Towards More Robust Speech Interactions for Deaf and Hard of Hearing Users

Raymond Fok
University of Michigan
Ann Arbor, MI
rayfok@umich.edu

Harmanpreet Kaur
University of Michigan
Ann Arbor, MI
harmank@umich.edu

Skanda Palani
University of Michigan
Ann Arbor, MI
spalani@umich.edu

Martez E. Mott
University of Washington
Seattle, WA
memott@uw.edu

Walter S. Lasecki
University of Michigan
Ann Arbor, MI
wlasecki@umich.edu

ABSTRACT

Mobile, wearable, and other ubiquitous computing devices are increasingly creating a context in which conventional keyboard and screen-based inputs are being replaced in favor of more natural speech-based interactions. Digital personal assistants use speech to control a wide range of functionality, from environmental controls to information access. However, many deaf and hard-of-hearing users have speech patterns that vary from those of hearing users due to incomplete acoustic feedback from their own voices. Because automatic speech recognition (ASR) systems are largely trained using speech from hearing individuals, speech-controlled technologies are typically inaccessible to deaf users. Prior work has focused on providing deaf users access to aural output via real-time captioning or signing, but little has been done to improve users' ability to provide *input* to these systems' speech-based interfaces. Further, the vocalization patterns of deaf speech often makes accurate recognition intractable for both automated systems and human listeners alike, causing traditional approaches to mitigating ASR limitations, such as human captionists, less effective. To bridge this accessibility gap, we investigate the limitations of common speech recognition approaches and techniques – both automatic and human-powered – when applied to deaf speech. We then explore the effectiveness of an iterative crowdsourcing workflow, and characterize the potential for groups to collectively exceed the performance of individuals. This paper contributes a better understanding of the challenges of deaf speech recognition and provides insights for future system development in this space.

ASSETS 2018, October 22–24, 2018, Galway, Ireland

ACM ISBN 978-1-4503-5650-3/18/10 ...$15.00.
http://dx.doi.org/10.1145/3234695.3236343

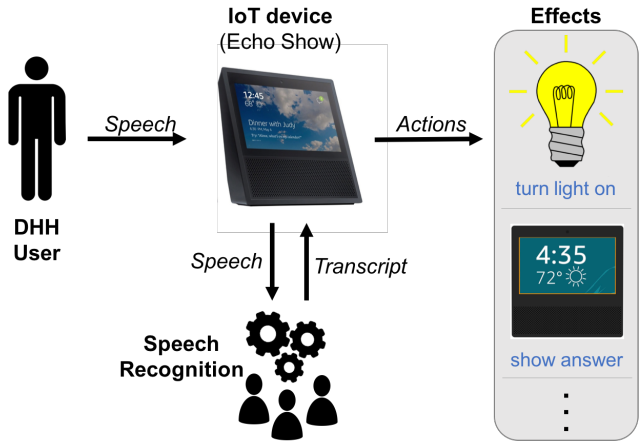

Figure 1. Example interaction setup for our work. Here, a deaf or hard-of-hearing (DHH) user interacts via speech with an intelligent agent (e.g., on a smartphone or Amazon Echo Show device). Based on the output of a speech recognition process, the system either performs an action (e.g., turns on a light) or provides on-screen feedback. Because most automatic speech recognition systems are trained on speech from hearing users, DHH users are often unable to use these devices effectively due to a "deaf speech" accent. Our work explores the viability of using current automatic and human-powered approaches to bridge this accessibility gap, and suggests directions and insights for future work to create more powerful and robust speech-based interfaces for DHH users.

Author Keywords

Accessibility; Automatic Speech Recognition; Human Computation; Deaf Speech; Deaf and Hard-of-Hearing

INTRODUCTION

Speech is becoming an increasingly common means of providing input to computing devices in our daily lives. Companies like Apple (Siri), Microsoft (Cortana) and Amazon (Alexa) have popularized digital personal assistants that simplify interactions around daily tasks — such as setting timers, accessing information on the weather, responding to messages, changing the temperature in a room, and much more — via spoken natural language. Smartphones and in-home Internet-of-Things

(IoT) devices like the Amazon Echo Show and Echo Spot provide visual feedback mechanisms that improve access for those who cannot hear spoken responses. However there are significant limitations to the *speech recognition* capabilities of these automated systems for people with uncommon speech patterns. As such, commands to these assistants are often met by responses like "Sorry, I'm having trouble understanding you right now. Please try again later."

Technical issues with speech-based interaction due to poor speech recognition are inconvenient to the typical user, but can be frequent enough to make these devices inaccessible to many people who are deaf or hard-of-hearing (DHH). This is because, depending on the recency and severity of their hearing impairment, the speech patterns of DHH individuals can vary significantly from existing large-scale speech datasets. As a result, their speech is not well-recognized by automated speech recognition (ASR) systems trained on these more common datasets. Initial results have demonstrated that ASR and crowdsourcing approaches are far from recognizing deaf speech accurately enough to provide transcriptions that are usable in a real-world setting [3, 11].

Prior research has attempted to make aural information accessible to DHH individuals, introducing real-time captioning in classrooms [14, 34, 16, 15], wearable assistive technologies [28, 37], and novel speech-to-sign systems [8]. In contrast, almost no work has been done to better handle the speech produced by DHH individuals or otherwise provide access to these increasingly-ubiquitous speech-based interfaces. To inform future research in deaf speech recognition, this paper explores the current scope of the problem, and seeks to better understand aspects that future work may be able to leverage.

We first test ASR and individual crowd workers on transcribing deaf speech, and find that crowd workers (individually) produce significantly better transcriptions than ASR (0.70 versus 0.54 word error rate, respectively). However, these word error rates are too high to be used in real-world settings. We then evaluate an iterative crowdsourcing approach to transcription, and find that crowd workers in an iterative process generate significantly better transcriptions than individual crowd workers for more intelligible deaf speech, but fail to improve quality for less intelligible deaf speech. Finally, we explore how context, task decomposition, and speech rate can be leveraged to potentially improve collective performance in future systems. Overall, this paper characterizes the problem of deaf speech transcription, and empirically explores various inroads towards a potential solution.

The remainder of this paper is structured as follows: (1) we present background on deaf speech and existing automated and crowd-powered systems for captioning; (2) we evaluate automated and individual crowd worker approaches to speech recognition as baselines for transcription of deaf speech; (3) we present three studies to understand how (i) modifying clip speed, (ii) breaking down audio into smaller segments, and (iii) surrounding linguistic context, affect transcription quality; (4) we evaluate the effectiveness of an iterative crowd-powered workflow; and (5) we evaluate the baseline and iterative approaches in the higher-context domain of Alexa commands.

This paper makes the following contributions:

- A characterization of existing approaches to deaf speech recognition that use fully-automated approaches, individual human contributions, or collective (crowd) input.
- An exploration of techniques used to improve human captioning performance on deaf speech input, i.e., speed modification, audio decomposition, and iteration.

BACKGROUND AND RELATED WORK

Our goal is to enable deaf speech recognition by speech-based UIs to better support deaf speech. To do this, we apply crowdsourcing workflows to deaf speech transcription such that the results can be fed as input to these devices. Our work extends the literature on deaf speech and speech captioning. Below, we discuss prior work in these domains.

Deaf Speech

Because deaf (and significantly hard-of-hearing) people are unable to hear the speech produced by themselves and others, and consequently lack direct feedback to their own vocalizations, their speech patterns often differ from those of hearing individuals. *Deaf speech* refers to accented speech produced by many individuals with partial or complete hearing loss. The severity of this accent often depends on when the individual lost their ability to hear and the level of hearing loss. Prior research studying the effects of deafness on speech have classified common errors made by deaf individuals. These include phonological errors such substitution, omission errors, and consonant-clustering errors [33]. These phonological and articulation errors contribute negatively to voice quality and speech intelligibility [29]. The pace of deaf speech is also considerably slower, on average, than speech produced by individuals without hearing impairments (hearing speech), due to vowel prolongation and the insertion of extraneous pauses [33]. This rhythmic inconsistency both within and between individuals has been shown to hurt speech intelligibility [13]. Though experience does have some effect on recognition, understanding deaf speech remains a challenge to both experienced and inexperienced listeners alike [27]. We build on this literature by providing empirical results for the application of existing transcription approaches to deaf speech, and show that while these errors render automated approaches inadequate, appropriate human computation workflows can be effective. Further, prior work has found that the amount of linguistic context present affected how well a clip could be understood [26, 27]. We extend this by exploring how to progressively build linguistic context using an iterative crowdsourcing workflow.

Automated Speech Recognition

Automated Speech Recognition (ASR) is popular for real-time captioning and is used in many current speech-based systems, including personal assistants and other IoT devices. It performs well in ideal situations with high-fidelity audio, but its accuracy deteriorates quickly in real-world settings. Since its underlying model is largely trained on hearing speech patterns, it does not adapt well to heavily accented or deaf speech. Prior work has studied the effectiveness of ASR on speech that is more difficult to understand; for example, a

study with dysarthric speech found that the performance of ASR on impaired speech was significantly affected by speech intelligibility, severity, and intra-speaker variability [31]. Various approaches have been attempted to improve the accuracy of ASR for dysarthic speech by reducing the influence of these factors, such as pooling hearing speech data to improve acoustic models [36] and training several neural networks in parallel to form *array learners* [35]. Approaches like these improve the performance of ASR, but require substantial computing power or special audio equipment, which is often unavailable in IoT or mobile devices. There are instances where ASR can be trained on the individual speaker's specific acoustic, pronunciation and language models to improve performance slightly, but this is not readily available and is untested for deaf individuals. We present results from testing Google's speaker-independent speech recognition system on deaf speech.

Human Captioning and Crowdsourcing

In cases when automated approaches fall short, people have also long been used for audio transcription and captioning tasks [6, 17]. Professional transcription services like Communication Access Real-Time Translation (CART) are reliable and produce transcriptions in real-time, but are too expensive for use in everyday settings (costing up to and exceeding $150/hr). Crowd-powered systems with non-expert captionists can lower this cost, supported by microtask marketplaces like Amazon's Mechanical Turk (MTurk). These platforms provide availability with a large pool of crowd workers who can be recruited on-demand [1] and engaged in continuous tasks for flexible periods of time [20]. In an accessibility context, crowds have been studied to provide intelligent access technology [4], answer visual questions [2, 12, 21, 38], and provide real-time captions [17, 19]. Other paradigms like iteration have been used to improve crowd performance in tasks like deciphering blurry text, where parallel, independent workers perform worse than workers in an iterative process [23]. This paradigm has been applied to hearing speech transcription, and shown to be successful (achieving 96.6% accuracy) [22]. In contrast to prior crowd-powered systems which provide captions for hearing speech, this paper explores how crowds can be used to transcribe deaf speech. Furthermore, we test an iterative approach, making the conscious decision to characterize crowd-based approaches that are not real-time, in an attempt to improve accuracy. Further, systems like Scribe [17] and Adrenaline [1] suggest that reducing transcription time to seconds, or even faster [24], is possible. However, given the difficulty of accuracy at any speed that we demonstrate in this paper, we leave real-time generation of deaf speech captions as future work.

EVALUATION METRICS

We use word error rate as our primary metric to measure the performance of current automatic and human-powered approaches to transcribing deaf speech. In our experiments, we use deaf speech examples gathered from the Clarke sentences dataset [25].

Word Error Rate

Word Error Rate (WER) is a common metric used to evaluate the performance of speech recognition and transcription systems. Derived from Levenshtein distance but on a word level, WER measures the number of modifications needed to turn a system's output transcription (hypothesis) into the ground truth transcription (reference), normalized by the number of words in the reference. The hypothesis and the reference can have different lengths, so they are dynamically aligned before the minimum number of modifications are calculated. While WER is technically an unbounded non-negative value, we limit WER between 0 and 1 inclusive, so WER can be interpreted as the percentage of incorrectness in a transcription.

$$\text{Word Error Rate } (\mathbf{WER}) = \frac{S+D+I}{N} \tag{1}$$

Equation 1 shows the formulation of WER, where S is the number of substitutions, D is the number of deletions, and I is the number of insertions, required to transform the target transcript to match the ground truth answer (which itself has N words). When calculating WER, we removed punctuation, converted all text to lowercase, converted numbers to their word representation, and removed any indications of worker uncertainty in our data (e.g., "..." or "<unintelligible>"). No other processing (e.g., stemming, lemmatization) was performed, and other equivalencies (e.g., abbreviation, contraction, and acronyms) were ignored since these were not common in the experimental dataset.

Clarke Sentences Dataset

We use the Clarke sentences dataset [25] to evaluate baseline approaches and our proposed workflow. The Clarke sentences dataset is a subset of a larger collection of audio recordings from 650 DHH individuals who took the Clarke sentences intelligibility test. Examples of Clarke sentences include "Bobby had hot cereal for supper" and "The water at the farm was very warm." The number of words per sentence varies, but each sentence has exactly 10 syllables. The dataset also includes an intelligibility score for each individual, which was measured as the number of words out of 50 pre-selected non-stop words recognized by a designated speech pathologist. The intelligibility scores range from 0 to 50, with an intelligibility score of 50 indicating the clip was generally intelligible, and 0 indicating the clip was completely unintelligible.

For our experiments, we selected five audio files from each of three levels of intelligibility–30, 40, and 50–for a total of 15 audio files. These files were split into 10 clips, 1 sentence per clip, for a total of 150 audio clips. We selected batches of clips at discrete intelligibility levels to broadly observe the relationship of clip intelligibility and transcription accuracy. We conducted a preliminary study with clips at intelligibility levels 10 and 20, and found that neither ASR nor crowdsourced approaches could generate transcriptions for these clips (WER for level 10: ASR=1.0, crowd=0.97; level 20: ASR=1.0, crowd=0.98). Further, of the 650 DHH individuals who participated in the Clarke sentences intelligibility test, about 50% fell between intelligibility levels 40 and 50, 25% fell between intelligibility levels 30 and 40, and 25% fell below intelligibility level 30 [11]. Given this, we leave intelligibility levels 10 and 20 for future work, as there is minimal signal for those in even baseline speech recognition methods.

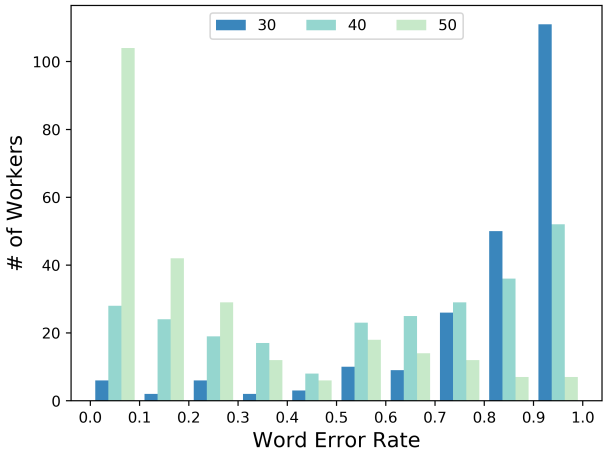

Figure 2. Word error rate (WER) distribution of transcriptions generated by individual crowd workers, separated by three levels of clip intelligibility. A lower WER is better and indicates a more accurate transcription. More intelligible clips tended to have lower WER, while less intelligible clips tended to have higher WER.

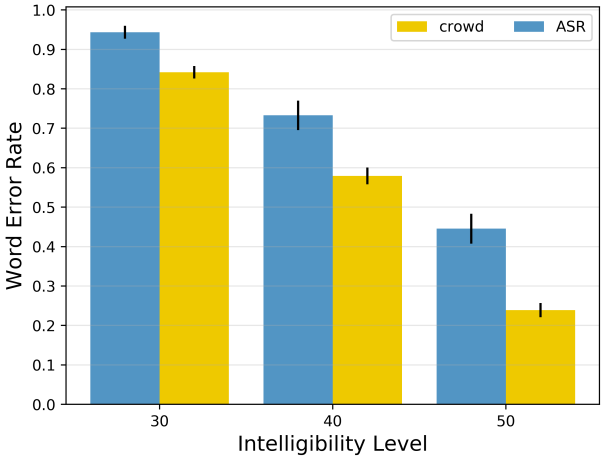

Figure 3. A comparison of average transcription WER with automated and individual crowd worker approaches. A lower WER is better and indicates a more accurate transcription. Overall, crowd workers generated better transcriptions than ASR, with the difference more evident at higher intelligibility levels.

We construct our experimental dataset with clips at three discrete intelligibility levels–30, 40, and 50–encompassing a total of ∼75% of DHH individuals.

BASELINE APPROACHES

We use automated speech recognition (ASR) and human transcription (via online crowds), two common methods for transcribing speech, as our baselines. Below, we present results from running these two approaches on our selected Clarke sentences dataset.

Baseline Study Setup

For the automated approach, we passed each clip to Google's Speech Recognition API and recorded the output transcriptions with the highest confidence scores. For the human computation approach, we recruited crowd workers on Amazon's Mechanical Turk platform, filtered to those with over 95% approval rating and located in the United States. Each worker was shown an interface with a small set of media controls for one audio clip, a text box, and a short snippet of instructions asking them to play and transcribe the clip. Workers were given no indication of clip content, but were told that the task was expected to be difficult and that they should provide a transcription to the best of their ability. In each crowd task, a worker transcribed a total of five different clips selected randomly from our test set. We did not allow the any workers to complete the task more than once to avoid learning effects. Each worker was paid $0.25 per task, an approximate pay rate of $8.00/hour. We collected 5 crowd worker transcriptions for each clip, for a total of 750 transcriptions (3 intelligibility levels $\times$ 50 clips per level $\times$ 5 crowd worker transcriptions).

Baseline Results

Transcriptions produced by ASR and individual crowd workers for deaf speech were poor quality, overall. ASR had an average WER of 0.70 ($\sigma = 0.31$) and the crowd-powered approach had an average WER of 0.54 ($\sigma = 0.38$). An independent-samples t-test shows that the crowd-powered approach has significantly lower WER ($t(807) = 4.94, p < .0001$). While crowd workers outperformed ASR, the WERs of both approaches are too high for any real-world scenario. To be usable in practice, transcripts should not have a WER of greater than 0.25 [30]; however for IoT and mobile devices, the WER benchmark for these transcripts may have to be significantly lower. To overcome the remaining gap between our baseline approaches and acceptable error rates, the next section of this paper explores more complex human workflows that engage groups of people (e.g., crowds) to improve collective performance.

The Effect of Intelligibility on Error Rate

Audio clips with higher intelligibility levels tended to result in better transcriptions. We conducted separate independent-samples t-tests with Bonferroni correction for both ASR and individual crowd worker approaches. Transcriptions produced by ASR had a significant difference in WER between intelligibility levels 30 and 40 ($t(98) = 5.16, p < .0001$), and levels 40 and 50 ($t(98) = 5.39, p < .0001$). Individual crowd worker transcriptions also had a significant difference in WER between intelligibility levels 30 and 40 ($t(484) = 9.64, p < .0001$), and levels 40 and 50 ($t(510) = 12.15, p < .0001$). Figure 2 shows the overall WER distribution of crowd workers on the Clarke sentences dataset for each intelligibility level, and Figure 3 compares the WER of transcriptions for automated and individual crowd worker approaches. The WER increased significantly with drops in intelligibility level (Figure 2)–transcriptions of both approaches at intelligibility level 30 had the highest WER.

Crowd workers performed increasingly better than ASR as intelligibility level increased (Figure 3). Crowd worker transcriptions had 11%, 20%, and 43% lower WER than automatically generated transcriptions at intelligibility levels 30, 40, and 50, respectively. An independent-samples t-test with Bonferroni correction conducted for each level showed these differences were significant at intelligibility levels 30

($t(295) = 2.96, p < .005$), 40 ($t(331) = 2.99, p < .005$), and 50 ($t(321) = 4.70, p < .0001$). A possible explanation is that crowd workers are able to recognize words and patterns that resemble hearing speech in more intelligible deaf speech, while ASR is less responsive to these similarities. This improvement lessens with lower intelligibility, as it was more dissimilar to hearing speech and thus less familiar to crowd workers.

Latency

Crowd workers are able to more quickly transcribe deaf speech with increasing intelligibility levels. Their transcriptions took $57.3s$ ($\sigma = 55.5s$), $47.2s$ ($\sigma = 36.7s$), and $36.2s$ ($\sigma = 36.5s$) at intelligibility levels 30, 40, and 50 respectively. A one-way ANOVA showed a significant effect of intelligibility level on transcription time ($F(2,781) = 15.19, p < .0001$). Post hoc comparisons using the Tukey HSD test indicated there was a significant difference in transcription time between two pairs of intelligibility levels, 30 and 50, and 40 and 50 ($p < 0.05$).

Random Sampling To Predict Worker Accuracy

Most crowdsourcing systems filter out responses from low-quality workers to improve average accuracy, often by inserting gold standard questions in the task and post hoc filtering workers who fall below an accuracy threshold for these questions [5, 9]. We observed that for deaf speech transcription, individual worker performance was fairly consistent across all five clips they were each asked to transcribe: some performed well across all five clips (WER < 0.1), and others poorly (WER = 1.0). Given this, we simulated the effect of using worker performance on one clip as a gold standard to filter or include the rest of their responses. We calculated a Pearson product-moment correlation coefficient analyzing the relationship between one randomly sampled crowd worker transcription (sample) and the rest of that worker's transcriptions (outer). We used the baseline results as the dataset for this analysis; since workers were given clips of varying intelligibility during that experiment, we measured performance relative to the average performance for clips of that intelligibility. There was a positive correlation between the two variables, sample and outer WER ($r = 0.578, p < .0001$). Figure 4 summarizes the results in a scatterplot. This suggests that filtering based on worker performance on one clip could be used to improve average transcription quality.

IMPROVING INDIVIDUAL WORKER PERFORMANCE

The baseline results show both automated and human transcription approaches are unable to recognize deaf speech, though human computation significantly outperforms ASR. In this section we study common techniques used in crowd-powered speech recognition systems to evaluate their effectiveness for improving improving the quality of deaf speech transcriptions. Specifically, we study clip speed modification, audio decomposition, and best-case iteration.

For the following studies, crowd workers were recruited from Amazon Mechanical Turk. Each crowd worker completed five transcription tasks per HIT, and was paid $0.25 per HIT based on a pay rate of $8.00/hour. To minimize learning effects, crowd workers could only do one HIT for each study.

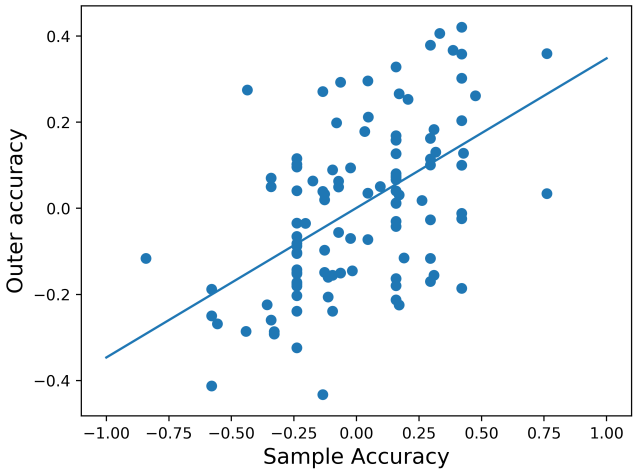

Figure 4. There was a positive correlation between the quality of one randomly sampled transcription generated by a unique crowd worker (sample), and the quality of the remainder of their transcriptions (outer). This suggests that filtering based on worker performance on one clip could be used to improve average transcription quality.

Speed Modification

We first explore whether changes in clip speed affect a crowd worker's ability to understand it. Prior crowd-powered captioning approaches have slowed down speech to allow crowd workers to keep up with the audio stream [18]. However deaf speech is generally 1.5× to 2× slower than hearing speech [33], and prone to timing errors such as pauses and irregular syllable duration [32]. Intuitively, transforming the audio to a temporal structure similar to that of hearing speech would make speech more familiar to non-expert listeners. We explore the effects of both on the transcription quality of crowd workers: slowing down and speeding up the audio clips.

Study Design

We selected five sentences from each of intelligibility levels 30, 40, and 50. New clips were generated by time stretching each clip by a speed-modification factor of 0.7 to 1.5, in increments of 0.1, giving a total of 135 clips. Each modified clip was transcribed by 5 crowd workers, and we measured the accuracy of the resulting transcriptions.

Results

We found that speed modification *did not* have a significant effect on the WER of crowd worker transcriptions at any intelligibility levels. We ran a one-way ANOVA for each intelligibility level: level 30 ($F(8,216) = 1.09, p = .373$), 40 ($F(8,223) = 0.30, p = .965$), and level 50 ($F(8,226) = 1.70, p = .099$). This suggests that slowing down deaf speech might not be needed because the main issue is not the ability to keep up with the audio stream, but rather the intelligibility of the audio itself. Similarly, for speeding deaf speech clips, our results imply that a naive stretching of deaf speech to normalize speaking rate is ineffective. This is reasonable since the slower rate of deaf speech is not due to an even prolongation of speech, but rather a variation in vocalization and pauses.

Audio Decomposition

We next study how decomposing an audio clip of deaf speech into shorter segments affects crowd worker transcription quality. Intuitively, there is a relation between a clip's length and the cognitive effort required for a worker to transcribe it. By creating shorter clips, crowd workers may better focus on each individual word. However, this decomposition can also remove linguistic context gained from the surrounding words, potentialy resulting in adverse effects: lowering transcription accuracy due to the need for workers to recognize words with little or no context.

Study Design

We selected two clips for each intelligibility level 30, 40, and 50; each clip corresponds to one sentence. Each clip was split manually at word boundaries into n clips with one word each, and $\lceil n/2 \rceil$ non-overlapping clips with two words each (n = the number of words in the sentence). In this way, a total of 71 decomposed clips were generated from our initial selection of 6 clips. Each decomposed clip was transcribed by 5 crowd workers. We evaluate the effectiveness of audio decomposition using recall, defined as a proportion of the number of words in the ground truth that were present in a worker's transcription to the total number of words in the ground truth.

Results

Overall, workers could recognize more given two-word segments than one-word segments. This difference was more pronounced as intelligibility level increased. Table 1 shows the average recall of worker transcriptions for clips split into one and two word segments. Our results suggest that trying to recognize deaf speech without sufficient linguistic context is difficult, and that fine-grained task decomposition may have adverse effects on deaf speech recognition.

Best-Case Iteration

Groups of crowd workers working together on a task are often able to perform a task better than any individual crowd worker. Little et al. first introduced this iterative crowdsourcing paradigm for crowd workers in TurKit, and tested it on the reconstruction of a hard-to-read handwriting sample [23]. Their results showed that while individuals performed poorly in parallel (i.e., independently), asking workers to iteratively build upon each other's responses enabled the reconstruction of most of the sample after about 15 steps. This approach is successful because *(i)* individual crowd workers relay at least some information to the next worker, despite poor overall performance at any one step, and *(ii)* people are able to synthesize disjoint context clues in forming their own response.

Since transcribing deaf speech is a similar style of task to reconstructing poor handwriting, i.e., difficult for any individual crowd worker but improved with additional context, we hypothesize that iteration may be a viable approach for improving our transcription accuracy. In this study, we test the effects of varying the amount of linguistic context on the resulting transcription. We provide crowd workers with a partial transcription and ask them to transcribe the remainder of the clip. This simulates a hypothetical, best-case iteration scenario in which a previous step in the iterative workflow produced an incomplete but otherwise correct transcription.

	I-30	I-40	I-50
1	.06 (.16)	.21 (.41)	.32 (.47)
2	.07 (.18)	.26 (.32)	.48 (.39)

Table 1. Results for the Audio Decomposition study, showing average recall and standard deviation by number of words in the audio clip. These results suggest audio decomposition hurts transcription quality.

	I-30	I-40	I-50
1	.58 (.50)	.79 (.41)	.92 (.28)
2	.58 (.45)	.72 (.37)	.87 (.27)
3	.56 (.36)	.70 (.37)	.91 (.15)
4	.57 (.36)	.67 (.25)	.94 (.13)

Table 2. Results for the Best-Case Iteration study, showing average recall and standard deviation by number of redacted words. These results suggest crowd workers are able to gather lingustic context from surrounding words in a partial transcription, with minimal changes in recall with 1 to 4 redacted words.

Study Design

We selected two clips from each of intelligibility levels 30, 40, and 50; each clip corresponds to one sentence. All selected clips contained exactly eight words. In addition to the baseline transcription interface, workers were given a partial transcription of the clip with k consecutive words redacted ($1 \leq k \leq 4$). For example, with the 6-word sentence "The dog ran on the grass" and $k = 4$ redacted words, possible partial transcriptions would be 1) "_ _ _ _ the grass", 2) "The _ _ _ _ grass", and 3) "The dog _ _ _ _". We collected responses from 5 crowd workers for each unique redacted set of words per clip, and measured recall for each response.

Results

We find that recall decreased only slightly as the number of redacted words increased. Table 2 shows the average recall for the redacted words. This is somewhat unexpected, but suggests that the number of redacted words in a partial transcription does not strongly affect crowd worker recognition. One reason for this may be that crowd workers infer the redacted words based on linguistic context alone, and use the audio clip to filter between the reasonable words they had in mind. A caveat to note here is that we provided the correct partial transcription and the number of words left to transcribe as a starting point, which is unlikely in a real system.

AN ITERATIVE TRANSCRIPTION WORKFLOW

The studies above show that iteratively refining transcriptions using groups of crowd workers results in high recall for deaf speech clips. However, the previous best-case iteration study was performed under idealized conditions where the previous step's transcription was incomplete but otherwise correct. This assumption affords information that would be unavailable in a real iterative workflow, such as the number of words in the clip and the relative positioning of words. Next, we test an iterative transcription workflow in a more realistic setting.

Workflow Design

To more thoroughly evaluate the efficacy of an iterative crowdsourcing approach, we designed a study in which crowd worker transcriptions were passed through a 10-step iterative workflow. In each iterative step, five workers independently

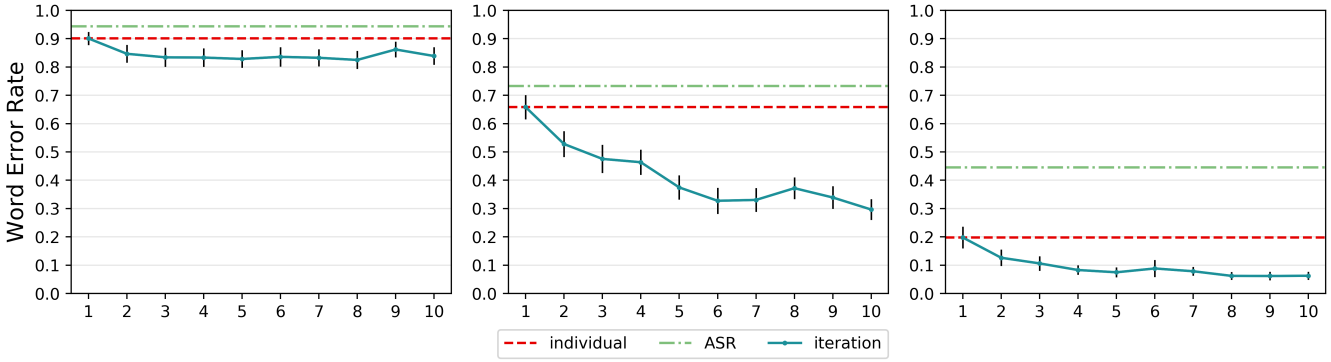

Figure 5. WER of transcriptions from iterative versus baseline approaches: intel 30 (left), intel 40 (center), intel 50 (right). The quality of iteratively-generated transcriptions quickly surpassed those from automated and individual crowd worker approaches for intelligibility levels 40 and 50. However the iterative approach failed to produce better transcription at intelligibility level 30.

transcribed a given clip. The worker interface was similar to the individual crowd worker baseline, except with a box containing the previous steps' five workers' transcriptions for that clip. We also provided instructions describing what the transcriptions were and their purpose for the current iteration step. In the first step of the iterative workflow, workers were not given any previous transcriptions, making it analogous to the baseline condition. We considered providing crowd workers with some starting transcription (e.g., ASR output) for the first iteration. However, prior studies have found that crowd workers tend to give worse transcriptions if given poor ASR output as a starting point [10, 17]. Therefore, we chose not to provide automatically generated transcriptions for the first set of workers.

In subsequent steps of the iterative workflow, crowd workers were given all five of the previous steps' transcriptions. We chose this fully-connected design purposefully to provide additional context to crowd workers as they transcribed the clip. With trade-offs in time and cost, this design allows for more robust transcriptions and minimizes the effects of one or more worker errors at each step. Some systems perform transcription aggregation via multiple sequence alignment [17], and these algorithms tend to provide minor performance gains when the transcriptions are reasonably accurate. We would expect these alignment algorithms to perform poorly when applied to deaf speech due to the transcriptions' high WER and variance.

Instead, our proposed iterative workflow uses human intelligence to perform implicit alignment. Prior work has shown crowd workers are capable of identifying correct transcriptions while listening to the clip, more so than producing their own correct transcription [3]. Human intelligence can readily recognize sentences that are grammatically, structurally, or semantically reasonable, quickly rejecting the sentences that fail those qualifications. By providing crowd workers with five previous transcriptions instead of just one, there is a higher possibility of at least one accurate transcription, which the crowd worker would be able to distinguish from the remaining inaccurate transcriptions. This also allows crowd workers to identify correct words within any of the transcriptions.

Experimental Setup

The original dataset tested in the baseline study consisted of 150 clips, 50 at each of three intelligibility levels. For testing the iterative approach, we subsampled 10 clips from each of the three levels, for a total experimental dataset of 30 clips. We recruited crowd workers on Amazon's Mechanical Turk platform, filtered to those with over 95% approval rating and located in the United States. Each worker was paid $0.25 per HIT, an approximate pay rate of $8.00/hour. Workers were not allowed to transcribe any one clip more than once throughout the ten steps of iteration.

All transcriptions were subject to the same post-processing process used in previous studies before the evaluation of WER. Results from iteration were compared against baseline results for transcriptions of the clips selected for the iterative study, rather than the entire dataset used in the baseline study. For significance testing, we compare against the individual crowd worker baseline approach because it had strictly lower WER than the automated approach.

Iteration Results

On average, the WER of crowd worker transcriptions after 10 steps of iteration was 3% lower than transcriptions by individual crowd workers for intelligibility level 30, 52% lower for intelligibility level 40, and 74% lower for intelligibility level 50. Independent-samples t-tests found that the difference was not significant at intelligibility level 30 ($t(93) = 1.00, p = .32$), but was significant at intelligibility levels 40 ($t(101) = 5.76, p < .0001$) and 50 ($t(99) = 4.18, p < .0001$).

For intelligibility levels 40 and 50, transcriptions generated by iteration had significantly lower WER than individual crowd workers by iteration step 5 (at $\alpha = .001$), and the rate of WER began to plateau with later iteration steps. The largest percentage decrease in WER occurred between iteration steps 1 and 2 for all intelligibility levels. Figure 5 shows the average WER at each iteration step against the WER of the automated baseline and the individual crowd worker baseline.

We tested whether an ideal number of iterations exists (i.e., a convergence point for worker transcriptions) by calculating cosine similarity for transcriptions between each iteration step.

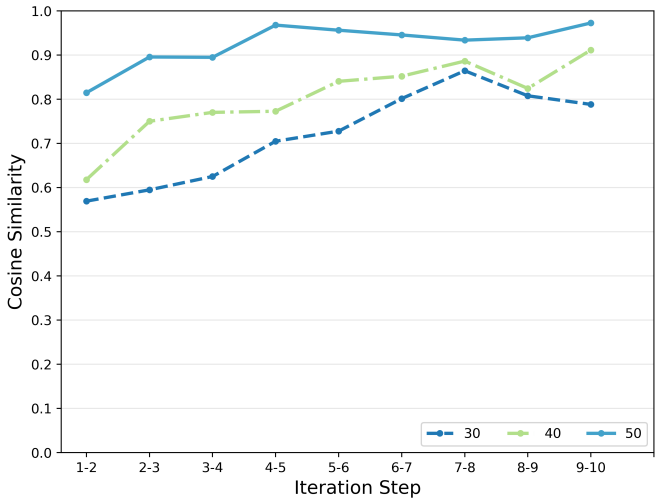

Figure 6. Average vector cosine similarity of crowd worker transcriptions at consecutive steps of iteration. Transcriptions tended to converge regardless of clip intelligibility and transcription accuracy.

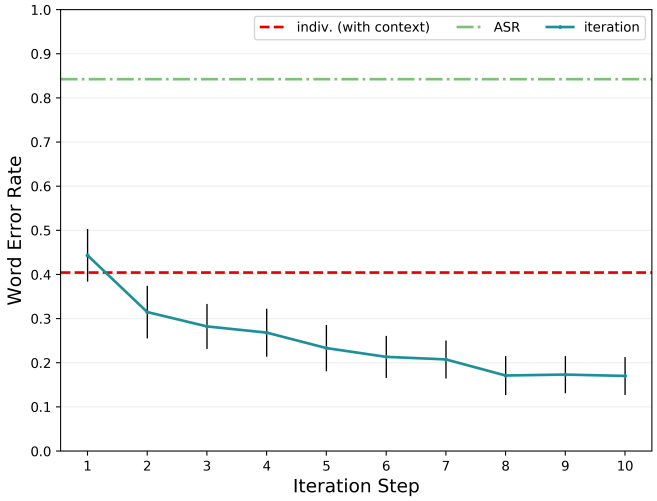

Figure 7. WER of iterative versus baseline approaches for the dataset of Alexa commands. The iterative approach produced transcriptions with significantly lower WER than those produced by individual crowd worker approaches. Further, the WER of transcriptions tended to decrease with each iteration step.

For each clip, we aggregated the five worker responses into a multiset, and used word multiplicity and ordering within sentences to construct a vector. We calculated vector cosine similarity between sentence vectors at each step, giving values in the range 0 to 1, where a similarity value of 0 represents no similarity between two sets of words, and a similarity value of 1 represents two sets of words are identical. The similarity between iteration steps 1 and 2 was .57 for intelligibility level 30, .62 for intelligibility level 40, and .81 for intelligibility level 50. The similarity between iteration steps 9 and 10 was .79 for intelligibility level 30, .91 for intelligibility level 40, and .97 for intelligibility level 50. Because worker transcriptions had more word variance at lower intelligibility levels at the beginning of iteration, the rate of similarity increase was greater for lower intelligibility levels. Overall, the similarity between steps had increasing trend, reaching 85% - 95% similarity for most clips after the tenth step of iteration. Figure 6 summarizes these results and illustrates the near monotonic convergence of worker transcriptions.

LEVERAGING DOMAIN-SPECIFIC CONTEXT

We have shown iterative crowdsourcing is effective at generating transcriptions with lower word error rates than baseline approaches for the Clarke sentences dataset. In this section we return to our original motivation and evaluate automated, individual crowd worker, and iterative crowdsourcing approaches for transcription of deaf speech in the domain of speech-controlled devices.

Alexa Dataset

To simulate a real-world scenario, we use a dataset collected by Bigham et al. with 10 of the most common commands to an Amazon Alexa personal assistant, spoken aloud by a DHH individual [3, 7]. Example commands include "Alexa, tell me a joke" and "Alexa, play music by Pearl Jam." Unlike the Clarke sentences dataset, these clips of Alexa commands have no corresponding intelligibility scores. We used the Alexa data because of the Amazon Echo's popularity as a personal assistant device, and because most of Alexa's functionality is accessed via speech without other alternatives.

The Impact of Thematic Context

In contrast to the general Clarke Sentences, Alexa commands are linked by a shared *thematic context*. More generally, personal assistants and other speech-controlled interfaces are situated in known domains, with bounded natural language inputs. These domain-specific devices can utilize their known thematic context in improving speech recognition. Thematic context can be similarly integrated in crowd-powered systems since understanding and transcribing an audio clip may be easier if a general sense of the clip's contents was known prior to listening. In this study we test the effects of thematic context on the accuracy of a crowd worker's transcriptions.

Study Setup

We used the same transcription UI described in the individual-worker *baseline* study. In addition, crowd workers were told the clips were Alexa commands in both the instructions on the task interface and in the task description on Mechanical Turk.

Results

Crowd workers who had thematic context, i.e., knew the clips were of Alexa commands, had improved transcription quality over crowd workers who did not, suggesting that crowd workers were able to internalize the additional context and use it to guide their transcriptions. On average, transcriptions by crowd workers without thematic context had a WER of 0.54 ($\sigma = 0.39$) and transcriptions by crowd workers with thematic context had a WER of 0.40 ($\sigma = 0.36$), a relative improvement of 26%. An independent-samples t-test shows that this improvement in WER was significant ($t(129) = 2.11, p < .05$).

Iterative Transcription for Alexa Commands

Next we compare our iterative approach to individual crowd workers who were given thematic context (modified baseline approach), using the Alexa dataset.

Results

On average, the iterative workflow generated 58% better transcriptions than the modified baseline approach on the Alexa dataset. The iterative approach achieved the lowest average WER of 0.17 ($\sigma = 0.30$) at step 10. An independent-samples t-test showed this difference in performance was significant ($t(127) = 5.415, p < .0001$). Figure 7 shows the average WER for each approach, illustrating how average WER decreased with each iteration step.

Automated speech recognition performed poorly on the dataset of Alexa commands, with an average WER of 0.84 ($\sigma = 0.32$). There was a greater difference between transcription accuracy by ASR and individual crowd workers for Alexa commands than for the Clarke sentences. This suggests that more specific domain and inherent familiarity of the Alexa commands makes it relatively easier for humans to transcribe deaf speech. In contrast, ASR cannot distinguish between the two datasets and generates transcriptions irrespective to variations in content of the deaf speech.

Each transcription took an average of 33.4s ($\sigma = 25.6s$) per clip in the crowd-powered modified baseline. For the iterative approach, the longest average transcription time was in iteration step 2 with 45.3s, which was 37% higher than step 1 and 51% higher than steps 3-5. This increased latency may be because crowd workers in step 2 were the first to receive a set of previous crowd worker transcriptions, which were often error-prone. Workers in subsequent steps had reduced cognitive load since they received a more refined set of transcriptions.

Like the Clarke sentences dataset, crowd worker transcriptions in the iterative approach converged quickly for the Alexa commands dataset. Cosine similarity started at .77 between steps 1 and 2, and peaked at .97 between steps 8 and 9.

LIMITATIONS AND FUTURE WORK

With iterative transcription, we observed a biasing effect among crowd workers between steps. Since workers received five previous transcriptions as a guide, a priming effect was introduced that made their transcriptions closely resemble the previous steps'. While this passing of context underlies the effectiveness of iteration, it can also "trap" transcription accuracy in local maximum. Our results show that similarity between worker transcriptions increases over multiple iteration steps, suggesting that workers independently converge towards a transcription that they believe to be correct. Interestingly, while worker transcriptions converged at intelligibility level 30, the transcription quality did not improve. This shows workers were converging to the same incorrect transcription for each clip. Since it is more difficult to provide an alternate transcription after primed with a "possible" transcription, there is little incentive or guidance for workers to escape from this local quality maximum. An example was with the clip "Alexa, play music by Pearl Jam," which crowd workers in early steps transcribed as "Alexa, play music by Burn Them." Workers in subsequent steps tended to forward these incorrect transcriptions, since it would require significantly more effort to reject the transcription momentum of previous steps and synthesize a new, isolated transcription. Future work may explore how to mitigate this bias while retaining the performance benefits of iteration. One possibility is to introduce a source of randomness at each step, either by providing no previous transcriptions to a small set of "unbiased" workers, or by including ASR transcriptions at each step. Albeit error-prone, these transcriptions could provide the necessary stimulus to escape the local performance maxima.

Though we provide transcription times for the crowd-powered approaches that we study, this paper focuses more on exploring approaches that may improve transcription quality. This has an understandable tradeoff in latency: the iterative approach we used required more time than ASR for example, but produced more accurate transcriptions. Given the latency, usability of the present transcriptions are generally insufficient for real-world use in interactive systems. However, we believe future work can build on the insights that we provide to reduce transcription time and while retaining high quality.

CONCLUSION

In this paper, we have explored the problem of deaf speech recognition through a series of empirical studies. Our experiments demonstrate that individual crowd workers produced higher quality transcriptions of deaf speech than automated speech recognition, though word error rates of both approaches were too high for real-world applications (0.70 versus 0.54 WER, respectively). Results from our studies of methods to improve individual transcription performance found that modifying the speed of a clip had no significant effect on quality, audio decomposition hurt transcription quality, and providing additional thematic context when present in the domain improved transcription quality by 26%. Lastly, we evaluate a state-of-the-art crowdsourcing approach by applying an iterative crowd-powered workflow as a means of improving collective performance. We evaluated this approach on a set of general sentences (Clarke dataset), as well as a dataset of Alexa commands spoken by a DHH user, to simulate a level of context that may present in real-world settings. While iteration improved transcription quality for intelligibility levels 40 and 50 (52% and 74%, respectively), it failed to improve transcription quality for lower levels of intelligibility. In summary, we have characterized the state-of-the-art in deaf speech transcription, evaluating methodologies ranging from automated approaches to both individual and iterative crowdsourcing approaches. Our results aim to inform future approaches that further improve both quality and latency to enable more robust, accessible interactions with speech-based interfaces.

ACKNOWLEDGEMENTS

The authors would like to thank Raja Kushalnagar and Kevin Zheng for their input on and contributions to this work.

REFERENCES

1. Michael S. Bernstein, Joel Brandt, Robert C. Miller, and David R. Karger. 2011. Crowds in Two Seconds: Enabling Realtime Crowd-powered Interfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST '11)*. ACM, New York, NY, USA, 33–42. DOI: http://dx.doi.org/10.1145/2047196.2047201

2. Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. 2010. VizWiz: Nearly Real-time Answers to Visual Questions. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology (UIST '10)*. ACM, New York, NY, USA, 333–342. DOI: http://dx.doi.org/10.1145/1866029.1866080
3. Jeffrey P. Bigham, Raja Kushalnagar, Ting-Hao Kenneth Huang, Juan Pablo Flores, and Saiph Savage. 2017. On How Deaf People Might Use Speech to Control Devices. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '17)*. ACM, New York, NY, USA, 383–384. DOI:http://dx.doi.org/10.1145/3132525.3134821
4. Jeffrey P. Bigham, Richard E. Ladner, and Yevgen Borodin. 2011. The Design of Human-powered Access Technology. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '11)*. ACM, New York, NY, USA, 3–10. DOI:http://dx.doi.org/10.1145/2049536.2049540
5. Chris Callison-Burch. 2009. Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1 (EMNLP '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 286–295. http://dl.acm.org/citation.cfm?id=1699510.1699548
6. Chris Callison-Burch and Mark Dredze. 2010. Creating Speech and Language Data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1–12. http://dl.acm.org/citation.cfm?id=1866696.1866697
7. CNET. 2017. The complete list of Alexa commands so far. (2017). https://www.cnet.com/how-to/amazon-echo-the-complete-list-of-alexa-commands/
8. 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 Fifth International ACM Conference on Assistive Technologies (Assets '02)*. ACM, New York, NY, USA, 205–212. DOI: http://dx.doi.org/10.1145/638249.638287
9. Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. 2010. Are Your Participants Gaming the System?: Screening Mechanical Turk Workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 2399–2402. DOI: http://dx.doi.org/10.1145/1753326.1753688
10. Yashesh Gaur, Walter S. Lasecki, Florian Metze, and Jeffrey P. Bigham. 2016. The Effects of Automatic Speech Recognition Quality on Human Transcription Latency. In *Proceedings of the 13th Web for All Conference (W4A '16)*. ACM, New York, NY, USA, Article 23, 8 pages. DOI: http://dx.doi.org/10.1145/2899475.2899478
11. Abraham T. Glasser, Kesavan R. Kushalnagar, and Raja S. Kushalnagar. 2017. Feasibility of Using Automatic Speech Recognition with Voices of Deaf and Hard-of-Hearing Individuals. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '17)*. ACM, New York, NY, USA, 373–374. DOI: http://dx.doi.org/10.1145/3132525.3134819
12. Anhong Guo, Xiang'Anthony' Chen, Haoran Qi, Samuel White, Suman Ghosh, Chieko Asakawa, and Jeffrey P Bigham. 2016. Vizlens: A robust and interactive screen reader for interfaces in the real world. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 651–664.
13. Clarence Virginius Hudgins and Fred Cheffins Numbers. 1942. An investigation of the intelligibility of the speech of the deaf. *Genetic Psychology Monographs* 25 (1942), 289–392.
14. Saba Kawas, George Karalis, Tzu Wen, and Richard E. Ladner. 2016. Improving Real-Time Captioning Experiences for Deaf and Hard of Hearing Students. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '16)*. ACM, New York, NY, USA, 15–23. DOI: http://dx.doi.org/10.1145/2982142.2982164
15. Raja S. Kushalnagar, Walter S. Lasecki, and Jeffrey P. Bigham. 2012. A Readability Evaluation of Real-time Crowd Captions in the Classroom. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '12)*. ACM, New York, NY, USA, 71–78. DOI: http://dx.doi.org/10.1145/2384916.2384930
16. Raja S. Kushalnagar, Walter S. Lasecki, and Jeffrey P. Bigham. 2014. Accessibility Evaluation of Classroom Captions. *ACM Trans. Access. Comput.* 5, 3, Article 7 (Jan. 2014), 24 pages. DOI: http://dx.doi.org/10.1145/2543578
17. Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. 2012. Real-time captioning by groups of non-experts. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. ACM, 23–34.
18. Walter S. Lasecki, Christopher D. Miller, and Jeffrey P. Bigham. 2013. Warping Time for More Effective Real-time Crowdsourcing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 2033–2036. DOI: http://dx.doi.org/10.1145/2470654.2466269

19. Walter S. Lasecki, Christopher D. Miller, Iftekhar Naim, Raja Kushalnagar, Adam Sadilek, Daniel Gildea, and Jeffrey P. Bigham. 2017. Scribe: Deep Integration of Human and Machine Intelligence to Caption Speech in Real Time. *Commun. ACM* 60, 9 (Aug. 2017), 93–100. DOI:http://dx.doi.org/10.1145/3068663

20. Walter S Lasecki, Kyle I Murray, Samuel White, Robert C Miller, and Jeffrey P Bigham. 2011. Real-time crowd control of existing interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 23–32.

21. Walter S Lasecki, Phyo Thiha, Yu Zhong, Erin Brady, and Jeffrey P Bigham. 2013. Answering visual questions with conversational crowd assistants. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 18.

22. Beatrice Liem, Haoqi Zhang, and Yiling Chen. 2011. An Iterative Dual Pathway Structure for Speech-to-Text Transcription. In *Human Computation*.

23. Greg Little, Lydia B. Chilton, Max Goldman, and Robert C. Miller. 2010. TurKit: Human Computation Algorithms on Mechanical Turk. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology (UIST '10)*. ACM, New York, NY, USA, 57–66. DOI: http://dx.doi.org/10.1145/1866029.1866040

24. Alan Lundgard, Yiwei Yang, Maya L. Foster, and Walter S. Lasecki. 2018. Bolt: Instantaneous Crowdsourcing via Just-in-Time Training. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 467, 7 pages. DOI: http://dx.doi.org/10.1145/3173574.3174041

25. Marjorie E Magner. 1972. *A speech intelligibility test for deaf children*. Clarke School for the Deaf.

26. Nancy S. Mcgarr. 1981. The Effect of Context on the Intelligibility of Hearing and Deaf Children's Speech. *Language and Speech* 24, 3 (1981), 255–264. DOI: http://dx.doi.org/10.1177/002383098102400305

27. Nancy S. McGarr. 1983. The Intelligibility of Deaf Speech to Experienced and Inexperienced Listeners. *Journal of Speech, Language, and Hearing Research* 26, 3 (1983), 451–458. DOI: http://dx.doi.org/10.1044/jshr.2603.451

28. Ashley Miller, Joan Malasig, Brenda Castro, Vicki L. Hanson, Hugo Nicolau, and Alessandra Brandão. 2017. The Use of Smart Glasses for Lecture Comprehension by Deaf and Hard of Hearing Students. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. ACM, New York, NY, USA, 1909–1915. DOI: http://dx.doi.org/10.1145/3027063.3053117

29. R. B. Monsen. 1983. Voice Quality and Speech Intelligibility Among Deaf Children. *American Annals of the Deaf* 128, 1 (1983), 12–19.

30. Cosmin Munteanu, Gerald Penn, Ron Baecker, Elaine Toms, and David James. 2006. Measuring the acceptable word error rate of machine-generated webcast transcripts. In *Ninth International Conference on Spoken Language Processing*.

31. Mumtaz Begum Mustafa, Fadhilah Rosdi, Siti Salwah Salim, and Muhammad Umair Mughal. 2015. Exploring the influence of general and specific factors on the recognition accuracy of an ASR system for dysarthric speaker. *Expert Systems with Applications* 42, 8 (2015), 3924 – 3932. DOI:http://dx.doi.org/https://doi.org/10.1016/j.eswa.2015.01.033

32. Mary Joe Osberger and Harry Levitt. 1979. The effect of timing errors on the intelligibility of deaf children's speech. *The Journal of the Acoustical Society of America* 66, 5 (nov 1979), 1316–1324. DOI: http://dx.doi.org/10.1121/1.383552

33. Mary Joe Osberger and Nancy S. McGarr. 1982. Speech Production Characteristics of the Hearing Impaired. Speech and Language, Vol. 8. Elsevier, 221 – 283. http://www.sciencedirect.com/science/article/pii/B9780126086089500139

34. R. Ranchal, T. Taber-Doughty, Y. Guo, K. Bain, H. Martin, J. Paul Robinson, and B. S. Duerstock. 2013. Using speech recognition for real-time captioning and lecture transcription in the classroom. *IEEE Transactions on Learning Technologies* 6, 4 (Oct 2013), 299–311. DOI: http://dx.doi.org/10.1109/TLT.2013.21

35. Seyed Reza Shahamiri and Sayan Kumar Ray. 2015. On the use of array learners towards Automatic Speech Recognition for dysarthria. In *Industrial Electronics and Applications (ICIEA), 2015 IEEE 10th Conference on*. IEEE, 1283–1287.

36. R Sriranjani, M Ramasubba Reddy, and S Umesh. 2015. Improved acoustic modeling for automatic dysarthric speech recognition. In *Communications (NCC), 2015 Twenty First National Conference on*. IEEE, 1–6.

37. Blake S. Wilson, Charles C. Finley, Dewey T. Lawson, Robert D. Wolford, Donald K. Eddington, and William M. Rabinowitz. 1991. Better speech recognition with cochlear implants. *Nature* 352, 6332 (18 July 1991), 236–238. DOI:http://dx.doi.org/10.1038/352236a0

38. Yu Zhong, Walter S. Lasecki, Erin Brady, and Jeffrey P. Bigham. 2015. RegionSpeak: Quick Comprehensive Spatial Descriptions of Complex Images for Blind Users. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 2353–2362. DOI: http://dx.doi.org/10.1145/2702123.2702437