($U = 26$, $p = .006$; $r = 0.49$), as shown in Figure 3 (and confirming Reference [3]). Across both groups, participants most commonly used speech input on a smartphone compared to other devices. In terms of specific tasks, blind participants regularly used speech input for writing text for various applications ($N = 7$), such as text messages, emails, and filling out online forms while only a few sighted participants used speech input for writing text messages ($N = 4$). It was more comfortable to write texts with speech input than keyboards for blind participants who wrote text with speech input. B2 said *"Probably my main reason I mean is really just the convenience of it (speech input) so I don't have to really type anything out unless I have to more so the quickness of it."* The majority of both blind ($N = 9$) and sighted ($N = 9$) participants used conversational interfaces such as calling, asking Siri questions, opening apps, and setting timers.

Regardless of whether they regularly used speech input for dictating text, to understand differences in how speech input is being used, we asked participants to describe the length of the longest text that they had experience dictating. Of the 10 blind participants who had experience dictating text, eight had entered text longer than two sentences and four of those eight had dictated several paragraphs at a time. In contrast, only one sighted participant had dictated an entire paragraph, whereas the remaining eleven reported dictating at most one to two sentences.

4.3 Experience with Detecting ASR Errors

As seen in Figure 4, the majority of participants in both groups felt that they encountered ASR errors at least occasionally when dictating text; there was no significant difference between the two groups on this measure ($U = 73$, $p = .976$). When participants were asked an open-ended question about how concerned they were about ASR errors, the majority of blind participants expressed deep concerns about ASR errors ($N = 9$) versus only some of the sighted participants

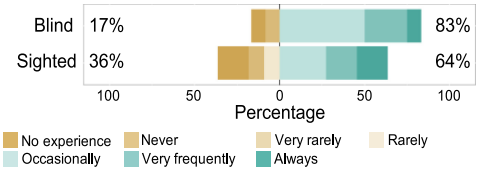

Fig. 4. Perceived frequency of encountering ASR errors when dictating text ($N = 24$).[8]

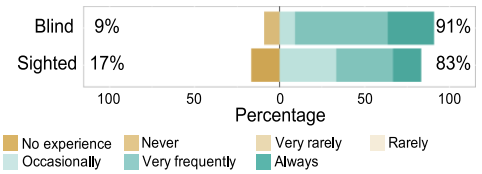

Fig. 5. Frequency with which participants reported reviewing and editing text after dictation ($N = 24$).

($N = 5$). In particular, B1 said, *"I care about them a lot because I don't want people to think that I'm stupid and I want them to understand what I'm talking about, what I'm trying to say to them,"* highlighting a previously studied misconception on the relation between spelling errors and cognitive abilities such as intelligence and logical ability [39, 63]. B10, one of the two blind participants that cared moderately, said *"To some extent. I wouldn't say I care extremely or I don't care just as much as I could have it correct."* The only blind participant, B3, who care a little said *"[I care] a little because if she can pick up 96% of what I'm saying, I'm happy with that."* No blind participant and four sighted participants reported not being concerned about ASR errors. Those participants did not necessarily feel that ASR was accurate. For example, S12 said, "*I mean I think it's a frustration but it's not a big deal. If it's an informal text it's fine [...] I wouldn't use it [speech input] to write something that's a little more important because it's not as reliable."*

As illustrated by the S12 comment, the importance of ASR errors also varied depending on the situation. To explore such use cases, as a follow up question we asked participants if there were some situations in which they were more concerned about ASR errors than others. When necessary, we further clarified this question by providing situation themes such as specific tasks, certain contents, communicating with different people, and being more rushed. Blind participants reported paying more attention when sending a message to someone in a professional relationship such as a work colleague or client ($N = 5$) or in a rushed situation to avoid wasting time in fixing ASR errors ($N = 3$). Blind participants also focused on punctuation marks, certain words that may be likely to be misrecognized by the speech recognizer (e.g., addresses, proper nouns, numbers), and content that may be hard to understand with incorrect speech recognition. B3 said, *"if you don't put a period of course it's one run-on sentence so again I guess that's user error because if I say period or comma it'll give the space."* B7 said, *"I need to speak a person's name, or a location that is something the speech recognition software is very unlikely to recognize and it's essential that the name or location be accurate."* However, sighted participants said they were most concerned about ASR errors when sending email to multiple people and when performing a voice search ($N = 6$). Some of the sighted participants ($N = 4$) mentioned rushed situations where they have limited time to review and fix ASR errors. For example, S6, said, "*If I'm more relaxed I don't really care but if I'm rushed and I need to like articulate a text message then I'm going to take the time to actually type it out.*" Like blind participants, sighted participants also mentioned concerns about ASR errors when sending a message to a person in a professional relationship compared to family or friends ($N = 6$).

The frequency of reviewing dictated text was not significantly different between blind and sighted participants ($U = 49$, $p = .168$), as shown in Figure 5. Unsurprisingly, blind participants were more likely to review dictated text via audio (synthesized speech output). Pertaining to the blind participants that reported having reviewed their dictated text ($N = 10$), the majority had used

[8]Participants in "No experience" had not entered text with speech input. Participants in "Never" had entered text with speech input, but never encountered any ASR error.

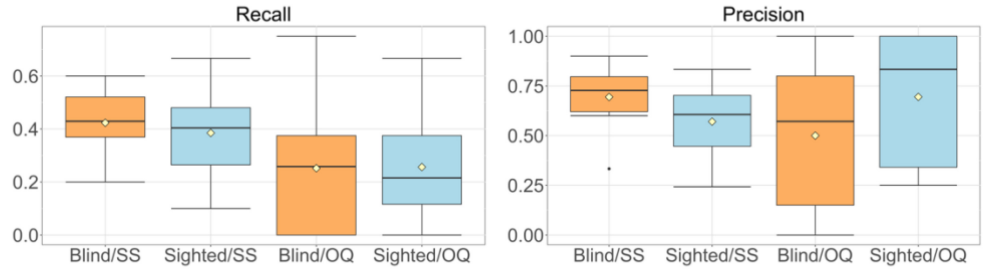

Fig. 6. Recall and precision for the blind and sighted participants in trials with short scenarios (SS) and open questions (OQ). The trials with open questions had longer messages with higher error rates.

primary audio ($N = 8$) and only (B10, B11) had used audio plus a magnifier. Of the sighted participants, only one (S9) had used audio to review dictated text and did that in conjunction with visual output by listening to the dictated text first, then visually checking if it sounded like there were ASR errors. The remaining sighted participants reported having reviewed dictated text only visually ($N = 9$) or had no experience with speech input for text entry at all ($N = 2$).

Though a prior study showed that the accuracy of identifying ASR errors by audio playback is only around 50% [33], when asked how difficult it is to identify ASR errors, participants were not aware of such challenge. All participants who had experience with reviewing dictated text thought that identifying ASR errors is not challenging. For example, B2 said, "*not challenging at all*"; S2, "*not challenging*"; S9, "*not that hard*"; and B4, "*not really challenging.*" Exceptionally, B11 pointed out that some ASR errors are not easy to detect due to the similar sounds with original input: *"you can easily hear an error, but you may not see it, you might not know it's an error. In other words, "to" and it might put two 'o's instead of one or something."* Perhaps, the rest of the participants did not realize the challenge of identifying ASR errors with synthesized speech due to difficulty in validating what they heard (for the blind participants) or due to limited experience with audio review (for the sighted participants).

5 RESULTS FROM SPEECH DICTATION TASKS

We report on WER and length of messages and our analysis of participants' performance in identifying ASR errors based on precision and recall. Our primary analysis compares blind and sighted participants in the short scenario (SS) trials. As a secondary analysis, we report on the open question (OQ) trials, comparing them to the SS trials. Given that we purposely chose to focus on the SS trials and did not counterbalance the SS and OQ trials and that there are many more SS than OQ trials, this analysis should be considered exploratory—useful for informing future research directions but not meant to be conclusive. We furthur analyzed the characteristics of speech input from the participants (i.e., speech rate and length of words) and the error instances (i.e., types of errors and the strategy of reporting errors). The analysis provides emprical findings about the patterns of entering a text using speech input and identifying errors.

The hypotheses of this task are as follows: (i) blind participants can identify the ASR errors with audio more accurately than sighted participants, and (ii) ASR error identification is harder with longer speech input.

5.1 Differences in Identifying ASR Errors

Figure 6 shows the average recall and precision of identifying ASR errors. The recall and precision data are normally distributed (Shapiro Wilk test, $p > .05$). They were analyzed using Welch's t-tests ($\alpha = 0.05$).

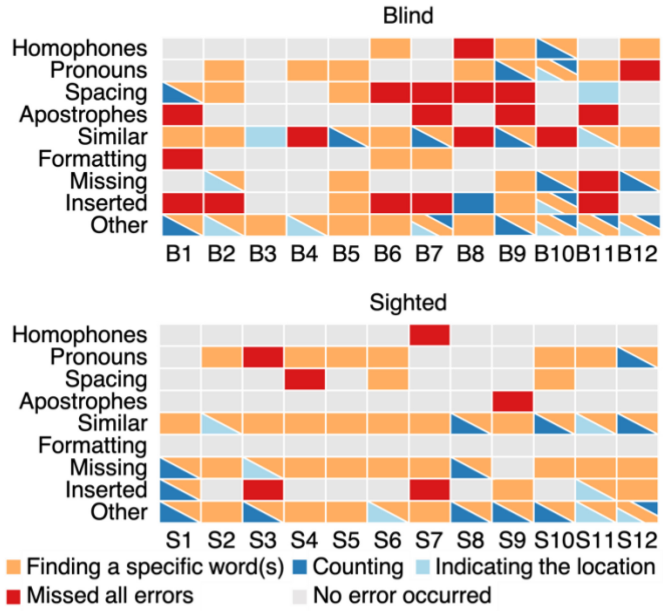

Fig. 7. The strategy used to report different types of ASR errors by blind and sighted participant. There is no strategy in a cell if no error occurred or a participant missed all errors.

Short scenario trials. While prior studies have shown that blind users comprehend synthesized speech better than sighted users [9, 45], this did not translate to a significantly improved ability to identify ASR errors through synthesized speech. Recall—the proportion of error instances correctly identified—was 0.42 ($SD = 0.13$) for blind users and 0.38 ($SD = 0.16$) for sighted users. This difference was not statistically significant ($t_{21} = -0.64$, $p = .529$). Precision—the proportion of ASR errors identified by the participants that were actually errors (not mistakes on the participant's part)—was also not significantly different across the two groups: on average 0.72 ($SD = 0.17$) for blind participants and 0.56 ($SD = 0.20$) for sighted participants ($t_{17} = -1.54$, $p = .140$).

Open question trials. Compared to the short scenario trials above, identifying ASR errors was more challenging with the three open question trials. The average recall of all 24 participants was 0.25 ($SD = 0.24$), which was significantly lower than SS trials at 0.40 ($SD = 0.15$) ($W = 42$, $p = .001$; $r = 0.45$). Specifically, the recall was 0.25 ($SD = 0.24$) and 0.26 ($SD = 0.21$) for blind and sighted participants, respectively, in OQ trials. The average precision in open question trials was 0.50 ($SD = 0.40$) for blind participants and 0.69 ($SD = 0.34$) for sighted participants. Average precision of all 24 participants in OQ trials was 0.59 ($SD = 0.37$) in OQ trials and 0.64 ($SD = 0.20$). There was no significant difference in precision between the two types of trials ($W = 91.5$, $p = .627$).

5.2 Strategies for Pointing to ASR Errors

The most common strategy, used by all blind participants in 156 trials and all sighted participants in 137 trials in both SS and OQ trials, was finding a specific word(s). Some participants counted the errors when they identify ASR errors. The eight blind and eight sighted participants reported the errors by counting the number of errors in 36 and 34 trials, respectively. Nine blind and eight sighted participants indicated the location of errors in 29 and 18 trials. Figure 7 shows that some participants (B6, S4, S5, S7, S8) tend to use the same strategies across different types of errors.

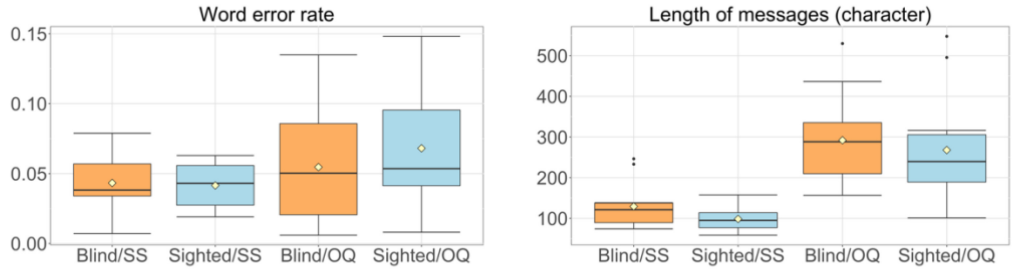

Fig. 8. WER and length of dictated messages for the blind and sighted participants in trials with short scenarios (SS) and open questions (OQ). Participants dictated longer messages in trials with OQ than SS. There was no significant difference of WER between sighted and blind participants.

The length of message and the number of errors in the ASR results may have influenced the strategies. When participants counted the ASR errors or indicated the location of errors, the length of message was 38.2 ($SD = 24.3$) and 40.6 ($SD = 27.7$) words on average, respectively. The length of message was only 30.4 ($SD = 19.2$) on average in trials where participants pointed out the specific word to report ASR errors. In trials where participants found a specific word, the ASR results had 1.75 ($SD = 1.5$) error instances on average while there were 2.4 ($SD = 1.9$) and 2.4 ($SD = 2.0$) error instances on average in trials where participants counted or indicated the location of words.

5.3 Characteristics of Dictated Messages

There are many characteristics of the dictated messages that could relate to the number of ASR errors that participants were able to indentify. To better contextualize our findings we report differences in word error rate, message length, speech rate, and word length across the recordings of blind and sighted participants as well as across short scenario and open question trials.

Word Error Rate and Length of Messages. Overall, no significant differences were found in WER or message length for the two user groups. Figure 8 shows the average WER and length of messages. While WER data did not violate the normality assumption (Shapiro Wilk test, $p > .05$), the length of dictated messages data were not normal ($p < .023$). Therefore, The WER and the length of dictated messages data were analyzed using Welch's t-tests and Mann-Whitney U test, respectively ($\alpha = 0.05$).

Short scenario trials (SS). In the 30 SS trials, the average WER of blind and sighted participants' speech input was 0.04 ($SD = 0.02$) and 0.04 ($SD = 0.02$), respectively, which is similar to the WER of state-of-the-art ASR engines [66]. We asked participants to keep their dictated messages to one to two sentences in length. Blind and sighted participants' dictated messages were 129.3 ($SD = 56.8$) and 98.5 ($SD = 29.5$) characters, which were 25.9 ($SD = 11.4$) and 19.9 ($SD = 6.1$) words, respectively; the medians were 121.3 characters (24.7 words) for blind participants and 95.1 characters (19.1 words) for sighted participants; this difference was not statistically significant (calculated in character; the mean ranks of blind and sighted participants were 14.4 and 10.6, respectively; $U = 49$, $Z = 1.33$, $p = .198$).

To examine whether the characteristics of the ASR results impacted the accuracy of identifying errors, we compared the trials with and without errors in terms of the message length and the number of errors in a trial. The average number of words was 32.7 ($SD = 20.4$) in trials with missed errors and 24.8 ($SD = 15.0$) in trials without missed errors. The average number of errors was 3.6 ($SD = 2.2$) in trials with missed errors and 2.2 ($SD = 0.8$) in trials without missed errors. The result shows that the length of message and the number of errors in the ASR result are potential

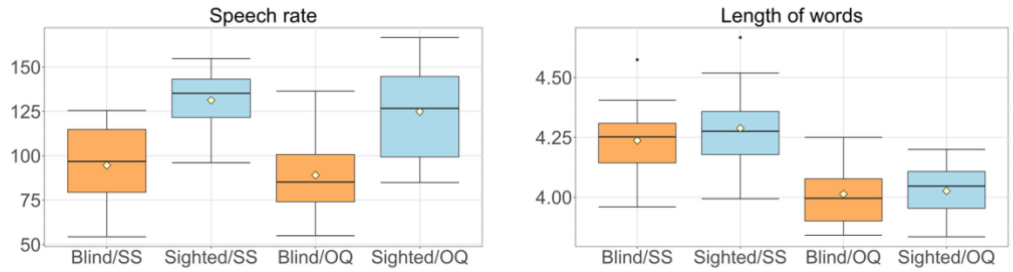

Fig. 9. Speech rate and length of words for the blind and sighted participants in trials with short scenarios (SS) and open questions (OQ). Blind participants spoke slower than sighted participants. The average length of words was shorter in OQ trials than SS trials.

factors that would affect the accuracy. That is, users would be able to identify ASR errors better in a shorter message and when there are fewer ASR errors.

Comparing SS and OQ trials. As expected based on the task instructions, the dictated messages were longer for the OQ trials than the SS trials, at on average 292.1 ($SD = 108.1$) and 268.1 ($SD = 134.4$) characters, which were 56.2 ($SD = 21.6$) and 51.3 ($SD = 25.5$) words, for blind and sighted participants, respectively. The average length of messages of all 24 participants in OQ trials was 280.1 ($SD = 120.0$) characters (53 words, $SD = 23.5$), which was longer than SS trials at 113.9 ($SD = 46.9$) characters (22.9 words, $SD = 9.4$); this difference between SS and OQ trials was significant (calculated in character; $W = 0$, $Z = -4.29$, $p < .001$; $r = 0.62$). Average WER of all 24 participants was also significantly higher in the OQ trials at 0.06 ($SD = 0.04$) than the SS trials at 0.04 ($SD = 0.02$) ($t_{23} = -2.63$, $p = .015$; $d = 0.60$).

Speech Rate and Length of Words. We analyzed the original speech input of the short scenario trials from the blind and sighted participants in terms of the speech rate and the length of words to examine if the experience with speech input influences on speech rate or complexity of words in the speech input. Figure 9 shows the speech rate and length of words. In the SS trials, blind participants spoke slower than sighted participants with 94.6 ($SD = 22.9$) and 131.3 ($SD = 17.2$) WPM speech rates, respectively ($t_{20} = 4.42$, $p < .001$, $d = 1.81$). However, we did not observe this to be reflected on the WER of the ASR. As shown in the previous section, there was no significant difference between blind and sighted participants. In the comparison between SS and OQ trials, there was no significant difference in the speech rate ($t_{45} = 0.72$, $p = .476$). The blind and sighted participants spoke at 113.0 ($SD = 27.3$) WPM speech rate in SS trials and 107.0 ($SD = 30.5$) WPM in OQ trials on average.

In the SS trials, the average length of words in the speech input was 4.2 characters ($SD = 0.2$) for blind participants and 4.3 characters ($SD = 0.2$) for sighted participants with no significant difference ($t_{21} = 0.22$, $p = .832$). However, the average length of words in SS trials for all participants—4.3 ($SD = 0.2$)—was longer than the length of words in OQ trials—4.0 ($SD = 0.1$) characters ($t_{40} = 5.06$, $p < .001$, $d = 1.46$). Considering that the OQ trials had higher WER than the SS trials, speaking shorter words would not have a positive impact on reducing ASR errors.

5.4 Error Analysis

In the SS trials, there were 340 error instances for blind participants and 236 ASR error instances in total for sighted participants. The average number of errors that participants identified, noticed, and missed (defined in Section 3.2) during a trial is shown in Table 2. Participants in both groups

Table 2. Average Number of ASR Errors That Participants Identified, Noticed, and Missed in a Trial

Vision	Task	Identified	Noticed	Missed	Total
Blind	SS	0.39	0.06	0.49	0.94
	OQ	0.58	0.72	1.69	2.47
Sighted	SS	0.23	0.08	0.35	0.66
	OQ	0.56	0.56	1.75	2.39

Table 3. Definition and the Number of Error Instances for the Types Where Error Instances Sounded Like the Original Words

Type	Description and Example	Occurred	Identified
Proper noun	Proper nouns were recognized as other words with similar sound (e.g., "Carol" and "Cara").	54	64.8% (35/54)
Spacing	The recognized words were incorrect because of the spacing (e.g., "prototype" and "proto type").	15	33.3% (5/15)
Homophones	The recognized words were homophones of the original words (e.g., "owe you" and "OU").	11	36.3% (4/11)
Apostrophes	The recognized words were incorrect because of an apostrophe mark (e.g., "doctor's" and "doctors").	5	0% (0/5)
Similar	Recognized words have similar sounds with the original words (e.g., "I'll" and "I will").	135	27.4% (37/135)
	Total	220	36.8% (81/220)

Identified column includes the proportion of identified errors (number of identified error instances/number of all error instances).

missed more than *half* of ASR error instances: *Missed* error instances represented 52.1% (SD = 11.1) and 52.1% (SD = 12.2) of all error instances on average for the blind and sighted participant groups, respectively. A further 42.3% (SD = 12.7) for blind participants and 38.5% (SD = 16.5) for sighted participants of ASR error instances were *identified*. Finally, only a small portion of errors, 5.6% (SD = 4.7) for blind participants and 9.4% (SD = 12.5) for sighted, were *noticed*.

To further understand error identification challenges, we assessed what types of ASR errors were missed in SS trials. All error instances belonged to one of the types in Tables 3 and 4, with the exception of three error instances that consisted of misrecognized words of two different types. As expected, based on past work [3], participants missed ASR errors when the errors sounded like the original words, as shown in Table 3. However, not all missed errors related to similar sounding words, as shown in Table 4. In general, the accuracy of identifying the error instances that sound similar or same with the original words were lower—36.8% than the accuracy of identifying error instances that did not sound like the original words at 44.8%. Though the sound of error instances in spacing, homophones, and apostrophes is almost same with the original words, participants could identify a few of them. For example, B6 guessed the misrecognition of "too" as "to" in a trial, saying "*I think it might've said the wrong version of too.*" Some participants picked up the small difference in synthesized speech caused by a space between words. For example, S10 distinguished "prototype" and "proto type," saying "*It just said 'prototype' like 'prawto type' do you guys care about how it says words? That's the difference.*" Participants identified the errors better when some words were missing in the recognized text than when additional words were inserted: 50.0% versus 17.6%.

Table 4. Definition and the Number of Error Instances for the Types Where Error Instances Did Not Sound Like the Original Words

Type	Description and Example	Occurred	Identified
Missing	The original word(s) was missed in the recognized text.	46	58.7% (27/46)
Inserted	The error word(s) was inserted in the recognized text though they were not spoken by the participants (e.g., recognized text of filler words).	51	17.6% (9/51)
Formatting	The recognized text has a different format (e.g., "2 o'clock PM" and "2:00 PM").	4	50% (2/4)
Others	The recognized words sound differently from the original words. We did not observe a common pattern among these errors (e.g., "we are" and "of years," "I think," and "Of think").	258	47.7% (123/258)
	Total	359	44.8% (161/359)

Identified column includes the proportion of identified errors (number of identified error instances/number of all error instances).

It is hard to identify the error instances of some types that have almost same sound with the original words (i.e., spacing, homophones, apostrophes) with audio only. To assess the ability of identifying errors that can be distinguished with audio, we measured the precision and recall of SS trials after excluding instances of spacing, homophones, and apostrophes error types. Still, there was no significant difference in precision and recall between blind and sighted participants ($t_{21} = -2.01$, $p = .056$; $t_{21} = -0.42$, $p = .677$).

5.5 Subjective Certainty

For each trial, participants were asked how certain they were that they had identified all ASR errors in their dictated text using a 4-point scale (very certain, certain, uncertain, very uncertain). In the short scenario trials, blind participants were confident, being very certain in 247 (68.6%) trials, certain in 104 (28.9%), and uncertain in only nine (2.5%). There was no trial where blind participants were very uncertain. Similarly, sighted participants were very certain in 237 (65.8%) trials, certain in 110 (30.6%), uncertain in 12 (3.3%), and very uncertain in 1 (0.3%). These numbers might not be surprising given that all but one participant, who had experience with reviewing dictated text, reported that they did not think that identifying ASR errors is challenging (Section 4.1).

One might expect that the participants would be more certain in cases where they are also able to accurately identify errors—resulting in higher precision and recall—but our results do not indicate that this was the case. While participants were confident in more than 96% of trials, the recall (very certain: 0.37; certain: 0.46) and precision (very certain: 0.67; certain: 0.61) were still low. Perhaps this could be explained by the fact that some ASR errors were difficult to detect, because the errors sounded like the participants' intended words, as in Reference [3]. Another plausible explanation could be that when interacting with a reliable ASR (with WER around 4% in our study), participants may have been less vigilant and less able to detect ASR errors when they occurred. Prior work, surveyed in Reference [43], indicates that complacency could explain why more reliable automation hurts the identification of system errors.

6 QUALITATIVE FEEDBACK

After completing the ASR dictation task, participants were still positive about the performance of ASR and their ability to identify ASR errors. When participants were asked if they agree that the system correctly recognized their input (5-point scale), 9 blind and 10 sighted participants agreed or strongly agreed; there was no significant difference between the two participant groups ($U = 76$, $p = .914$). Participants also disagreed when they were asked if it was difficult to identify ASR errors: eight blind and eight sighted participants disagreed or strongly disagreed. Again, there was no significant difference between the two groups ($U = 90$, $p = .375$).

When asked about any other difficulties they had during the task, seven blind participants reported no difficulty at all, while the remaining five blind participants mentioned challenges in remembering ASR errors in long text, checking punctuation marks, and distinguishing words with similar sounds. For example, B3 said: *"I knew there was a mistake in the beginning and the end but anything in the middle was fuzzy because these were like I said random tasks."* Contrastingly, 11 of the 12 sighted participants said they had difficulties, including remembering ASR errors in long text, imperfect pronunciation of synthesized speech, and the fast rate of synthesized speech. S12 said: *"If there are a couple little errors in larger text then you kind of lose track of them. [...] another is, is it me who's like am I creating and I saying it incorrectly or is the system picking it up incorrectly?"*

7 DISCUSSION

The semi-structured interview showed differences between blind and sighted participants with respect to their experience with speech input and error identifcation. In the speech dictation task, blind participants spoke slower than sighted participants when they used speech input. We also found that the length of the speech input impacts the accuracy of error identifcation. Furthur analysis of the errors characterized the patterns of identifying ASR errors. The emprical findings from the user study provides some insights for future research.

7.1 Implications

Our analysis of subjective responses and dicatation task data from blind and sighted participants provide insights for the design of text entry interfaces through audio-only interaction as it pertains to communicating and correcting ASR errors. Specifically, we observed the following:

Need for accessible ASR error reviewing through audio-only interactions. Our findings reinforce the importance of improving text-entry through audio only for blind users, confirming past studies that show that blind users are more likely to use speech dictation than sighted users [3, 67]. However, when it comes to reviewing their dictated text, our interview findings show sighted participants use visual output, which is only available to blind users through the text-to-speech audio. Perhaps this explains why blind participants were more concerned about ASR errors than sighted participants, given the difficulty of reviewing the ASR results through audio. For both groups, context relates to their concerns about ASR errors (i.e., kinds of tasks, content, the recipient of the dictated message, rushed or relaxed situations), suggesting that in some cases, users may be willing to use a more time-consuming but accurate reviewing process than simply hearing back their dictated message.

Mismatch between ability and perception of challenges in finding ASR errors. Neither participant group felt it was challenging to identify ASR errors by just listening to the dictated message. However, when asked to perform this task, they missed more than half of the ASR errors. This contradiction suggests that users may be making more errors than they are aware of in their dictated text motivating future work in assessing

real-world error rates in dictated and reviewed messages. Therefore, future work is needed to develop an interface enabling blind users to check the final text after revision in an efficient way rather than going through the text letter by letter.

Higher chance of missing errors with longer text. Comparing the results from the trials with short scenarios to the trials with open questions also showed that identifying ASR errors is more difficult with longer input. With longer input and higher WER in the trials with open questions, participants had to identify more errors with the long text than the short text. This would have increased the mental load of the task, requiring participants to remember more ASR errors. Since blind participants were more likely to have experience dictating longer passages of text than sighted participants, this challenge may unduly affect blind users. It is important to consider whether mechanisms to support users in reviewing and editing speech dictation via synthesized speech output need to differ for shorter versus longer passages of text, such as supporting users in reviewing only one sentence at a time.

Little impact of experience with screen reader on ability of finding ASR errors. Contrary to our hypothesis, no significant differences were found between blind and sighted participants' ability to identify ASR errors through synthesized speech. Though our interviews showed that the blind participants had more experience than sighted participants in reviewing dictated text via audio, only two blind participants who also used magnifiers had the opportunity to confirm what they heard through synthesized speech by checking visually. This lack of visual confirmation may have led them to overly trust the ASR results compared to sighted users who had, on average, substantial exposure to visual feedback of ASR results. The relatively low WERs seen in the task, though reflective of state-of-the-art automatic speech recognizers [66], may have also made it more difficult to detect a statistically significant difference between the two user groups.

Distinct strategies for reporting ASR errors can lead to novel interactions. We found three distinct strategies of identifying ASR errors by analyzing how participants reported the ASR errors during the speech dictation task. The most common strategy was finding a specific word(s) of the ASR errors. The other two strategies were indicating the location of errors and counting the errors. We found that the average length of messages was shorter and the number of errors were fewer in trials where participants found a specific word(s) than the trials where participants counted or indicated the location of errors.

The selection of strategy was potentially related to the length of message and the number of errors in the ASR result. When the text is long or the ASR result includes many ASR errors, participants would have counted or remember the location of errors rather than memorizing words to reduce the mental load. A future study on designing an accessible interface for reviewing ASR result needs to consider that the strategy of identifying ASR errors can be influenced by the length of message and the number of ASR errors.

Variation of speech input in different contexts. The analysis of speech rate provides empirical evidence that blind users speak slower than sighted users when they enter text with speech input. The difference in speech rate might have been caused by blind participants' caution to avoid the ASR errors. A prior study showed that users articulate speech more precisely when they want to enter text with speech input without errors [47]. In this case, blind participants compensate for potential limitations of the ASR system by speaking slowly.

7.2 Limitations

The speech dictation task in the user study was designed to make it realistic by employing the free-form text entry task. Though we were able to measure the ability of identifying ASR errors and characterize the use of speech input, the study also had some limitations.

Limitation of the free-form text entry. For our speech dictation task, we employed a free-form text entry task that allowed participants to compose text for themselves. Though a prior study evaluated the ability of identifying ASR errors using reference phrases [33], free-form text entry was adopted in our study because of the advantages mentioned in Section 3.1. However, using free-form text entry has a few drawbacks compared to using reference texts. Free-form text entry can result in ambiguity during error coding by the research team given that the team has only access to the spoken messages by the participant and not the ground truth text phrases. For example, some proper nouns were accurately recognized by the ASR engine (e.g., city and product names) but others were ambiguous (e.g., whether the user intend to spell the name Steven or Stephen). In these cases, if the proper noun in the synthesized speech has *one* correct spelling, then we marked it as correct. However, proper nouns (e.g., "Barbara") that were recognized as a common nouns (e.g., "barber") were considered as ASR errors. The participants dictated 13.6 proper nouns on average throughout the task.

Missing the semantic change in metrics of performance. In this work, we analyzed the performance of an ASR system for text entry through speech only, both in terms of WER and in participants' ability in identifying these errors using metrics such as recall and precision. A limitation of these metrics is that they focus on the number of error instances instead of the degree of change in the meaning of the text. For example, if "want" and "can" are recognized as "wanted" and "can't," the latter usually changes the meaning of the text more significantly than the former. However, WER, recall, and precision cannot reflect such differences in error analysis [32]. Metrics reflecting the semantic change of the original text due to the semantic differences between ASR errors (e.g., ACE metric by Kafle et al. [36]) would also be useful to examine.

Small sample size. The small number of participants in this study limits the statistical power to detect the significant difference with small effect size, though 24 participants is a common sample size in the CHI and ASSETS community. Therefore, in the analysis of precision and recall, this limitation may have resulted in no statistically significant difference between blind and sighted participants. The small number of participants also make the statistical analysis subject to change by potential outliers. Considering this limitation, we conducted another statistical analysis of the data from speech dictation task where any outliers[9] were excluded. Specifically, there were four outliers (S3, S11, B10, B11) in terms of dictated message length and one outlier (S11) in terms of precision. Removing these outliers did not change any of our results.

8 CONCLUSION

We explored the experience of speech input, synthesized speech, and ASR error identification through semi-structured interviews and evaluated the ability of identifying ASR errors through a task of entering and reviewing text using speech-only. From the semi-structured interviews, we found that sighted and blind participants' experiences differ in many aspects such as tasks,

[9]Outliers are greater than 1.5 * IQR (interquartile range) above the upper quartile or less than 1.5*IQR below the lower quartile.

devices, and frequency of using speech input, as well as employed methods for reviewing the dictated text. Though most participants reported that identifying ASR errors is not a challenging task, participants in both groups identified only around 40% of the ASR errors. This indicates that identifying ASR errors is challenging even for blind users who may have more experience with speech input and synthesized speech compared to sighted participants. We also characterized how participants identified ASR errors through analysis of the speech input, the ASR errors, and strategies for pointing to ASR errors in the speech dictation task. These findings enable us to better understand and quantify the challenges in identifying ASR errors for both sighted and blind users. More so, they reveal the need for further research on improving user interaction for speech-only text input that relies on inherently error-prone systems such as ASR.

APPENDICES

A SHORT SCENARIO PROMPTS

1. You co-worker possess the latest electronic versions of your vendor's price lists. You require these lists. Make a request to your co-worker.
2. You will be out of the office for the next week. Your co-worker James will handle any crucial issues. Inform people of the situation.
3. Your best customer is coming to visit and needs to be picked up at the airport. Make a request to the office administrator to handle the situation.
4. The prototype of the new hydraulic press will not be ready for testing for another two weeks. Inform your co-worker of the situation.
5. You are meeting with a customer on the twelfth flour of your hotel in the Golden Rose room. Give your customer directions.
6. Your phone number is 123-4567. You want to communicate with somebody in accounts payable. Leave a message.
7. You have booked a holiday next Thursday and Friday. Your boss Mary is asking for people's upcoming availability. Inform Mary of the situation.
8. You need to provide food for a meeting for ten people, three of whom are vegetarians. Tell the caterer what you need.
9. You have an all day dentist's appointment on February 28th. Inform your colleagues.
10. Your company has bought new adjustable height desks. Your knees are hitting the bottom of your new desk. Make a request to the maintenance department.
11. You have arranged to donate blood today between one and two pm. Inform your co-workers when you will not be present.
12. Michael handles procurement of supplies in your office. Your printer no longer contains magenta or cyan ink. Make a request to Michael.
13. You would like your co-worker to give a half-hour presentation at the annual company meeting on his research project. Make a request to your co-worker.
14. In the third quarter, your company's sales decreased to 2.7 million. Inform your management of the situation.
15. You can no longer recollect your computer password. Mark administers your company's computers. Ask Mark for help.
16. Your housemate has been sick for the last week. You are currently shopping downtown. See if he requires anything.
17. You want to have lunch with Laura and would also like to see her trekking photos from Nepal. Arrange to do both.

18. Your electricity bill must be paid every month. It is your housemate's responsibility this month. Inform him of the situation.
19. Your friend is picking you up at the airport but you are still waiting for your bags. Inform your friend of the situation.
20. Your friend Carol is at the local sandwich shop. Request a ham and cheese baguette.
21. Your friend is in hospital with a dislocated shoulder. Send her your sympathies.
22. Your household is trying to reduce its electric bill. Compact fluorescent light bulbs utilize significantly less energy. Make a suggestion.
23. Your friend is in trouble with the law. You have had good experience with the law firm of David and Linda. Make a suggestion.
24. You are expecting an important package to be delivered to your home. You are currently not at home but your housemate is. Make a request to your housemate.
25. You normally feed and walk your dog after work. You have to work late but your housemate is home. Make a request to your housemate.
26. You are taking your friend Barbara to your parents for dinner. She is allergic to all types of shellfish. Inform your parents of the situation.
27. Let your housemate know that the concert was sold out and you couldn't get any more tickets.
28. Your classmate wants to go out drinking tonight. You have a big Spanish exam on Monday and need to study. Inform your classmate of the situation.
29. You have just left on holiday. You are worried you did not turn the oven off. Make a request to your housemate.
30. You are building a shed behind your house and a battery powered drill is needed. Your neighbor William has such a drill. Make a request to William.

B OPEN QUESTION PROMPTS

1. You noticed one of your friend is streessed out thesedays. You want to let him or her know your method to relieve stress by email.
2. You are filling out an online questionnaire about customer reviews of products. Describe how much you trust online reviews and why.
3. You are writing a short description of a person who inspires you. Describe his or her achievement and why you are inspired by the person.

C QUESTIONS IN THE INTERVIEW

Experience with Synthesized Speech

1. How often do you use text-to-speech output? (7-point scale)

Never	Once a month	Several times a month	Once a week	Several times a week	Once a day	Several times a day

2. On what devices do you use text-to-speech, like a computer, smartphone, kindle, Amazon Echo, or other device? (Open question)
3. With what applications or tasks do you use text-to-speech output? For example, screen reader, audio books, email, etc. (Open question)
   a. Why do (or don't) you use text-to-speech with these applications or tasks? (Open question)
4. Do you have a preferred speech rate for the text-to-speech output? If so, what is it and why do you like that rate? (Open question)

5. How often do you adjust the speech rate? (6-point scale)

Never	Very rarely	Rarely	Occasionally	Very frequently	Always

   (a) [if not "never"] Please describe for what tasks and why. (Open question)

Experience with Speech Input

1. How often do you use speech input to dictate text or do other actions like opening applications, asking Siri a question, and so on? (7-point scale)

Never	Once a month	Several times a month	Once a week	Several times a week	Once a day	Several times a day

   (a) (if not "never) On what devices do you use speech input? (Open question)
   (b) For what tasks do you use speech input? Why? (Open question)
2. How useful do you find speech input to be? (Open question)
3. How often do you use speech dictation when you're not able to review the dictated text by reading it visually (or with Braille)? (6-point scale)

Never	Very rarely	Rarely	Occasionally	Very frequently	Always

   (a) (If not "never") Can you provide some examples of when this occurs? (Open question)
4 What was the longest text that you entered by speech input? How long was it? (Open question)
   1. How did you review the text? (Open question)

Experience with Detecting ASR Errors

1. How often do you encounter speech recognition errors when you dictate text? (6-point scale)

Never	Very rarely	Rarely	Occasionally	Very frequently	Always

2. How much do you care about speech recognition errors? Why? (Open question)
   (a) Are there some situations in which you care more than others (follow-up with asking for a specific example if they didn't give one for each of the following)? (Open question)
      1. What about across different kinds of tasks like email, text messages, or notes?
      2. What about based on the content or specific words that you're dictating?
      3. What about when you're communicating with different people, like family members or work colleagues?
      4. What about when you're more rushed for time or more relaxed? (follow-up with asking for a specific example if they didn't give one)
      5. Anything else?
3. How often do you review your dictated text and try to fix speech recognition errors? (6-point scale)

Never	Very rarely	Rarely	Occasionally	Very frequently	Always

   (a) Why? (Open question)

4. How do you identify or find speech recognition errors? I'm not talking about correcting those errors, just finding them. Do you have a strategy? (Open question)
   (a) How challenging is it for you to find the speech recognition errors? (Open question)
5. How do you fix speech recognition errors? Do you have a strategy? (Open question)

D ASR AND SYNTHESIZED SPEECH IN IOS

The information from the websites in footnote 2 and 3 is as follows.

Automatic Speech Recognition in iOS

Speech. Perform speech recognition on live or prerecorded audio, and receive transcriptions, alternative interpretations, and confidence levels of the results.

Overview. Use the Speech framework to recognize spoken words in recorded or live audio. The keyboard's dictation support uses speech recognition to translate audio content into text; this framework provides similar behavior, except that you can use it without the presence of the keyboard. For example, you might use speech recognition to handle recognize verbal commands or handle text dictation in other parts of your app. You can perform speech recognition in many languages, but each SFSpeechRecognizer object operates on a single language. On-device speech recognition is available for some languages, but the framework also relies on Apple's servers for speech recognition. Always assume that performing speech recognition requires a network connection.

Synthesized Speech in iOS

Speech Synthesis. Convert text to spoken audio.

Overview. The Speech Synthesis framework manages voices and speech synthesis for iOS, tvOS, and watchOS. (To perform text-to-speech tasks in macOS, use the NSSpeechSynthesizer class.) Synthesizing speech requires two main steps:

1. Create one or more AVSpeechUtterance objects containing text to be spoken. Optionally, configure speech parameters (such as voice and rate) for each utterance.
2. Pass utterances to a AVSpeechSynthesizer object to produce spoken audio. Optionally, use that object to control or respond to ongoing speech.

REFERENCES

[1] Ahmed Ali and Steve Renals. 2018. Word error rate estimation for speech recognition: E-WER. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 20–24.
[2] Vikas Ashok, Yevgen Borodin, Yury Puzis, and I. V. Ramakrishnan. 2015. Capti-speak: A speech-enabled web screen reader. In *Proceedings of the 12th Web for All Conference*, 22.
[3] Shiri Azenkot and Nicole B Lee. 2013. Exploring the use of speech input by blind people on mobile devices. In *Proceedings of the ACM SIGACCESS Conference on Computers and Accessibility*. Article No. 11.
[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] Bhakti Bhikne, Anirudha Joshi, Manjiri Joshi, Shashank Ahire, and Nimish Maravi. 2018. How much faster can you type by speaking in hindi?: Comparing keyboard-only and keyboard+ speech text entry. In *Proceedings of the 9th Indian Conference on Human Computer Interaction*. 20–28.
[6] Bhakti Bhikne, Anirudha Joshi, Manjiri Joshi, Charudatta Jadhav, and Prabodh Sakhardande. 2019. Faster and less error-prone: Supplementing an accessible keyboard with speech input. In *Proceedings of the IFIP Conference on Human-Computer Interaction*. 288–304.

[7] Jeff A. Bilmes, Xiao Li, Jonathan Malkin, Kelley Kilanski, Richard Wright, Katrin Kirchhoff, Amarnag Subramanya, Susumu Harada, James A. Landay, Patricia Dowden, and others. 2005. The vocal joystick: A voice-based human-computer interface for individuals with motor impairments. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. 995–1002.

[8] Yevgen Borodin, Jalal Mahmud, I. V. Ramakrishnan, and Amanda Stent. 2007. The hearsay non-visual web browser. In *Proceedings of the 2007 International Cross-disciplinary Conference on Web Accessibility (W4A'07)*. 128–129.

[9] Danielle Bragg, Cynthia Bennett, Katharina Reinecke, and Richard Ladner. 2018. A large inclusive study of human listening rates. In *Proceedings of the 36th Annual ACM Conference on Human Factors in Computing Systems*. 444:1–444:12. DOI: https://doi.org/10.1145/3173574.3174018

[10] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qual. Res. Psychol.* 3, 2 (2006), 77–101.

[11] Sorrel Brown. 2010. Likert scale examples for surveys. Retrieved from http://beinspired.no/wp-content/uploads/2019/04/likertscaleexamplesforsurveys.pdf.

[12] Kelly Caine. 2016. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 981–992.

[13] Zhehuai Chen, Mahaveer Jain, Yongqiang Wang, Michael L Seltzer, and Christian Fuegen. 2019. End-to-end contextual speech recognition using class language models and a token passing decoder. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'19)*. 6186–6190.

[14] E. Colin Cherry. 1953. Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am* 25, 5 (1953), 975–979.

[15] Junhwi Choi, Kyungduk Kim, Sungjin Lee, Seokhwan Kim, Donghyeon Lee, Injae Lee, and Gary Geunbae Lee. 2012. Seamless error correction interface for voice word processor. In *Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'12)*. 4973–4976.

[16] Eric Corbett and Astrid Weber. 2016. What can I say?: Addressing user experience challenges of a mobile voice user interface for accessibility. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 72–82.

[17] Stephen Cox, Michael Lincoln, Judy Tryggvason, Melanie Nakisa, Mark Wells, Marcus Tutt, and Sanja Abbott. 2002. Tessa, a system to aid communication with deaf people. In *Proceedings of the 5th International ACM Conference on Assistive Technologies*. 205–212.

[18] Rahhal Errattahi, Asmaa El Hannani, Thomas Hain, and Hassan Ouahmane. 2018. Towards a generic approach for automatic speech recognition error detection and classification. In *Proceedings of the 2018 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP'18)*. 1–6.

[19] Rahhal Errattahi, Asmaa El Hannani, and Hassan Ouahmane. 2018. Automatic speech recognition errors detection and correction: A review. *Proc. Comput. Sci.* 128 (2018), 32–37.

[20] Rahhal Errattahi, Asmaa El Hannani, Hassan Ouahmane, and Thomas Hain. 2016. Automatic speech recognition errors detection using supervised learning techniques. In *Proceedings of the 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA'16)*. 1–6.

[21] W. Feng. 1994. Using handwriting and gesture recognition to correct speech recognition errors. *Urbana* 51 (1994), 61801.

[22] Arnout R. H. Fischer, Kathleen J. Price, and Andrew Sears. 2005. Speech-based text entry for mobile handheld devices: An analysis of efficacy and error correction techniques for server-based solutions. *Int. J. Hum. Comput. Interact.* 19, 3 (2005), 279–304.

[23] Kazuki Fujiwara. 2016. Error correction of speech recognition by custom phonetic alphabet input for ultra-small devices. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 104–109.

[24] Sahar Ghannay, Nathalie Camelin, and Yannick Esteve. 2015. Which ASR errors are hard to detect. In *Proceedings of the Errors by Humans and Machines in Multimedia, Multimodal and Multilingual Data Processing (ERRARE'15) Workshop*, 11–13.

[25] Sahar Ghannay, Yannick Esteve, and Nathalie Camelin. 2015. Word embeddings combination and neural networks for robustness in asr error detection. In *Proceedings of the 2015 23rd European Signal Processing Conference (EUSIPCO'15)*. 1671–1675.

[26] Sharon Goldwater, Dan Jurafsky, and Christopher D. Manning. 2008. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase ASR error rates. In *Proceedings of the Conference of the Association for Computational Linguistics and Human Language Technologies (ACL'08/ HLT'08)*. 380–388.

[27] Michael Gonchar. 2016. 650 Prompts for narrative and personal writing. *New York Times* 20 (2016).

[28] Grammarly. Grammarly: Free Writing Assistant. Retrieved from www.grammarly.com/.

[29] João Guerreiro and Daniel Gonçalves. 2016. Scanning for digital content: How blind and sighted people perceive concurrent speech. *ACM Trans. Access. Comput.* 8, 1 (2016), 2.

[30] Susumu Harada, James A. Landay, Jonathan Malkin, Xiao Li, and Jeff A. Bilmes. 2006. The vocal joystick: Evaluation of voice-based cursor control techniques. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*. 197–204.

[31] Mark S. Hawley, Stuart P. Cunningham, Phil D. Green, Pam Enderby, Rebecca Palmer, Siddharth Sehgal, and Peter O'Neill. 2013. A voice-input voice-output communication aid for people with severe speech impairment. *IEEE Trans. Neur. Syst. Rehabil. Eng.* 21, 1 (2013), 23–31.

[32] Xiaodong He, Li Deng, and Alex Acero. 2011. Why word error rate is not a good metric for speech recognizer training for the speech translation task? In *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'11)*. 5632–5635.

[33] Jonggi Hong and Leah Findlater. 2018. Identifying speech input errors through audio-only interaction. In *Proceedings of the 36th Annual ACM Conference on Human Factors in Computing Systems*. 567:1–567:12. DOI: https://doi.org/10.1145/3173574.3174141

[34] David Huggins-Daines and Alexander I. Rudnicky. 2008. Interactive asr error correction for touchscreen devices. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*. 17–19.

[35] Hui Jiang. 2005. Confidence measures for speech recognition: A survey. *Speech Commun.* 45, 4 (2005), 455–470.

[36] Sushant Kafle and Matt Huenerfauth. 2017. Evaluating the usability of automatically generated captions for people who are deaf or hard of hearing. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'17)*. 165–174. DOI: https://doi.org/10.1145/3132525.3132542

[37] Vidyashree Kanabur, Sunil S. Harakannanavar, and Dattaprasad Torse. 2019. An extensive review of feature extraction techniques, challenges and trends in automatic speech recognition. *Int. J. Image Graph. Sign. Process.* 11, (2019), 1–12. DOI: https://doi.org/10.5815/ijigsp.2019.05.01

[38] Clare-Marie Karat, Christine Halverson, Daniel Horn, and John Karat. 1999. Patterns of entry and correction in large vocabulary continuous speech recognition systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 568–575.

[39] David S. Kreiner, Summer D. Schnakenberg, Angela G. Green, Michael J. Costello, and Anis F. McClin. 2002. Effects of spelling errors on the perception of writers. *J. Gen. Psychol* 129, 1 (2002), 5–17.

[40] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174. DOI: https://doi.org/10.2307/2529310

[41] Yuan Liang, Koji Iwano, and Koichi Shinoda. 2014. Simple gesture-based error correction interface for smartphone speech recognition. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'14)*. 1194–1198.

[42] Bill Manaris, Renée McCauley, and Valanne MacGyvers. 2001. An intelligent interface for keyboard and mouse control. In *Proceedings of the 14th International Florida AI Research Symposium (FLAIRS-01)*. 182–188.

[43] Sara E. McBride., Wendy A. Rogers, and Arthur D. Fisk. 2014. Understanding human management of automation errors. *Theor. Issues Ergon. Sci.* 15, 6 (2014), 545–577.

[44] Yoshiyuki Mihara, Etsuya Shibayama, and Shin Takahashi. 2005. The migratory cursor: Accurate speech-based cursor movement by moving multiple ghost cursors using non-verbal vocalizations. In *Proceedings of the 7th International ACM SIGACCESS Conference on Computers and Accessibility*. 76–83.

[45] Anja Moos and Jürgen Trouvain. 2007. Comprehension of ultra-fast speech–blind vs."normally hearing" persons. In *Proceedings of the 16th International Congress of Phonetic Sciences*. 677–680.

[46] Jun Ogata and Masataka Goto. 2005. Speech repair: Quick error correction just by using selection operation for speech input interfaces. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'05)*. 133–136.

[47] Sharon Oviatt, Margaret MacEachern, and Gina-Anne Levow. 1998. Predicting hyperarticulate speech during human-computer error resolution. *Speech Commun.* 24, 2 (1998), 87–110.

[48] Konstantinos Papadopoulos and Eleni Koustriava. 2015. Comprehension of synthetic and natural speech: Differences among sighted and visually impaired young adults. In *Proceedings of the International Conference on Enabling Access for Persons with Visual Impairment (ICEAPVI'15)*. 147–151.

[49] Bhagath Parabattina and Pradip Das. 2004. Acoustic phonetic approach for speech recognition: A review. *Language* 77 (2004), 93.

[50] Thomas Pellegrini, Lionel Fontan, Julie Mauclair, Jérôme Farinas, Charlotte Alazard-Guiu, Marina Robert, and Peggy Gatignol. 2015. Automatic assessment of speech capability loss in disordered speech. *ACM Trans. Access. Comput* 6, 3 (2015), 8.

[51] Marcelo Philip. 2017. Technology seeks to preserve fading skill: Braille literacy. Retrieved from https://apnews.com/7c9cf97fbdde47a3a262985ecfc2c564/Technology-seeks-to-preserve-fading-skill:-Braille-literacy.

[52] 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*. 459.
[53] William Revelle. 2018. Psych: Procedures for Psychological, Psychometric, and Personality Research. Retrieved from https://cran.r-project.org/package=psych.
[54] Larry D. Rosen, Kelly Whaling, L. Mark Carrier, Nancy A. Cheever, and J. Rokkum. 2013. The media and technology usage and attitudes scale: An empirical investigation. *Comput. Hum. Behav.* 29, 6 (2013), 2501–2511.
[55] Sherry Ruan, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James A. Landay. 2018. Comparing speech and keyboard text entry for short messages in two languages on touchscreen phones. *Proc. ACM Interact. Mobile Wear. Ubiq. Technol.* 1, 4 (2018), 159.
[56] Oscar Saz, Shou-Chun Yin, Eduardo Lleida, Richard Rose, Carlos Vaquero, and William R. Rodriguez. 2009. Tools and technologies for computer-aided speech and language therapy. *Speech Commun.* 51, 10 (2009), 948–967.
[57] Amanda Stent, Ann Syrdal, and Taniya Mishra. 2011. On the intelligibility of fast synthesized speech for individuals with early-onset blindness. In *Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*. 211–218.
[58] Bernhard Suhm, Brad Myers, and Alex Waibel. 2001. Multimodal error correction for speech user interfaces. *ACM Trans. Comput. Interact.* 8, 1 (2001), 60–98.
[59] Yik-Cheung Tam, Yun Lei, Jing Zheng, and Wen Wang. 2014. ASR error detection using recurrent neural network language model and complementary ASR. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'14)*. 2312–2316.
[60] Graham Upton and Ian Cook. 1996. *Understanding Statistics*. Oxford University Press.
[61] Arul Valiyavalappil Haridas, Ramalatha Marimuthu, and Vaazi Gangadharan Sivakumar. 2018. A critical review and analysis on techniques of speech recognition: The road ahead. *Int. J. Knowl.-based Intell. Eng. Syst.* 22, 1 (2018), 39–57. DOI: https://doi.org/10.3233/KES-180374
[62] Keith Vertanen and Per Ola Kristensson. 2014. Complementing text entry evaluations with a composition task. *ACM Trans. Comput. Interact.* 21, 2 (2014), 8.
[63] Jane A. Vignovic and Lori Foster Thompson. 2010. Computer-mediated cross-cultural collaboration: Attributing communication errors to the person versus the situation. *J. Appl. Psychol.* 95, 2 (2010), 265.
[64] D. Walker. 1975. The SRI speech understanding system. *IEEE Trans. Acoust.* 23, 5 (1975), 397–416.
[65] Lijuan Wang, Tao Hu, Peng Liu, and Frank K Soong. 2008. Efficient handwriting correction of speech recognition errors with template constrained posterior (TCP). In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'08)*. 2659–2662.
[66] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2017. The microsoft 2016 conversational speech recognition system. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17)*. 5255–5259.
[67] Hanlu Ye, Meethu Malu, Uran Oh, and Leah Findlater. 2014. Current and future mobile and wearable device use by people with visual impairments. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computer Systems (CHI'14)*. 3123–3132. DOI: https://doi.org/10.1145/2556288.2557085
[68] Yu Zhong, T. V. Raman, Casey Burkhardt, Fadi Biadsy, and Jeffrey P. Bigham. 2014. JustSpeak: Enabling universal voice control on android. In *Proceedings of the 11th Web for All Conference*. 36.

Received February 2019; revised December 2019; accepted February 2020