Identifying Speech Input Errors Through Audio-Only Interaction

Jonggi Hong
Inclusive Design Lab | HCIL
University of Maryland, College Park
jhong12@umd.edu

Leah Findlater
Human Centered Design and Engineering
University of Washington
leahkf@uw.edu

ABSTRACT
Speech has become an increasingly common means of text input, from smartphones and smartwatches to voice-based intelligent personal assistants. However, reviewing the recognized text to identify and correct errors is a challenge when no visual feedback is available. In this paper, we first quantify and describe the speech recognition errors that users are prone to miss, and investigate how to better support this error identification task by manipulating pauses between words, speech rate, and speech repetition. To achieve these goals, we conducted a series of four studies. Study 1, an in-lab study, showed that participants missed identifying over 50% of speech recognition errors when listening to audio output of the recognized text. Building on this result, Studies 2 to 4 were conducted using an online crowdsourcing platform and showed that adding a pause between words improves error identification compared to no pause, the ability to identify errors degrades with higher speech rates (300 WPM), and repeating the speech output does not improve error identification. We derive implications for the design of audio-only speech dictation.

Author Keywords
Speech dictation; error correction; synthesized speech; text entry; eyes-free use; audio-only interaction.

ACM Classification Keywords
H.5.2. Information interfaces and presentation (e.g., HCI): User Interfaces.

INTRODUCTION
The past few years have yielded vast improvements in speech recognition due to advances in machine learning. Speech input is now faster and more accurate than mobile touchscreen keyboards [21], and is the primary means of text input on devices that have a small or no visual display, such as smartwatches or voice-based intelligent personal assistants (e.g., Google Home, Amazon Echo). Speech input is also particularly useful during eyes-free interaction (e.g., walking or driving) or as accessible input for blind users [1,32].

Reviewing and editing the inputted text, however, is a bottleneck [12]. Speech recognition errors can arise from several sources: the ambiguity of words (*e.g.*, homophones and pronouns), background noise, and mistakes from users [10]. When visual output is available, users can read the recognized text and more easily identify these errors than when they can only hear an audio synthesis of the same text. Error detection and correction techniques that rely on visual output have been proposed and studied for desktops and mobile devices (e.g., [4,5,6]). However, the challenge of identifying and correcting speech recognition errors is greatly magnified when both the text input and the review to identify errors is done through audio only. Azenkot et al. [1] showed that while speech is a primary text input method for blind users, 80% of the time is spent reviewing and correcting errors using screenreader (audio) output.

In this paper, we quantify the problem of identifying speech recognition errors through audio-only feedback and investigate potential solutions. While researchers have examined understanding of and ability to transcribe text-to-speech output (e.g., [16,22]), the impacts of different text-to-speech manipulations on the user's ability to identify speech recognition *errors* have not been investigated.

We report on a series of four controlled studies. The goal of the first study was to characterize the problem of identifying errors based on audio-only output. For this in-lab study, native and non-native English speakers dictated and listened to the recognized version of a series of phrases in silent and noisy conditions. Overall, participants were unable to identify more than 50% of recognition errors when listening to audio of the recognized text, with the most common difficulty being with multiple-word errors (e.g., "mean" to "me in", or "storm redoubles" to "stormy doubles"). Studies 2 through 4 then investigated the effect of three text-to-speech manipulations—pauses between words, speech rate, and speech repetition—on the user's ability to identify those recognition errors. Inserting pauses, in particular, could help to address the multiple-word errors identified in Study 1. Studies 2 and 3 showed that adding a pause between words resulted in significantly higher error identification rates than no pause, and that fast speech (i.e., 300 WPM) made identification more difficult. Finally, Study 4 evaluated another alternative—repeating the audio

CHI 2018, April 21–26, 2018, Montreal, QC, Canada

ACM ISBN 978-1-4503-5620-6/18/04 $15.00
https://doi.org/10.1145/3173574.3174141

output twice—and found that repetition did not improve participants' ability to identify errors over simply listening to the audio once.

The contributions of this paper include: (i) identifying the types of speech recognition errors that are likely to be missed by users interacting only through audio; (ii) evaluating the effects of synthesized speech with different speech rates and pauses between words on this error identification; (iii) evaluating the effect of listening to synthesized speech twice on error identification.

RELATED WORK

Speech interfaces are increasingly common with the advance of deep learning techniques such as recurrent neural networks [33]. Here, we focus on the comprehension of synthesized speech and the correction of dictated text.

Comprehension of Synthesized Speech

Many studies have examined the impacts of speech and user characteristics on comprehension of synthesized speech. For example, synthesized speech rates of 150-200 WPM are the most comfortable for younger and older adults [24]. Naturally produced speech has traditionally been easier to understand than synthesized speech (e.g., [29]). However, linearly time-compressed synthesized speech results in faster cognitive processing time compared to fast, naturally produced speech, due to the less careful articulation of the naturally produced speech [9]. Another approach to speeding up listening is to generate a summary of the synthesized speech, removing some words. This summarized synthesized speech has been found to be more comprehensible than time-compressed synthesized speech when the lengths of the synthesized speech from both compression methods are identical [26].

User characteristics can also impact comprehension levels. Native English speakers have higher comprehension than non-native speakers when listening to synthesized speech [7], although comprehension decreases for both native and non-native speakers as the speech rate increases (from 155 to 178 WPM) [11]. Moreover, while sighted and visually impaired users are both better able to comprehend natural speech than synthesized speech, visually impaired users fare better than the sighted users with synthesized speech, perhaps due to greater experience with it (e.g., from audio-based screenreaders) [20]. In a related study, Stent *et al*. [22] measured the intelligibility of fast synthesized speech for users with early-onset blindness, testing rates of 300 to 500 WPM. As with slower speeds, transcription accuracy decreased as speech rate increased, and participants' experience with synthesized speech impacted accuracy. Blind users are also better than sighted users at understanding ultra-fast synthesized speech, at a rate of 17-22 syllables per second, or around 680-880 WPM [16].

While the above studies examine how to design synthesized speech output for intelligibility, comprehensibility, and naturalness, they do not examine the user's ability to detect when there are *errors* in the speech output. This error identification task could be impacted differently by factors such as speech rate and elision (i.e., dropping the beginning or ending syllables of words during speech).

Correcting Dictated Speech

Correcting speech recognition errors requires first reviewing the recognized text to identify errors, then performing edits. The focus of our study is on the former, but because a common assumption is that users will *visually* review the recognized text, more research effort has been expended on the latter, editing step. In general, methods to edit recognized text can be categorized into two types: unimodal and multi-modal [23]. Many studies have examined multi-modal methods, combining speech input with other modalities such as touchscreen, pen gestures, or spelling out individual letters [4,5,6]. Some have combined speech with touchscreen gestures or keyboard input for editing text [4,5]. Automatically predicting alternative word candidates when the user identifies a recognition error has also been used, though typically these approaches have assumed touchscreen or keyboard input [8,14,17,27]. Fujiwara developed a custom phonetic alphabet to enter words accurately by spelling, which the authors considered to be a separate modality from speech (whole-word) input [6]. The above examples, however, have all relied on visual output, and by definition have included input other than speech, as opposed to our audio-only focus.

Unimodal methods, in contrast, use only speech input to enter the main text *and* to correct words, and can be combined with visual or non-visual output. Unimodal correction, however, suffers from cascading side effects, where the speech input commands used to correct recognition errors result in further recognition errors [12]. Addressing this problem (though still employing visual output), Choi *et al*. developed a system that attempts to automatically predict whether speech input is intended for main text dictation or for correction, showing in offline experiments that their classifier is able to make this prediction with 83% accuracy [3].

Compared to the above approaches, which employ visual output for the recognized text, audio-only interaction where speech is used for dictation *and* for error detection has received little attention. Highlighting the need for further work, Azenkot and Lee [1] showed that speech input is popular for blind users on mobile devices, but that reviewing and editing the dictated text is a substantial challenge. Our work further explores and begins to address this problem of audio-only error detection and correction.

STUDY 1: UNDERSTANDING AUDIO-BASED ERROR IDENTIFICATION IN RECOGNIZED SPEECH

Though previous studies have shown that reviewing dictated text using non-visual output is a challenge [1], the extent of that challenge and the specific difficulties that users encounter have not been quantitatively assessed. How many misrecognized words do users miss when reviewing

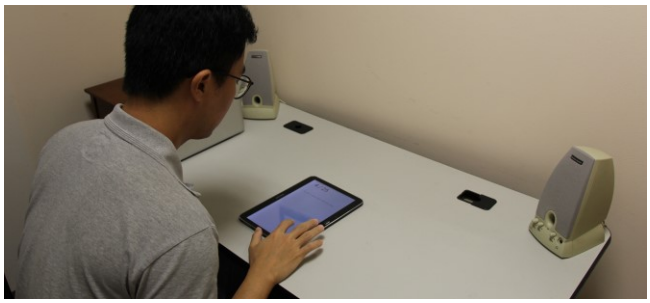

Figure 1. The experimental setup of Study 1.

only through audio? What kind of errors are hardest for users to identify? To answer these questions, we conducted a lab-based study where participants dictated a set of phrases using a mobile device, and reviewed the system's recognition of each phrase by listening to audio output. To increase generalizability of the findings, we manipulated the level of background noise and participants' fluency levels, two factors that are known to impact speech recognition accuracy [2].

Method

This controlled experiment measured the impacts of background noise level and the user's English proficiency, on the word error rate (WER) of the speech recognizer and on the user's ability to identify recognition errors based on text-to-speech output.

Participants

We recruited 12 native English speakers (5 male and 7 female) and 12 non-native English speakers (8 male and 4 female) through campus email lists. The native English speakers ranged in age from 18 to 36 (M=23.4, SD=5.3), while the non-native English speakers were 22 to 38 years old (M=26.3, SD=4.4). None reported having hearing loss. Non-native speakers had lived in United States for 0.3 years on average (SD=2.3). Ten native speakers and eight non-native speakers had experience with speech input before, while the remaining participants did not.

Procedure

Study sessions took 30 minutes and were conducted in a quiet room. As shown in Figure 1, participants sat at a table on which a Galaxy Tab 4 and two speakers were placed. We first collected demographic information and experience with using speech input. The silent and noisy conditions were then presented in counterbalanced order. The tablet's audio output was set to 75% of maximum volume, which was approximately 60db with the synthesized speech audio. For the noisy condition, the speakers played street noise (audio from [25]) at ~50db. A custom Android application guided participants through 30 trials per condition, where each trial consisted of: (1) reading a phrase displayed on the tablet screen, (2) dictating the phrase, which included double tapping the screen to indicate the start and end of dictation, (3) listening to synthesized speech output of the recognized phrase, and (4) identifying discrepancies, if any, between the dictated and recognized text. This lattermost step involved reporting words that had been incorrectly recognized and locations where extra words were inserted. Participants viewed the reference phrase while listening to the synthesized speech, and verbally reported errors they heard to the experimenter.

The phrases were randomly selected without replacement from a set with 200 phrases extracted from the LibriSpeech ASR corpus [19]. Of the 2703 phrases in the Librispeech development subset, 600 had 10 or fewer words, of which we randomly selected 200 that were of a complete sentence form, comprehensible, and contained no proper nouns which would increase ASR errors. The IBM Speech-to-Text API[1] was used for speech recognition because it provides functions to analyze the speech recognition results (*e.g.*, confidence scores and timing of words). The speech was synthesized on the tablet device using the TextToSpeech[2] function in Android 5.0 with the default speech rate of 175 WPM (which is within the range recommended in the research literature as well [24]).

Study Design

This study used a mixed factorial design with a within-subjects factor of *Noise* (silent vs. noisy) and a between-subjects factor of *Fluency* (native vs. non-native). The silent and noisy conditions were presented in counterbalanced order. Participants were randomly assigned to orders.

Measures and Data Analysis

To provide a baseline understanding of how well the speech recognizer performed, we computed word error rate (WER) on the recognized text [15]; lower rates are better. To assess the user's ability to identify errors, we computed *precision*—that is, when a participant thinks they hear an error, how often is it actually an error—and *recall*—that is, how the proportion of true errors participants were able to identify. We also employed *phrase-level accuracy* as a secondary measure, that is, whether a participant identified at least one error in a phrase that contains one or more errors, or no errors in a correct phrase. For this exploratory study, we focused on accuracy and did not measure speed.

To compute these measures, we needed to judge whether each instance where the participant pointed out an incorrectly recognized or inserted word was a true positive, or that the lack of an error label was a true negative. Ambiguity arose when a single word was recognized as multiple words (e.g., "meet" to "me it"). Is this (i) one "incorrect word – meet" or (ii) one "incorrect word – meet" plus one "inserted word – it"? We considered both responses to be correct, with (i) counted as a true positive and (ii) counted as two true positives. As a third case, if the participant marked this error as simply one "inserted word – it", we judged the response to include one false negative (the word "meet" should have been marked as incorrect) and one true positive (for the word "it" being added).

[1] https://www.ibm.com/watson/services/speech-to-text/

[2] https://developer.android.com/reference/android/speech/tts/TextToSpeech.html

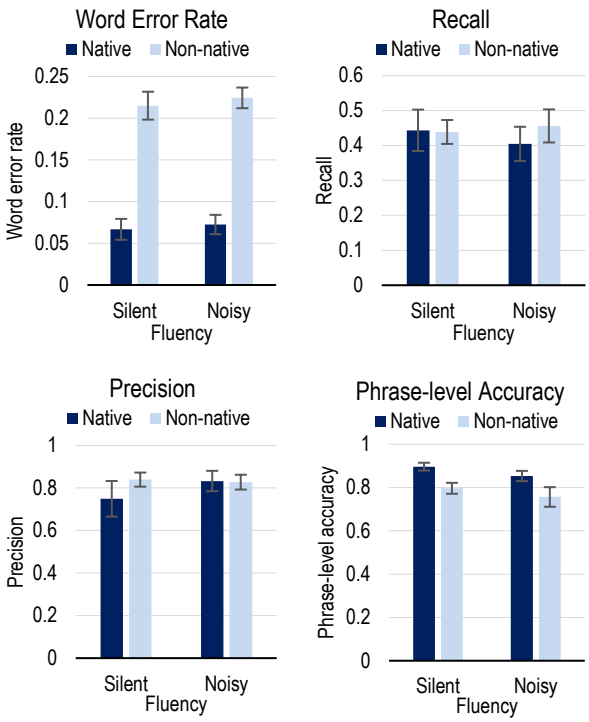

Figure 2. WER, precision, recall, and phrase-level accuracy in Study 1. Recall results showed that participants missed identifying more than *half* of the speech recognition errors. Error bars show standard error (*N*=12 per group).

WER and precision violated the normality assumption of an ANOVA (Shapiro Wilk tests, $p<.05$), so we instead used 2-way repeated measures ANOVAs with aligned rank transform (ART) for these measures, a non-parametric alternative to a factorial ANOVA [30]. Recall was analyzed in the same way, but without the ART adjustment.

Result

Figure 2 shows the WER, precision, recall, and phrase-level accuracy in silent and noisy conditions with native and non-native speakers.

Fluency affected speech recognition accuracy. Impacts of fluency and noise on WER have been previously studied, so our intention in including these factors in the experience was simply to increase the generalizability of our main measures (i.e., precision and recall in identifying recognition errors). For completeness, however, we still examined whether fluency and noise impacted WER. As expected based on past work [28], fluency did impact WER. WER was higher with non-native speakers than native speakers, at 0.22 (*SD*=0.05) compared to 0.07 (*SD*=0.04); this difference was significant (main effect of *Fluency*: $F_{1,22}=123.00$, $p<.001$, $\eta^2=0.74$). The WER with native speakers was close to typical WERs achieved by recent speech recognition engines at 0.05-0.10 WER [31]. Different levels of background noise did not significantly impact WER (main effect of *Noise*: $F_{1,22}=0.79$, $p=.384$, $\eta^2<0.01$), nor was there a significant interaction effect between *Fluency* and *Noise* ($F_{1,22}=0.16$, $p=.693$, $\eta^2<0.01$).

Participants missed more than half of the errors. In terms of participants' ability to identify the speech recognition errors based on audio output, across all conditions, precision was 0.81 (*SD*=0.18), meaning that 19% of the errors that participants marked were not true errors. Of greater importance for being able to produce accurate text input, however, are the relatively low recall rates: on average across all four conditions, only 0.44 (*SD*=0.16) of true errors were identified—more than half the errors were undetected. Phrase-level accuracy, which could allow a user to at least know they should re-dictate an entire phrase even if they are not aware of all detailed errors, was higher, at 0.90 (*SD*=0.06) in the best case (native speakers + silent).

ANOVA (with ART if applicable) results revealed no significant main or interaction effects of *Fluency* or *Noise* on precision or recall. There was a significant main effect of *Fluency* on phrase-level accuracy ($F_{1,22}=7.48$, $p=.009$, $\eta^2=0.14$), whereby native speakers had higher accuracy than non-native speakers, at 0.85 (*SD*=0.08) compared to 0.76 (*SD*=0.12). However, the main effect of *Noise* and the interaction effect between *Fluency* and *Noise* on phrase-level accuracy were not significant.

Multiple-word errors most difficult to identify. To better understand what types of errors participants had trouble identifying, we qualitatively analyzed the 183 errors that native speaker participants missed (i.e., instances of false negatives). Native speakers who are most likely to use speech input in English were target participants in Studies 2-4, so we focused on native speakers in this analysis. One research team member coded the missed errors into the categories below. For validation, a second coder also independently coded all missed errors, and Cohen's kappa showed strong interrater agreement (*kappa*=0.82, 95% CI: [0.76, 0.88]). The categorizations were as follows:

- *Multiple-word errors (N=107; 58.5%).* Multiple sequential words sometimes sounded like another word or words. We included cases where multiple words were recognized as a single word (e.g., 'a while' and 'awhile'), multiple words were recognized as other multiple words (e.g., 'storm redoubles' and 'stormy doubles'), and single words were recognized as multiple words (e.g., 'meet' and 'me it').
- *Single word errors (N=57; 31.1%).* This type of error includes single words that were replaced with homophones or other single words with similar sounds (*e.g.* 'inquire' and 'acquire', 'he' and 'she').
- *Punctuation mark errors (N=7; 3.8%).* There is typically no explicit indication of punctuation marks such as apostrophes in text-to-speech output. If the recognized word is exactly the same as the intended word except for a punctuation mark, we classified it as a punctuation error (e.g., 'state's' and 'states').

- *Other (N=12; 6.6%).* In some cases, the type of error was unclear. For example, when there were many errors in a phrase the participant may simply have been unable to remember them all.

Summary and Discussion

Across both user groups, participants missed over 50% of recognition errors when listening to the audio playback. Phrase-level accuracy, which would allow a participant to know they should re-dictate an entire phrase, was higher but still left many unidentified errors (10% of phrases). The majority of the errors that participants did not notice were classified as multiple-word errors. A potential solution to address this type of error is to emphasize the individual words in the text-to-speech output by adding pauses between words—an approach that we focus on in Studies 2-4 alongside other simple output manipulations. That there were no differences in WER or participants' ability to identify recognition errors between different background noise levels suggests that we may have needed a wider range of noise levels to properly assess that factor.

STUDY 2: EFFECT OF SPEECH RATE AND PAUSE ON ERROR IDENTIFICATION

Study 1 showed that participants missed a substantial number of errors when listening to the confirmation audio clips, with the most common type of missed error being a multiple-word error. In Study 2, we focused on a straightforward potential means of addressing this problem: adding artificial pauses between words in the speech output, which should allow the user more easily distinguish individual words. Inserting pauses in synthesized speech affects prosody and elision—the latter being when successive words are strung together while speaking, causing omission of an initial or final sound in a word. While this change is not ideal for many uses of text-to-speech, it is potentially useful for helping users to correct recognition errors with audio-only interaction.

This study isolated the error identification component of speech input and correction. Fifty-four crowdsourced read a series of phrases that had been dictated in Study 1, listened to corresponding confirmation audio clips (i.e., text-to-speech output of what the system had recognized), and identified discrepancies (recognition errors) between the presented text and the audio output under varying conditions: no/short/long pause and three speech rates.

Method

For this study and the two subsequent ones, we recruited crowdsourced participants on Amazon's Mechanical Turk to be able to run a series of studies with a larger and more diverse sample than would have been feasible in the lab.

Participants

The 54 participants (33 male, 21 female) ranged in age from 21 to 58 (M=33.4, SD=8.8). All participants were native English speakers, and none reported hearing loss. Just over half (N=29) had experience with speech input. All participants reported completing the study in a quiet room.

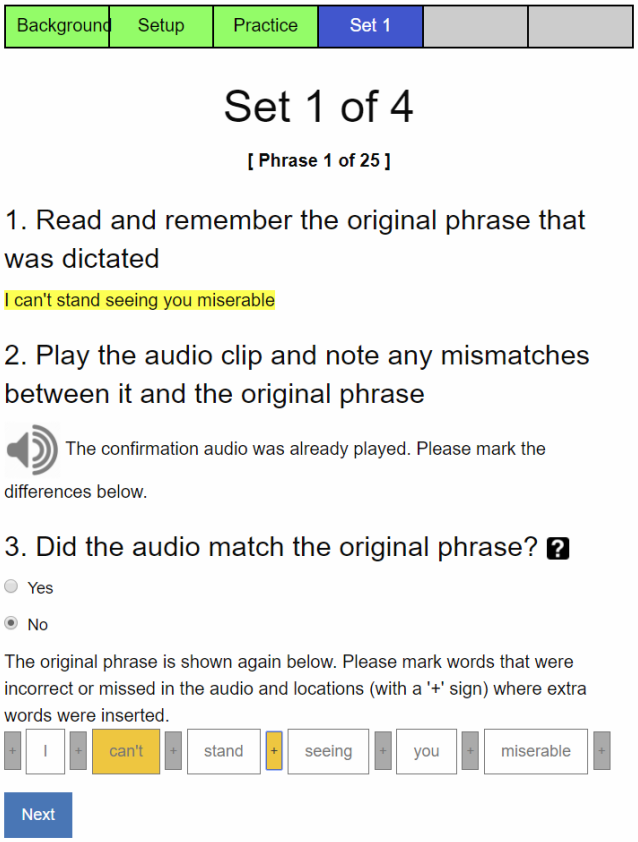

Figure 3. Screenshot of the online testbed used for Studies 2, 3, and 4, showing a single trial. A trial consisted of reading a presented phrase, listening to an audio clip of what a speech recognition engine had heard, and marking errors in the recognized version (i.e., discrepancies between text and audio).

Procedure

Participants were directed to an online testbed that guided them through the 45-minute study procedure. The procedure began with a background questionnaire, followed by instructions about the overall tasks. Participants were then shown a sample phrase and asked to adjust the sound volume to ensure they could easily hear the audio clip.

The task consisted of identifying discrepancies between presented text phrases and audio clips, where the audio clips may contain errors made by a speech recognizer. To test realistic speech recognition errors, the pairings of presented phrases and audio clips were taken from the speech input collected during Study 1. That study resulted in 600 pairs of presented and recognized phrases, where 32.4% of the recognized phrases included at least one error. The *Say*[3] app in Mac OS X was used to generate the synthesized speech, including pauses.

Figure 3 shows an example trial, with the presented phrase and an audio clip widget. After clicking to listen to the audio clip once (a single time; no replays allowed), the

[3]https://developer.apple.com/legacy/library/documentation/Darwin/Reference/ManPages/man1/say.1.html

participant answered (yes/no) whether the audio clip had matched the presented phrase. The page included boxes that mapped to each word in the presented phrase as well as locations before and after words where extra words could appear. Participants marked all discrepancies between the presented text and the audio by clicking the corresponding boxes. The boxes were only enabled after the audio clip finished playing, so participants could not mark errors while actively listening to the audio. The presented phrase was visible for the duration of the trial. The 'next' button was enabled only after the participant had reported whether the audio contained any errors.

Participants first completed six practice trials to familiarize themselves with the task. Practice trials used a typical text-to-speech output setting of 180 WPM and no pauses between words. After each practice trial, participants were shown the correct answer as feedback. The experimental conditions were then presented in counterbalanced order, with 20 test trials per condition. Phrases were randomly selected from the set of 600 with no replacement, and different phrases were used for practice and test trials. After finishing all conditions, participants had to answer questions about easiness and preference of conditions.

Study Design

Study 2 used a 3x3 within-subjects design with factors of *Speech Rate* (100, 200, and 300 WPM) and *Pause Length* (no pause, 1ms, and 150ms). Order or presentation for the nine conditions was counterbalanced using a balanced Latin square (in fact, two squares due to having an odd number of conditions). Participants were randomly assigned to orders.

The 1ms pauses, while too short to cause a detectable silence in the output, were used to eliminate elision in contrast to the 'no pause' condition. The 150ms pause length was selected based on pilot testing different lengths (1 to 200ms) to identify a short, yet distinguishable pause. Because the effectiveness of pause lengths and error identification in general may be impacted by the speech rate, we included three speech rates: one close to default rates in commercial text-to-speech systems (200 WPM), a slower rate (100 WPM) and a faster rate (300 WPM).

Measures and Data Analysis

As in Study 1, we computed precision, recall, and phrase-level accuracy of identifying errors in the audio clips. Two participants were excluded from analysis because they did not mark any words as errors in one condition, making it impossible to calculate precision. Although our focus is on how *well* participants identify errors, for completeness we also report on trial completion time (time from the start of a trial to clicking the 'next' button). However, low trial completion times are not necessarily our goal, since they could be due to not noticing and thus not taking the time to mark errors. Perhaps more importantly, the length of the audio clips varies by condition, so we also report on descriptive statistics for audio clip length.

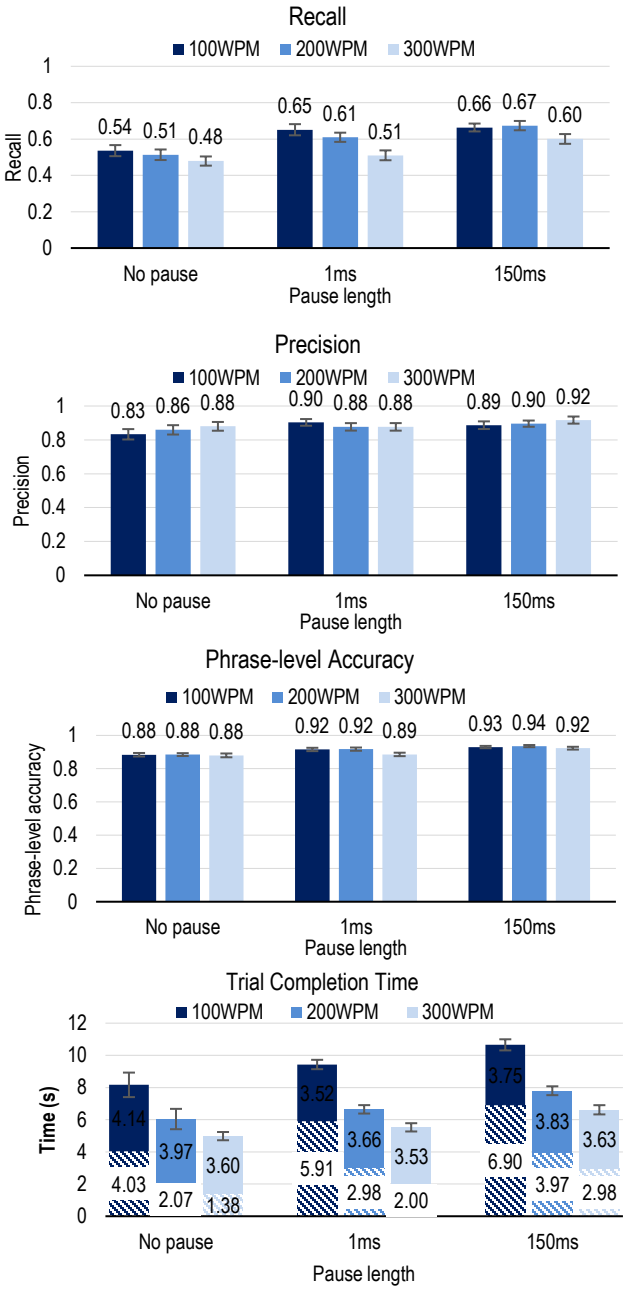

Figure 4. Precision, recall, phrase-level accuracy, and trial completion time in Study 2. The shaded portion in trial completion time indicates the average length of audio clips in that condition. Participants identified errors most accurately with the 200 WPM speech rate and 150ms pause. Error bars show the standard error (*N*=52)

Precision, phrase-level accuracy, and trial completion time violated the normality assumption of ANOVA (Shapiro-Wilk tests, $p<.05$). Therefore, 2-way repeated measures ANOVAs with ART were used, with Wilcoxon signed rank tests and a Bonferroni correction for posthoc pairwise comparisons. For recall, a 2-way RM ANOVA was used with paired t-tests for posthoc pairwise comparisons.

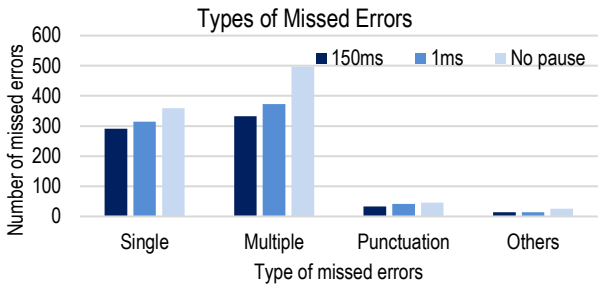

Figure 5. Types of errors participants missed identifying in Study 2. Participants missed 33% fewer multiple-word errors with a 150ms pause compared to no pause.

Result

Figure 4 shows our primary measures of precision, recall, and phrase-level accuracy, along with trial completion time for completeness.

Pauses and slower speech improve recall. Recall ranged from 0.48 to 0.67 across the nine condition. *Pause Length* significantly impacted recall ($F_{2,408}$=1.47, p<.001, η^2=0.07). All posthoc pairwise comparisons were significant (p<.05), showing that as pause length increased, so did recall. *Speech Rate* also significantly affected recall ($F_{2,408}$=0.66, p<.001, η^2=0.03). Posthoc pairwise comparisons showed that the 300WPM speech rate resulted in significantly lower recall than the other two speeds (both comparisons p<.05). The interaction between *Speech Rate* and *Pause Length* was not significant ($F_{4,408}$=0.89, p=.467, η^2<0.01).

Pauses also impact precision. Precision ranged from 0.83 to 0.92 across the nine conditions. Precision was significantly impacted by *Pause Length* ($F_{2,408}$ = 3.71, p=.025, η^2=0.01), although after a Bonferroni correction no posthoc pairwise comparisons were significant. There was no significant main effect of *Speech Rate* on precision ($F_{2,408}$=1.22, p=.297, η^2<0.01), nor was the *Pause Length* x *Speech Rate* interaction effect significant ($F_{2,408}$=1.53, p=.193, η^2<0.01).

Secondarily, pauses improve phrase-level accuracy. Overall, *Pause Length* significantly impacted phrase-level accuracy ($F_{2,408}$=22.36, p<.001, η^2=0.07), with posthoc pairwise comparisons showed that the differences between all pairs of pause lengths were significant (all p<.05). *Speech Rate* also significantly impacted phrase-level accuracy ($F_{2,408}$=6.46, p<.001, η^2=0.01), but no posthoc pairwise comparisons were significant after a Bonferroni correction. The interaction effect between *Speech Rate* and *Pause Length* was not significant ($F_{4,408}$=2.31, p=.058, η^2<0.01).

Trial completion times and audio lengths as expected. The length of time to play the audio clip consisted of a substantial portion of the trial completion time on average, as shown in Figure 4. The downside of inserting pauses between words and slowing down speech playback is that these changes lengthen the audio clip time. Accordingly, there was wide variation in both trial completion times and audio clip length. Even the 1ms pause added 10-15% to trial completion times across the three speech rates compared to *no pause*, and 44-47% if just examining the length of the audio clips, because the pauses eliminate overlaps between words (eliminating elision).

	Speech rate (WPM)			Pause length (ms)		
	100	200	300	No	1	150
Ease	21	27	4	17	21	14
Preference	7	31	14	20	22	10

Table 1. Subjective vote tallies in Study 2. The 200 WPM speech rate and shortest two pause lengths were the most preferred, while 300 WPM was least likely to be voted easiest.

Identifying multiple-word errors improved the most. To examine the effect of pauses on specific types of errors, we manually coded 2341 missed errors from all participants. Figure 5 shows the number of errors of all types. The overall trend shows that all three types of errors decreased as the pause increased. However, the most substantial reduction was for multiple-word errors, which dropped 33.2% from the no pause condition (497 missed errors) to the 150ms pause condition (332 errors). In contrast, missed single-word errors only dropped 18.9%, from 359 to 291, and punctuation errors dropped 28.2%, from 46 to 33.

Speech rate impacted perceived ease and preference. The subjective responses differed from the objective measures. Table 1 shows vote tallies for easiest and most preferred speech rates and pause lengths. Pearson Chi-Square test of independence showed that *Speech Rate* significantly impacted ease ($X^2_{(2,N=52)}$=16.42, p<.001) and preference votes ($X^2_{(2,\ N=52)}$=17.58, p<.001). The 200 WPM speech rate received the most votes for both ease and preference. *Pause Length* did not significantly impact either measure. In open-ended comments, participants said that 200 WPM felt natural because it was close to normal speech rate. While the accuracy with 150ms was highest, nine participants felt that it sounded unnatural compared to the other two pause lengths. Four participants reported that the 1ms pause, however, gave a moment to think as well as being more natural than the 150ms pause.

Summary and Discussion

Recall and phrase-level accuracy were highest with the longest pause length (150ms), while the fastest speech rate (300 WPM) negatively affected recall. An important consideration, however, is that inserting pauses and slowing down speech increases audio clip length and thus overall task time. Compared to the baseline condition (i.e., 200 WPM, no pause), the best combination (200 WPM, 150ms pause) resulted in a 31% increase in recall and a 4% increase in precision, though also almost doubled the playback length. Even the 1ms pause made the audio 0.6-1.9s longer than no pause audio because it removed the elision in the phrase. In terms of subjective responses, most participants preferred the 200 WPM speech rate (which corresponds to [24]) and felt that 300 WPM made the task harder. However, there was no impact of pause length on

subjective measures, suggesting that these short pauses (1ms, 150ms) may be acceptable compared to no pause even though they add time to the task.

STUDY 3: DETAILED IMPACT OF PAUSE LENGTH

Study 2 showed that inserting pauses between words in the speech output enables users to identify errors more accurately, but only included two pause lengths that were greater than 0ms. Because adding pauses increases overall task time, we would ideally be able to pinpoint the shortest pause length that is still effective, and use that during audio-only speech input. To more precisely identify an ideal pause length than was possible in Study 2, here we evaluate seven pause lengths ranging from 1ms to 300ms.

Method

The study method is similar to Study 2 with the exceptions described here. The speech rate was fixed at 200 WPM because there were no significant error identification differences between 100 and 200 WPM in Study 2, but participants preferred 200 WPM. We recruited 42 participants (23 male, 19 female). Participants were on average 37.2 years old (*SD*=11.7; range 21-68). All were native English speakers and none had hearing loss. Twenty had previously used speech input. Four participants reported completing the study with light background noise (e.g., light street noise or office), while the remaining 38 participants reported using a quiet room.

This study employed a within-subjects design with the single factor of *Pause Length* (1, 50, 100, 150, 200, 250, or 300ms). This range spans from imperceptible pauses to highly obvious pauses. The seven conditions were presented in counterbalanced order using a balanced Latin square, similar to Study 2. Participants were randomly assigned to orders. Precision, recall, phrase-level accuracy, and trial completion time all violated the normality assumption of ANOVA (Shapiro-Wilk tests, $p<.05$), so 2-way RM ANOVAs with ART were used. Two participants who marked no errors in one condition were excluded from analysis because their precision could not be calculated.

Result

Figure 6 shows results for the four main measures. Unlike in Study 2, there were no significant main effects of *Pause Length* on recall, precision, or phrase-level accuracy (respectively: $F_{6,234}=2.12$, $p=.052$, $\eta^2=0.04$; $F_{6,234}=0.68$, $p=.667$, $\eta^2=0.02$; $F_{6,234}=2.09$, $p=.06$, $\eta^2=0.04$). Average audio clip length ranged from 3.0s per trial with the 1ms pause to 5.0s per trial with the 300ms pause. The trial completion time was shortest with 1ms pause at 6.4s and longest at 8.1s for both the 250ms and 300ms pauses. Following the performance results, there were no statistically significant difference in easiness and preference due to *Pause Length* (Chi-square tests, $p>.05$).

Discussion and Further Data Collection

These results are unexpected and appear to contradict Study 2, where we had concluded that the 150ms pauses resulted in significantly higher recall and phrase-level accuracy than the 1ms pause. (Note that the worst-performing condition from Study 2 – *no pause* – is not included in this study.)

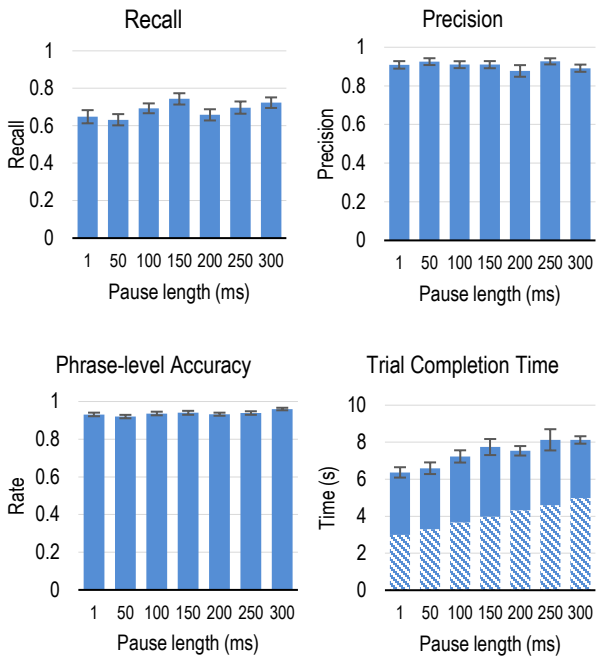

Figure 6. Graphs of precision, recall, phrase-level accuracy, and trial completion time in Study 3. The shaded portion in trial completion time indicates the length of audio clips. There were no significant differences in accuracy measures due to pause length. Error bars show standard error (*N*=40).

To confirm that the result of Study 3 was not obtained by chance and to better understand this unexpected result, we conducted two additional studies, which we report on briefly. First, we approximately replicated Study 3, but with 30 participants and two adjustments to increase statistical power: only four pause length conditions (1, 75, 150, and 225ms), and 40 trials per condition instead of 20. This replication yielded a similar result to what is reported above: no significant effects of pause length on recall, precision, and phrase-level accuracy. A subsequent closer examination of the Study 2 results, however, revealed that an important yet not statistically significant interaction effect may have affected those earlier conclusions: the 1ms vs. 150ms pause difference may have arisen primarily from the 300 WPM speech rate condition, rather than the 100 WPM or 200 WPM conditions. As such, because we used only 200 WPM in Study 3, we revisited the 200 WPM data from Study 2. A simple paired t-test showed that there was no significant difference between the 1ms and 150ms pause for recall; similarly, Wilcoxon signed rank tests were not significant for precision or phrase-level accuracy. As such, Study 3 does confirm Study 2, but also provides more nuance on the conclusions.

Again, the worst-performing pause length from Study 2 was the *no pause* condition, which allowed us to conclude that

inserting even 1ms pauses was better than no pause. To confirm that this conclusion still held for a 200 WPM speech rate alone, we first conducted a t-test and Wilcoxon signed rank tests on the 200 WPM data from Study 2. The 1ms pause resulted in significantly higher recall and phrase-level accuracy than *no pause* (all $p < .05$). We then conducted a short follow-up replication: we collected new data from 28 participants who completed 25 trials in two conditions: 200 WPM with a 1ms pause and 200 WPM with no pause. The 1ms pause resulted in significantly higher recall (t-test, t_{27}=2.73, p=.011, d=0.59) and phrase-level accuracy (Wilcoxon signed rank test, W=216.5, Z=2.43, p=.014, r=0.32) than no pause.

Considering the results from both Studies 2 and 3, we can conclude that inserting a pause between words *does* help significantly in identifying speech recognition errors at the preferred speech rate of 200 WPM, but the length of that pause does not matter. What is most important is the existence of a pause, perhaps because it eliminates elision.

STUDY 4: EFFECT OF LISTENING TO SPEECH TWICE

Inserting pauses between words lengthens the time for audio playback. As already mentioned, even with only a 1ms pause, there was an additional ~45% for playback time over *no pause* with the text-to-speech engine we used in Study 2. In this final study, we conducted an initial assessment of an alternative approach to making use of extra time: simply repeating the audio clip twice compared to listening to it only once. Participants in the earlier studies had only been allowed to listen to each audio clip once, to assess their first-pass ability to identify errors. However, repeating the audio twice could improve error identification. While it may be useful to assess the effects of repetition in more detail in future work, for this first evaluation, we compared clips at 200 WPM played only once versus played twice, with no pauses between words.

Method

The method is the same as for Study 2 except as follows.

Participants

We recruited 30 participants (17 male, 13 female). Participants ranged in age from 23 to 66 (M=36.6, SD=11.5). All participants were native English speakers with no hearing loss. All participants reported completing the task in a quiet room, except for one who reported light background noise. Seventeen participants had previously used speech input on their phone or computer.

Study Design

We used a within-subjects design with two conditions: *Default* or *Repeat*. With *Default*, the audio feedback played once at 200 WPM with no pause between words, whereas with *Repeat* the audio played twice at 200 WPM with no pause between words and a chime sound (1.1s long) between repetitions. The two conditions were presented in counterbalanced order. Participants were randomly assigned to orders. Precision, recall, phrase-level accuracy, and trial completion time data all violated the normality assumption

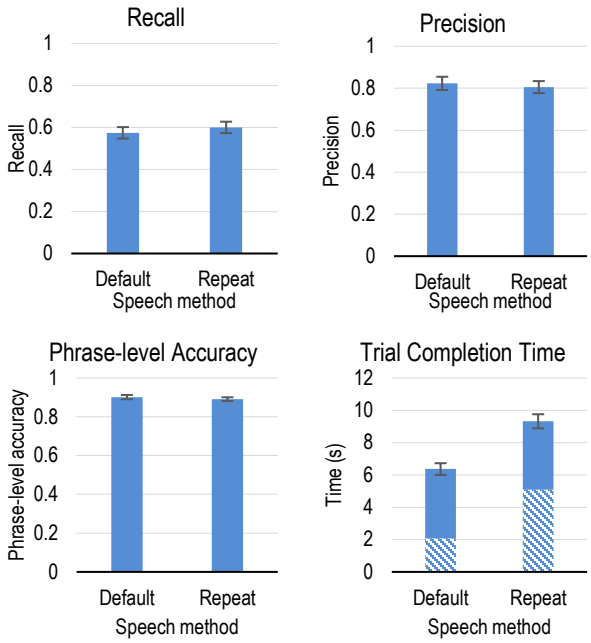

Figure 7. Graphs of precision, recall, and phrase-level accuracy in Study 4. The shaded portion in trial completion time indicates the average length of audio clips in that condition. Only time was significantly different between the two conditions. The error bars are standard errors (N=30).

(Shapiro-Wilk test, p<.05). Therefore, Wilcoxon signed rank tests were used to compare the two conditions.

Result

Figure 7 shows the measures for Study 4. There was no significant difference in recall between the two conditions (W=192, Z=-0.56, p=.591, r=0.07). The differences in precision and phrase-level accuracy were also not significant (respectively: W=184, Z=0.49508, p=.633, r=0.06; W=213, Z=1.0238, p=.315, r=0.132). The average length of the audio clips was 2.1s (SD=0.01) in the *Default* condition and 5.1s (SD=0.02) in the *Repeat* condition. Due to the longer length of audio, trial completion time in the *Repeat* condition was also longer than in the *Default* condition, at 9.3s (SD=0.4) compared to 6.4s (SD=0.4); the difference was significant (W=1, Z=-4.76, p<.001, r=0.61).

In summary, listening to audio clips twice did not improve the accuracy measures, although it added length to the audio clip. However, this result should be considered to be a preliminary exploration of the repetition approach, with more work needed to evaluate potential interactions with speech speeds, pause between words, and repetition.

DISCUSSION AND FUTURE WORK

Combined, the four studies show, first, that identifying speech recognition errors through audio-only interaction is hard: participants missed identifying over 50% of errors in Study 1, the majority of which included speech sounds strung across multiple words (e.g., one word recognized as two separate words that sound similar to the original word). Studies 2 to 4 then explored three straightforward speech

output manipulations, showing that adding even an imperceptibly brief pause (1ms) between words increases recall and phrase-level accuracy. In terms of speech rate, a high speech rate of 300 WPM reduced the ability identify errors compared to slower and more subjectively comfortable rates. Finally, repeating the audio output (i.e., playing it twice instead of once) did not impact error identification, at least at a 200 WPM speech rate.

Designing Audio-Only Speech Input

With the widespread adoption of speech interfaces, enabling highly accurate text input through audio-only interaction will need to be addressed. Our studies point to promising and simple manipulations that should increase text input accuracy if used during text dictation and correction. Adding a short (even 1ms) pause to eliminate elision and using a speech rate of ~200 WPM or slower was the best combination (this rate matches recommendations for speech synthesis in general [24]). However, even with this improvement, the best average recall rates in Studies 2 and 3 were ~0.70, which highlights an opportunity for more sophisticated audio-only error identification support. Phrase-level accuracy is more promising, in that many phrase-level accuracy rates were above 90%, indicating that most often a user would know that there is at least one error and they could re-dictate the entire phrase. Still, improvement is needed to bring audio-only text input more in line with the accuracy that can be achieved with visual feedback. One possibility is to explore audio techniques that are analogous to visually underlining words that the recognition system deems to be potentially incorrect.

The manipulations we evaluated all add time to the audio output. One design choice would be to add very short inter-word pauses to all dictated text output. However, it may be preferable to provide user control over whether to achieve higher input accuracy at the cost of this extra time. Users could listen to the text output using typical speech settings (e.g., no pause), then if they detect the possibility of an error, they could review the text again in more detail using pauses and slower speech. Depending on the speech recognizer's accuracy, in fact, this could be overall the most efficient interaction style, achieving both high speed and text input accuracy. More work is needed.

Users may also have individual preferences regarding the tradeoff between speed and text input accuracy, where some may be more concerned than others about missed errors. The level of concern will also vary based on task context, similar to how the acceptability of handwriting input recognition errors varies based on context [13]. For example, sending an informal text message to a spouse likely does not require the same level of attention to accuracy as writing an email to one's work supervisor.

Previous work has shown that experience with synthesized speech impacts comprehensibility (e.g., [20]). As such, it will be important to investigate how experience may interact with the effectiveness of pauses, speech rate, and repetition on audio-only error identification. For example, users with visual impairments who are experienced with screen readers, will likely perform differently than the sighted users included in our study. Another factor that could impact the ability of users to identify errors when reviewing audio is the attentional demand in many settings where speech is used, such as mobile interaction [18]. Related, while we did not see an impact of background noise level on error identification rates in Study 1, we hypothesize that our background noise was simply too quiet and that louder noise would cause lower identification rates.

Limitations

Our study has limitations that should be addressed in future work. First, though we purposely chose a single-factor design for Study 3 to bolster statistical power, the follow-up data collection showed that there is an interaction effect between pause length and speech rate, an interaction that should be examined further. Second, we used a transcription task where participants dictated a presented phrase. This approach allows for precise measurement of error identification rates (by comparing errors and participant responses to the presented phrases) but is less realistic than a free-form dictation task would be. Third, we reused the phrase set and errors from Study 1 for all subsequent studies. It will thus be important to generalize the findings to different phrase sets. Fourth, participants were only provided with feedback on whether they had correctly identified speech recognition errors during the practice trials in Studies 2-4 but not during test trials. It is possible that such feedback, while not representative of real use, would have impacted performance and subjective responses. Finally, while we did not observe any impacts due to the quality of the synthesized speech, future work should examine potential impacts of different types of speech synthesis engines on error identification.

CONCLUSION

We reported on four studies to characterize and address the difficulty of identifying speech recognition errors when using audio-only speech input. Study 1 revealed that by listening to audio clips alone, users could identify less than half of the speech recognition errors. We then addressed the most common type of error that participants had missed in Study 1—errors where multiple words blended together—by inserting pauses between each word and varying speech rate in the audio output. The simple solution of inserting even a 1ms pause between words improved the ability to identify errors, while a fast speech rate made the task more difficult, and repeating the audio output had no effect. These findings have implications for speech-based text input for a variety of non-visual contexts, and an important avenue of future work will be to extend the investigation to accessibility for blind and visually impaired users.

REFERENCES

1. Shiri Azenkot and Nicole B Lee. 2013. Exploring the use of speech input by blind people on mobile devices. *Proceedings of the ACM SIGACCESS Conference on*

Computers and Accessibility, ACM, Article No. 11.

2. Ann R Bradlow and Jennifer A Alexander. 2007. Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *The Journal of the Acoustical Society of America* 121, 4: 2339–2349.
3. Junhwi Choi, Kyungduk Kim, Sungjin Lee, et al. 2012. Seamless error correction interface for voice word processor. *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 4973–4976.
4. W Feng. 1994. Using handwriting and gesture recognition to correct speech recognition errors. *Urbana* 51: 61801.
5. 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. *International Journal of Human-Computer Interaction* 19, 3: 279–304.
6. Kazuki Fujiwara. 2016. Error Correction of Speech Recognition by Custom Phonetic Alphabet Input for Ultra-Small Devices. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 104–109.
7. Beth G Greene. 1986. Perception of synthetic speech by nonnative speakers of English. *Proceedings of the Human Factors Society Annual Meeting*, 1340–1343.
8. David Huggins-Daines and Alexander I Rudnicky. 2008. Interactive asr error correction for touchscreen devices. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, 17–19.
9. Esther Janse. 2004. Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech. *Speech Communication* 42, 2: 155–173.
10. Hui Jiang. 2005. Confidence measures for speech recognition: A survey. *Speech communication* 45, 4: 455–470.
11. Caroline Jones, Lynn Berry, and Catherine Stevens. 2007. Synthesized speech intelligibility and persuasion: Speech rate and non-native listeners. *Computer Speech & Language* 21, 4: 641–651.
12. Clare-Marie Karat, Christine Halverson, Daniel Horn, and John Karat. 1999. Patterns of entry and correction in large vocabulary continuous speech recognition systems. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 568–575.
13. Mary LaLomia. 1994. User Acceptance of Handwritten Recognition Accuracy. *Conference Companion on Human Factors in Computing Systems*, ACM, 107–108.
14. Yuan Liang, Koji Iwano, and Koichi Shinoda. 2014. Simple gesture-based error correction interface for smartphone speech recognition. *INTERSPEECH*, 1194–1198.
15. Iain A McCowan, Darren Moore, John Dines, et al. 2004. *On the use of information retrieval measures for speech recognition evaluation*. .
16. Anja Moos and Jürgen Trouvain. 2007. Comprehension of ultra-fast speech--blind vs."normally hearing" persons. *Proceedings of the 16th International Congress of Phonetic Sciences*, 677–680.
17. Jun Ogata and Masataka Goto. 2005. Speech repair: quick error correction just by using selection operation for speech input interfaces. *INTERSPEECH*, 133–136.
18. Antti Oulasvirta, Sakari Tamminen, Virpi Roto, and Jaana Kuorelahti. 2005. Interaction in 4-second bursts: the fragmented nature of attentional resources in mobile HCI. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 919–928.
19. Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 5206–5210.
20. Konstantinos Papadopoulos and Eleni Koustriava. 2015. Comprehension of Synthetic and Natural Speech: Differences among Sighted and Visually Impaired Young Adults. *Enabling Access for Persons with Visual Impairment*: 147.
21. Sherry Ruan, Jacob O Wobbrock, Kenny Liou, Andrew Ng, and James Landay. 2016. Speech Is 3x Faster than Typing for English and Mandarin Text Entry on Mobile Devices. *arXiv preprint arXiv:1608.07323*.
22. Amanda Stent, Ann Syrdal, and Taniya Mishra. 2011. On the intelligibility of fast synthesized speech for individuals with early-onset blindness. *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, 211–218.
23. Bernhard Suhm, Brad Myers, and Alex Waibel. 2001. Multimodal error correction for speech user interfaces. *ACM transactions on computer-human interaction (TOCHI)* 8, 1: 60–98.
24. Brenda Sutton, Julia King, Karen Hux, and David Beukelman. 1995. Younger and older adults' rate performance when listening to synthetic speech. *Augmentative and Alternative Communication* 11, 3: 147–153.
25. TheMSsoundeffects. City sound effect 1 - downtown. Retrieved from https://youtu.be/LZbEIxhiJRM.

26. Simon Tucker and Steve Whittaker. 2005. Novel techniques for time-compressing speech: an exploratory study. *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, I--477.

27. Lijuan Wang, Tao Hu, Peng Liu, and Frank K Soong. 2008. Efficient handwriting correction of speech recognition errors with template constrained posterior (TCP). *INTERSPEECH*, 2659–2662.

28. Zhirong Wang, Tanja Schultz, and Alex Waibel. 2003. Comparison of acoustic model adaptation techniques on non-native speech. *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, I--I.

29. Stephen J Winters and David B Pisoni. 2004. Perception and comprehension of synthetic speech. *Progress Report Research on Spoken Language Processing* 26: 1–44.

30. Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 143–146.

31. Wayne Xiong, Jasha Droppo, Xuedong Huang, et al. 2017. The Microsoft 2016 conversational speech recognition system. *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, 5255–5259.

32. Hanlu Ye, Meethu Malu, Uran Oh, and Leah Findlater. 2014. Current and future mobile and wearable device use by people with visual impairments. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*: 3123–3132.

33. Geoffrey Zweig, Chengzhu Yu, Jasha Droppo, and Andreas Stolcke. 2017. Advances in all-neural speech recognition. *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, 4805–4809.