The Effects of Automatic Speech Recognition Quality on Human Transcription Latency

Yashesh Gaur
Language Technologies Institute
Carnegie Mellon University
yashesh@cs.cmu.edu

Walter S. Lasecki
Computer Science & Engineering
University of Michigan, Ann Arbor
wlasecki@umich.edu

Florian Metze
Language Technologies Institute
Carnegie Mellon University
fmetze@cs.cmu.edu

Jeffrey P. Bigham
HCI and LT Institutes
Carnegie Mellon University
jbigham@cs.cmu.edu

ABSTRACT

Transcription makes speech accessible to deaf and hard of hearing people. This conversion of speech to text is still done manually by humans, despite high cost, because the quality of automated speech recognition (ASR) is still too low in real-world settings. Manual conversion can require more than 5 times the original audio time, which also introduces significant latency. Giving transcriptionists ASR output as a starting point seems like a reasonable approach to making humans more efficient and thereby reducing this cost, but the effectiveness of this approach is clearly related to the quality of the speech recognition output. At high error rates, fixing inaccurate speech recognition output may take longer than producing the transcription from scratch, and transcriptionists may not realize when transcription output is too inaccurate to be useful. In this paper, we empirically explore how the latency of transcriptions created by participants recruited on Amazon Mechanical Turk vary based on the accuracy of speech recognition output. We present results from 2 studies which indicate that starting with the ASR output is worse unless it is sufficiently accurate (Word Error Rate of under 30%).

Categories and Subject Descriptors

K.4.2 [**Computers and Society**]: Social Issues—*Assistive technologies for persons with disabilities*

Keywords

Captioning, Human Computation, Automatic Speech Recognition, Crowd Programming.

W4A'16, April 11-13, 2016, Montreal, Canada

DOI: http://dx.doi.org/10.1145/2899475.2899478

1. INTRODUCTION

Audio captions provide access to aural content for people who are deaf and hard of hearing in a wide range of different domains, from classroom lectures to public speeches to entertainment, such as movies and television. Unfortunately, producing accurate captions automatically is not yet possible. As a result, human computation has become a popular approach to generating captions [8]. On crowdsourcing platforms like Amazon's Mechanical Turk, Crowdflower, Crowdsource.org, and more, transcription tasks are one of the most commonly available types of task. Several successful approaches for organizing people to complete this task have been proposed for both the online and offline settings [1, 19]. However, while automated systems alone cannot provide accurate captions in *all* settings, decades of work on automatic speech recognition (ASR) has resulted in systems that can effectively handle a variety of audio content with reasonable accuracy [13, 28, 31]. This paper explores how mixing these two sources of captions can be traded off by studying the effect of error rates on workers' ability to edit partially-correct ASR captions.

To *generate* captions from scratch, workers must hear the audio content and convert this content to text by typing. It is important to note that, often, this can be done with minimal cognizant processing of the content or meaning of the speech itself. *Editing* content generated by an ASR system requires less direct motor effort (typing), but requires more cognitive activities involved in understanding the text relative to the speech that is being listened to in order to identify and correct mistakes. We expect this difficulty to rise as the word error rate (WER) of automatically generated captions increases, because the difference between the audio and caption requires more complex mental mapping.

Figure 1 shows different quality ASR transcripts for a small clip from a TED talk [30]. It also shows the deletions (red strike-through) and insertions (green words) steps that the captionist must perform on these ASR transcripts to match them to the words in the speech. We observe that for a transcript with WER of 15%, very few insertions/ deletions are required, while for a transcript with WER of 45%, a lot of effort will be required in deciding which parts of the ASR transcripts are to be deleted and where the new words are to be added. In addition to a more complex map-

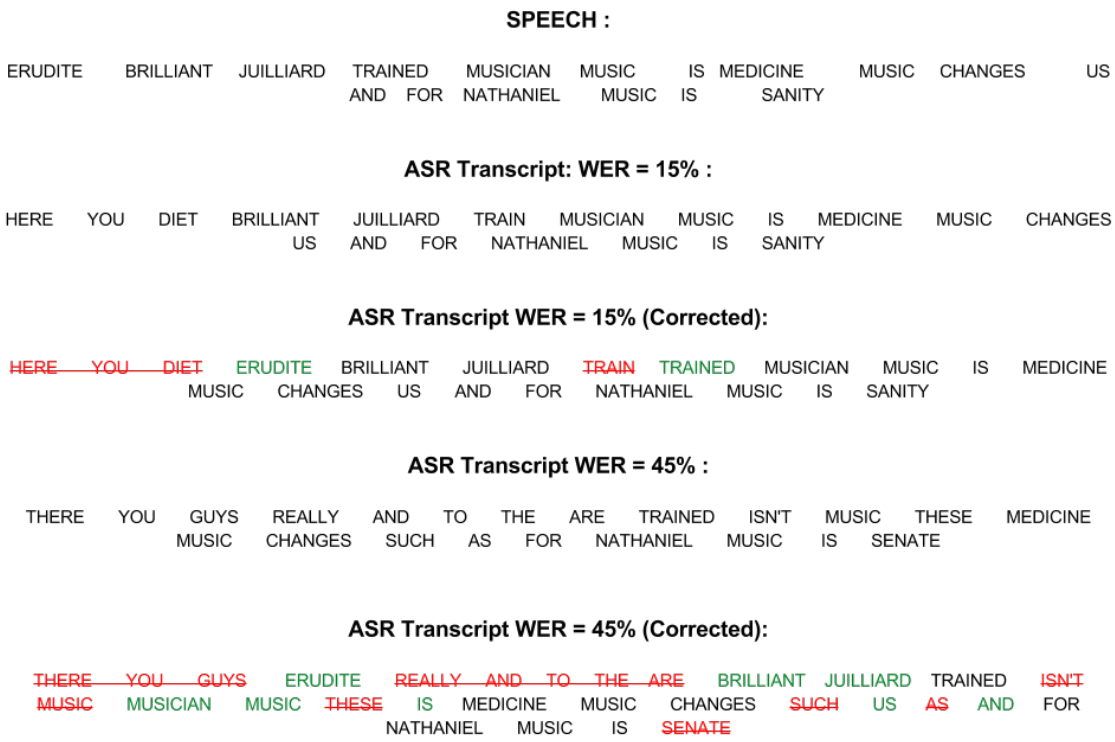

Figure 1: A typical 10 second utterance from a TED talk with corresponding ASR generated transcripts of different Word Error Rates (WERs). These clips with different WERs were created in an ecologically valid way by modifying the beam width of the ASR, which means the errors are similar to what we should see in practice. A lower WER means higher quality. Possible correction steps to convert the erroneous transcripts into the ground truth (at the word level) are also shown. Red (crossed out) words need to "deleted" and green words need to be "inserted" into the ASR transcript. "Substitution" errors are corrected as a sequence of "deletions" and "insertions". While transcripts with low WER can be easily corrected, correcting a transcript with high WER turns out to be time consuming. If WER is high enough, it can be better to transcribe from scratch.

ping from the ASR transcript to the ground truth, increasing WERs also make it difficult to *spot* mistakes. This is because most ASRs are phoneme based and consequently, the incorrect words are often homonyms of the correct words, which might trick the captionists into believing they are correct. In such a scenario, it would seem that the ASR transcripts have stopped being useful and the captionists are better off transcribing everything from scratch, without any ASR support.

To explore the effects of ASR quality on captionist performance, we designed and conducted 2 studies with participants on Amazon's Mechanical Turk. Importantly, we modified the ASR directly to control its WER, which provided an ecologically valid set of clips with different WERs. For each WER, we compare starting with the ASR output to simply generating captions for the same audio segment from scratch.

We find that while editing ASR output is significantly faster when the accuracy is high enough, editing latency quickly rises above the level at which workers can accurately generate the content themselves. If the WER is high enough (45% in our study), then workers identify how bad it is, erase the text generated from the ASR, and simply type from scratch.

Our work directly speaks to the likelihood of success in prior systems in which people are tasked with correcting the output of ASR [16, 9, 15]. Furthermore, it gives practictioners a cut-off for expected WER after which they should simply have people type from scratch.

Our contributions are:

- an ecologically valid result for the captioning task explored in this paper that might guide practictioners
- insight into how captionists interact with error-prone ASR, suggesting that they are capable of detecting WERs that are too difficult for them to correct

2. RELATED WORK

Our work is related to *(i)* automatic speech recognition, and *(ii)* human computation and crowdsourcing.

2.1 Automatic Speech Recognition

Automatic Speech Recognition aims to generate text from human speech with the help of computers. It has been a topic of research for quite some time now and has seen huge improvements in the past few decades [28]. Consequently, it has evolved from desktop recognition in ideal recording conditions, to mobile based recognition in real world settings [14]. Most ASR systems in use today are statistical systems trained on huge amounts of audio and corresponding

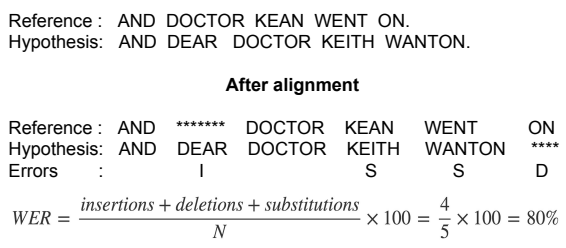

Figure 2: Calculation of Word Error Rate. Hypothesis is the word sequence generate by the speech recognizer while reference is the word sequence present in the speech. First, the reference and hypothesis are aligned using dynamic alignment. This alignment allows us to calculate the insertions (I), deletions (D) and substitutions (S) errors. Word Error Rate is calculated by taking ratio of the sum of these errors with the total number of words in the reference.

texts. These systems learn a mapping from the audio signal to phonemes, words and sentences. Modern state-of-art ASR systems have shown impressive performance on established datasets [31], can be adapted rapidly to new tasks [11], and can process live speech from mobile devices [25]. Despite this, ASR tends to not give satisfactory results in practical situations when no or only few resources (Human effort, computation, and data) is available for development and tuning, which prevents its use as a crucial aid for the deaf and hard of hearing.

The most commonly used metric for measuring the performance of a speech recognition system is Word Error Rate (WER). Figure 2 depicts how WER is calculated. The word sequence that is put out by the speech recognizer (*hypothesis*) can have a different length from the *reference* word sequence (assumed to be the correct one). To calculate WER, first the reference is aligned with the hypothesis using dynamic programming. This alignment is not unique, but allows us to calculate the minimum number of insertions, deletions and substitution operations that need to be performed on the *reference* in order to turn it into *hypothesis*. The insertions, deletions and substitution are treated as errors and the ratio of the sum of these errors to the number of words in the reference gives us WER. This kind of a metric, however, does not provide any information about types of errors that are made and therefore, further error analysis is required to focus research efforts. It is generally believed that lower WER corresponds to better performance in a given application. However, [35] shows that this may not always be true.

One of the reasons for ASR's poor performance in use cases like captioning is the constraint of producing real-time output. This allows the ASR to work with only small models and it tends to produce sub-optimal output. Poor acoustics [11] at the venue, use of unadapted models, out of vocabulary words (words not seen by the system before) and bootstrapping a recognizer for a new domain are other major causes for poor performance, which are hard for non-expert users of ASR to overcome [18, 17]. Natural language understanding, which is at the core of ASR technology, has been categorized as an AI-complete problem [32]. It is considered that AI-complete problems can not be completely solved by present day computer technology alone, but would require human computation as well.

2.2 Human Computation and Crowdsourcing

Human computation [33] is the process of engaging people in computational processes, typically to solve problems that computers cannot yet solve alone, such as those involving understanding image or audio content [2]. Human-powered systems have become more feasible and impactful in recent years due to the rise of readily-available workers on crowdsourcing platforms, who can be recruited within a matter of seconds [23, 3, 7].

The accessibility community has long engaged support from other people to help provide accommodations [4, 5]. Crowdsourcing for human computation allows this notion to be expanded beyond one's local community, to online platforms of people who help both out of generosity and to earn a living [6]. People with disabilities have also found advantages working on crowd platforms, such as the ability to work from home, when they are able, etc. [36]

To generate captions even in settings where ASR cannot do so reliably, recent work has leveraged human computation to provide high quality transcription [10, 12, 24, 22, 20]. Large online crowdsourcing platforms such as Amazon's Mechanical Turk [1] provide general purpose platforms for human computation, and see tens of thousands of tasks related to transcription posted. Due to its scale, some crowdsourcing platforms even have special-purpose portals for transcription, such as Crowdsource [2], and other platforms have arisen as special-purpose crowdsourcing services for transcription, such as Casting Words [3].

Research in this area has sought to improve the efficiency, accuracy, and scope of these crowd-powered methods. For instance, Liem *et al.* [1] use an iterative, "dual pathway" approach to generate and evaluate worker input based on the input of others. This resulted in captions with over 96% accuracy.

Scribe [19] goes beyond offline captioning services and allows even non-expert typists to provide text captions with less than a five second latency, where years of training would have been required in the past. By synchronously coordinating groups of workers, Scribe's work-flow allowed for new approaches to making workers' individual tasks easier to accomplish, such as decreasing the audio playback speed without increasing latency [21].

[12] has looked at how crowd workers can be used to correct the captions directly and [34] demonstrate a tool that facilitates crowdsourcing correction of speech recognition captioning errors. However, to the best of our knowledge, there has been no study till now that explores whether or when it actually beneficial to manually edit ASR's output.

3. METHODOLOGY

We conducted two studies to explore the importance of

[1]https://mturk.com
[2]http://www.crowdsource.com/solutions/transcription/
[3]https://castingwords.com/

ASR quality on its utility as a starting point for human transcription. The first study used a between-subjects design, while the second study used a within-subjects design. Each of these studies had two conditions. The first condition asked the participants to enter transcriptions for the audio, with no ASR output provided. This time measurement served as a baseline for the second condition, where participants were asked to edit ASR output of varying error rates, instead of needing to write it from scratch.

3.1 Apparatus

To facilitate data collection, we developed a web-based captioning platform (Figure 3). This interface consisted of an audio player and a text box into which workers could type their transcription. The web page was also instrumented to record the interactions that participants had with the interface, including click behavior and keystroke dynamics. We also tracked how the participants navigated the audio during the study. To make audio navigation more convenient, we provided key shortcuts that would allow the participants to play/pause or go back 5 seconds in the audio. This allowed the workers to navigate the audio without needing to directly interact with the audio tag's play/pause button or progress bar. Shortcuts keys for audio navigation also make our time measurements less noisy, as the participants can now control the audio without using the mouse and keep their hands on the keyboard all the time. When a participant submits the transcription, we compare it against the reference transcripts. The transcription is accepted only if it is sufficiently accurate, otherwise they are asked to continue to improve it further. This mechanism ensures that our time measurements compare similar-quality transcriptions.

3.2 Data set and Preparation

For this study, we used the TEDLIUM [30] dataset, which is a collection of TED talks. The popular Kaldi speech recognition toolkit [29] was used for generating ASR transcriptions of different quality. We have worked extensively with both the TEDLIUM dataset and Kaldi, and currently achieve a baseline WER performance of 15%.

ASR transcripts of varying quality were generated by changing the 'beam width' parameter in the speech decoder. For the purpose of this study, 16 one minute long clips were selected randomly from the TEDLIUM development test set. While relatively short, these clips are long enough that participants cannot simply remember what is being said. They will also likely lose context and need to search within the audio stream to find their place. We decoded this set of 16 clips 9 times, with different 'beam width' parameter resulting in real ASR transcripts with WER ranging from 15% to 55% at rough intervals of 5%. We believe that this range is realistic and ecologically valid: the best real-time systems on standard English tasks such as TEDLIUM or Switchboard perform at about 15% WER [29, 27], while 35% can be achieved on distant speech [11], and 55% WER is still the reality for new domains or low resource languages [26].

3.3 Study Procedures

Participants were first asked to carefully read the instructions, which summarized the purpose and structure of our study. The shortcut keys designed for convenient audio navigation were described to them both in text and through a visual diagram. Owing to the time sensitive nature of our study, participants were given the opportunity to start the experiment on their own accord. The audio and the text box were only accessible after the participant agreed to start the experiment. Once the participant submits their transcription, the systems compares the input with the reference. The worker's response was accepted only if it had a WER less than 10% as compared to the dataset's ground truth. We adopted this qualifying step to discourage spammers on Mechanical Turk and also get time measurements that correspond to high-accuracy transcriptions. A qualifying WER threshold of 0% was not chosen to allow for minor errors by the workers. Threshold of 10% WER was decided based on our initial trials of the study which indicated that threshold values lesser than 10% resulted in too many worker submissions being rejected while with greater threshold values, risk of accepting low quality transcriptions was foreseen.

Table 1: Between subjects design

Types of Tasks	Number of tasks per clip	Number of clips	Number of Tasks
from-Scratch	1	16	160
Editing ASR captions of varying WER (15%-55%)	9		

Table 2: Within subjects design

Types of Tasks	Number of tasks per participant	Total number of participants
from-Scratch	1	16
Editing ASR captions of varying WER (15%-55%)	9	
Control Task	1	

3.3.1 Between Subjects Study

Our between-subjects study, summarized in Table 1, was conducted on Amazon's Mechanical Turk with 160 participants. As mentioned in Section 3.2, we conducted our study on 16 randomly chosen one-minute clips from the TEDLIUM dataset. For each clip we have 9 real ASR transcripts with WER ranging from 15% to 55%. This resulted in 9 "editASR" tasks per clip where the participants had to correct ASR generated captions. For each clip, we also had one "from-Scratch" task, where the participant had to enter the transcription for the audio without any ASR support. Consequently, we had 10 tasks for each of the 16 clips, resulting in a total of 160 tasks each of which were assigned to distinct crowd participants.

3.3.2 Within Subjects Study

The within-subjects study (Table 2) was conducted on Amazon's Mechanical Turk with 16 participants. Latin square design was used to control for variation across transcription speeds. Each participant did 9 "editASR" tasks over the complete spectrum of WER (15% - 55%), one "fromScratch" task and one "control-task". The "control task" was to transcribe a minute long control clip without any ASR support.

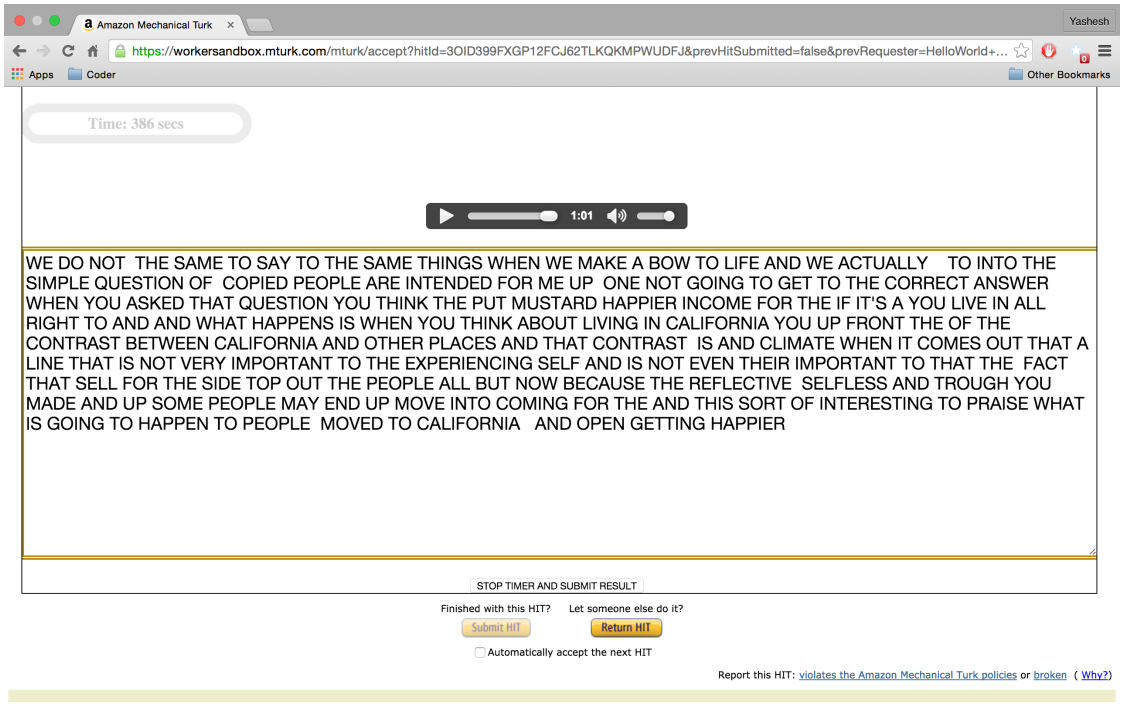

Figure 3: The transcription interface used in our study. The text-box is pre-populated with ASR captions. The participants were were previously informed of the shortcut keys that they could use to control the audio playback using both text and visual diagram.

This control clip was kept common across all the participants. Participant's latency measurement from the control task was used to normalize their latency measurements obtained for other tasks. This normalization enabled us to compare latencies across participants with potentially different transcription speeds. We also ensured that the participants worked on a different audio clip for every tasks to safeguard against any potential learning effects. The 9 editASR tasks were presented in a randomized order to the participants.

4. RESULTS

In this section we describe the findings from our study.

4.1 Between Subjects Study

Figure 4 shows the results from our experiments with one-minute clips for our between study. By the time the ASR transcripts exceed 30% WER, workers are again faster just typing the content from scratch. This matches our results from the five-second clips, and suggests that not all automated assistance is helpful when captioning audio. Note, however, that after the 50% WER mark, the average latency actually begins to *decrease* relative to the maximum seen at approximately 45% WER.

To answer the question of why this initially unexpected decrease in latency occurs, we looked at the logs that tracked how workers interacted with our captioning task. Specifically, one aspect of the interaction we logged was the length of the current transcript that the worker was editing. Change of 1-2 words indicate typical editing of a document or splitting of an existing word, while more represents multi-word pasting or removal of text.

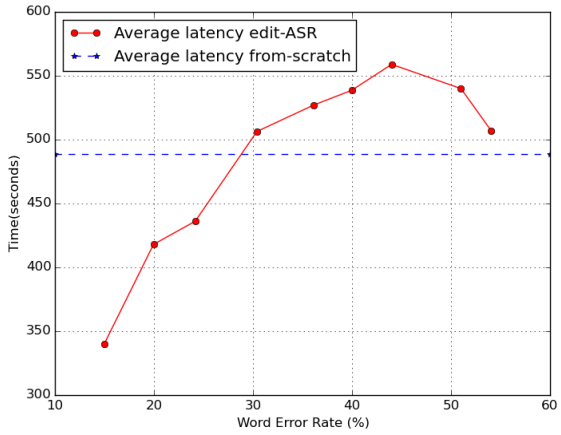

Figure 4: Average latency vs. WER for the between subjects design. Latency increases along with the WER until it surpasses the from-scratch latency at about 30 WER. After about 45% workers realize that the ASR captions are not helpful anymore and start deleting big chunks of ASR transcripts. Hence the latecnies at higher WERs decrease and start approaching the from-scratch measurements.

An important component of our setup is that workers are provided with an ASR transcript to edit, however, there is no guarantee that they actually *do* edit the existing text. They could, instead, remove the provided text from the text-

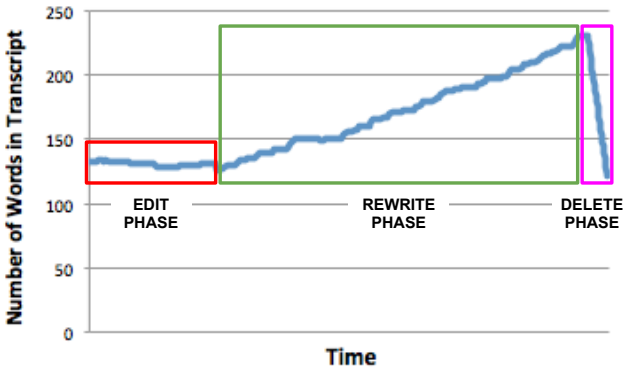

Figure 5: Word count log from one worker session in our trials. At first, the worker changes words or replaces them, resulting in the word count remaining mostly stable. Next, the worker begins to add words to the transcript, without removing or replacing others. This is essentially a rewrite step. Finally, the worker removes the incorrect content (about 50% of the total words at that point), and ends in a transcript with under 10% WER. A similar and equivalent pattern can be seen when workers first delete all content, then rewrite it based on the audio they heard – this results in a mirrored version of this plot.

box and then enter their own from scratch. These types of 'clearing' behaviors can appear in multiple forms. One example would be to first clear the text then rewrite it, while another option would be to write the text then clear the initial content once done. Figure 5 shows an example of the latter type of clearing. The actual clear event might be preceded by initial in-place edits, which appear as changes of few or no words in the logs.

To see if this behaviors could be responsible for the decrease in latency as workers see higher levels of initial error in the text, we compared the higher and lower half of our WER range. Figure 6 shows a side-by-side comparison of the word count traces over time.

Looking at the number of occurrences of workers removing more than 25% of the transcript in a given time step, we observe that for WERs under 30%, just 7.12% of workers cleared a large portion of text. For WERs under 50%, this ratio increases to 12.28%. For WERs of over 50% (where the latencies begin to decrease from their maximum), we see that 42.52% of workers clear the text they were given. This very large jump may explain why workers are less slowed down by increased error rates as the overwhelming number of errors in a given transcript becomes more evident up front.

4.2 Within Subjects Study

Figure 7 shows the variation of averaged normalized latencies for "editASR" tasks corresponding to different word error rates. We normalize the latency measurements of participants with their latency measurements on the control clip. This takes care of any variation that may arise because of different transcription speeds between participants. We again see that the participants benefit from ASR captions if the word error rate performance is low ($< 30\%$). Similar to the first study, we observe that ASR captions stop

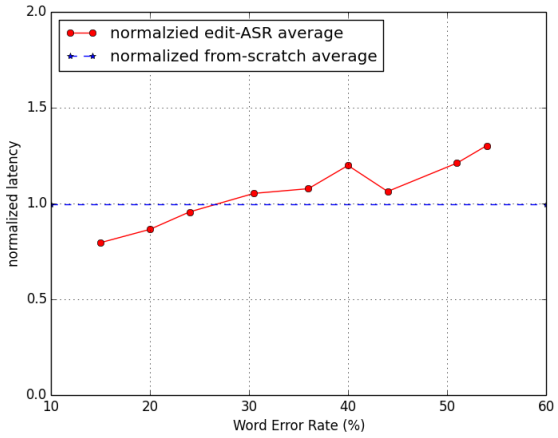

Figure 7: Normalized average latency vs. WER for within subjects design. Latency measurements for each participant were normalized with their latency measurement on the control clip. This allowed us to compare results across participants without worrying about any variation that may arise because of different transcription speeds. Notice that ASR captions are helpful only when WER $< 30\%$.

becoming helpful at word error rates around 30%. Above 30% WER, participants take more time correcting the ASR captions than writing them "from scratch" without any ASR support. As opposed to the first study, here we do not observe any drop in latencies with increasing WERs. The curve is generally non-decreasing, except for a downward "kink" at around WER of 40%. One potential explanation for this behavior can be attributed to the fact that every participant is exposed to all the levels of word error rates. Being exposed to good transcripts and having seen that they can be helpful, the participants do not give up on bad transcripts soon enough. A one way ANOVA performed on the normalized data from this study yielded a value of 0.01, which suggests that this increasing trend is significant.

5. DISCUSSION

Both studies show that speech recognition can save captionists' time if it is accurate enough. More specifically, we observe that there exists a threshold around 30% WER. If the WER of the speech recognition system is much greater than 30%, humans are better off writing everything from scratch rather than getting help from automatic speech recognition. While we expect that the specific WER will vary depending on such variables as the content of the speech, the speech rate, and the individual captionist's typing speed, establishing this average rate may help drive the design of future technologies to improve captioning.

One challenge for workers was recognizing the quality (and a resulting best transcription strategy). In Figure 4, we observe that even though automatically generated transcripts stop becoming useful after WER of 30%, the workers are only able to realize that after 45%. Automated approaches could help not just by providing accurate transcriptions as a starting point, but may also be usefully employed to understand what part of the transcription is accurate enough to

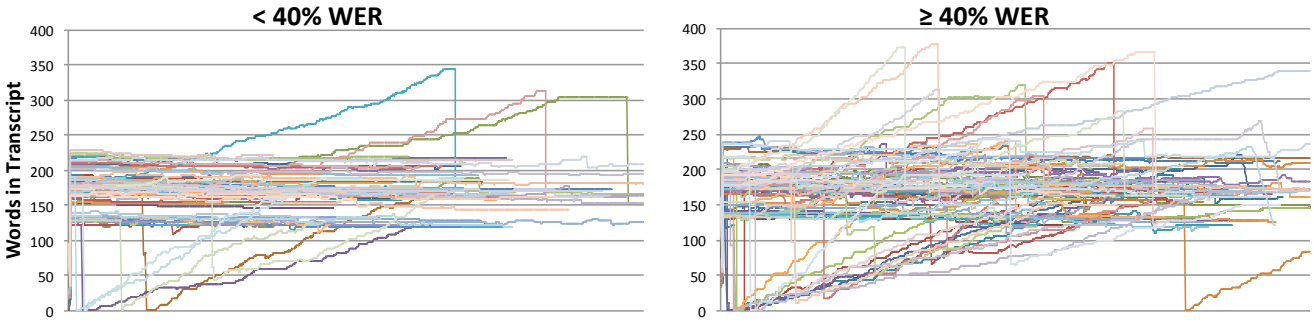

Figure 6: 144 participants' word count plots. On the left, the lower WER half of our experiments, and on the right, the higher WER half. The drastic increase in workers who remove large blocks of text (sudden drops in word count when a worker deletes the initial ASR transcript they were given, as shown in Figure 5) is tied to increased WER of the initial ASR transcript provided.

be kept as a starting point. This is challenging because confidence estimates output by ASR technology are unreliable, and ASR quality can vary considerably over short spans.

However, even if we cannot automatically detect when these variations occur, our results show that given high error rates, workers can identify when it would be easier to just work from a blank slate, and avoid the inefficiencies associated with large sets of corrections.

6. CONCLUSION AND FUTURE WORK

In this paper we explored the effects of ASR quality on its utility as a starting point for human transcription. Our results match our expectations that ASR is most useful as a starting point when it is fairly accurate. By examining worker behavior, we identified common strategies that workers use with transcription starting points of different qualities. Insight into how workers correct transcript will enable us to design captioning interface to aid their productivity. For example, captioning interfaces might try to "detect" when the WER increases above a certain threshold and stop providing ASR support altogether. This will be a challenge as "ground truth" is generally needed to calculate WER. However, ASR also generates confidence measures for every word is puts out. This suggests that future research should focus on predicting WER from the confidence scores generated by ASR. The same type of thresholds we define here may be definable in terms of confidence measures, instead of WER. Unfortunately the quality of a recognizer's word confidence measures can vary even more than the quality of the words themselves.

Our findings also suggest new methods for integrating high-error ASR output with human input. In our study, we have shown that ASR can get most of the words correct and still hurt performance. Future work will explore how we can design and develop captioning interfaces that may make error prone ASR more helpful. For example, using ASR to suggest words (like an auto-complete function), so that the worker does not have to correct erroneous transcripts but still some helpful information/words can be suggested. Being able to trade off human and machine ability, and how one affects the other, is a key aspect of effective hybrid intelligence systems, and can greatly benefit human-powered access technology.

7. ACKNOWLEDGMENTS

This research was supported by National Science Foundation awards #IIS-1218209 and #IIS-1149709, a Sloan Fellowship, and the National Institute on Disability Independent Living Rehabilitation Research.

8. REFERENCES

[1] Y. C. Beatrice Liem, Haoqi Zhang. An iterative dual pathway structure for speech-to-text transcription. In *Proceedings of the 3rd Workshop on Human Computation (HCOMP '11)*, HCOMP '11, 2011.

[2] J. P. Bigham, M. Bernstein, and E. Adar. Human-computer interaction and collective intelligence, 2015.

[3] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, and T. Yeh. Vizwiz: Nearly real-time answers to visual questions. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pages 333–342, New York, NY, USA, 2010. ACM.

[4] J. P. Bigham, R. E. Ladner, and Y. Borodin. The design of human-powered access technology. In *Proc. of Computers and Accessibility*, ASSETS '11, pages 3–10, 2011.

[5] E. Brady and J. Bigham. Crowdsourcing accessibility: Human-powered access technologies. *Foundations and Trends in Human-Computer Interaction*, 8(4):273–372, 2015.

[6] E. Brady, M. R. Morris, and J. P. Bigham. Gauging receptiveness to social microvolunteering. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '15, New York, NY, USA, 2015. ACM.

[7] M. A. Burton, E. Brady, R. Brewer, C. Neylan, J. P. Bigham, and A. Hurst. Crowdsourcing subjective fashion advice using vizwiz: Challenges and opportunities. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '12, pages 135–142, New York, NY, USA, 2012. ACM.

[8] C. Callison-Burch and M. Dredze. 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, pages 1–12, Stroudsburg, PA; U.S.A., 2010. Association for Computational Linguistics.

[9] R. Dufour and Y. Esteve. Correcting asr outputs: Specific solutions to specific errors in french. In *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*, pages 213–216, Dec 2008.

[10] Y. Gaur, F. Metze, Y. Miao, and J. P. Bigham. Using keyword spotting to help humans correct captioning faster. In *Proc. INTERSPEECH*, Dresden, Germany, Sept. 2015. ISCA.

[11] M. Harper. The automatic speech recognition in reverberant environments (ASpIRE) challenge. In *Proc. ASRU*, Scottsdale, AZ; U.S.A., Dec. 2015. IEEE.

[12] R. P. Harrington and G. C. Vanderheiden. Crowd caption correction (ccc). In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '13, pages 45:1–45:2, New York, NY, USA, 2013. ACM.

[13] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

[14] X. Huang, J. Baker, and R. Reddy. A historical perspective of speech recognition. *Communications of the ACM*, 57(1):94–103, 2014.

[15] D. Huggins-Daines and A. I. Rudnicky. Interactive asr error correction for touchscreen devices. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, HLT-Demonstrations '08, pages 17–19, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[16] H. Kolkhorst, K. Kilgour, S. Stüker, and A. Waibel. Evaluation of interactive user corrections for lecture transcription. In *Proc. IWSLT*, pages 217–221, 2012.

[17] A. Kumar, F. Metze, and M. Kam. Enabling the rapid development and adoption of speech-user interfaces. *IEEE Computer Magazine*, 46(1), Jan. 2014.

[18] A. Kumar, F. Metze, W. Wang, and M. Kam. Formalizing expert knowledge for developing accurate speech recognizers. In *Proc. INTERSPEECH*, Lyon; France, Sept. 2013. ISCA.

[19] W. Lasecki, C. Miller, A. Sadilek, A. Abumoussa, D. Borrello, R. Kushalnagar, and J. Bigham. Real-time captioning by groups of non-experts. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*, UIST '12, pages 23–34, New York, NY, USA, 2012. ACM.

[20] W. S. Lasecki and J. P. Bigham. Online quality control for real-time crowd captioning. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '12, pages 143–150, New York, NY, USA, 2012. ACM.

[21] W. S. Lasecki, C. D. Miller, and J. P. Bigham. Warping time for more effective real-time crowdsourcing. In *Proc. SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2033–2036, New York, NY, USA, 2013. ACM.

[22] W. S. Lasecki, C. D. Miller, R. Kushalnagar, and J. P. Bigham. Legion scribe: Real-time captioning by the non-experts. In *Proc. 10th Int. Cross-Disciplinary Conference on Web Accessibility*, W4A '13, pages 22:1–22:2, New York, NY, USA, 2013. ACM.

[23] W. S. Lasecki, P. Thiha, Y. Zhong, E. Brady, and J. P. Bigham. Answering visual questions with conversational crowd assistants. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '13, pages 18:1–18:8, New York, NY, USA, 2013. ACM.

[24] C.-y. Lee and J. R. Glass. A transcription task for crowdsourcing with automatic quality control. Florence; Italy, Aug. 2011. ISCA.

[25] X. Lei, A. Senior, A. Gruenstein, and J. Sorensen. Accurate and compact large vocabulary speech recognition on mobile devices. In *INTERSPEECH*, pages 662–665, 2013.

[26] F. Metze, A. Gandhe, Y. Miao, Z. Sheikh, Y. Wang, D. Xu, H. Zhang, J. Kim, I. Lane, W. K. L. ee, S. Stüker, and M. Müller. Semi-supervised training in low-resource ASR and KWS. In *Proc. ICASSP*, Brisbane; Australia, Apr. 2015. IEEE.

[27] Y. Miao, M. Gowayyed, and F. Metze. EESEN: End-to-End Speech Recognition using Deep RNN Models and WFST-based Decoding. In *Proc. ASRU*, Scottsdale, AZ; U.S.A., Dec. 2015. IEEE. https://github.com/srvk/eesen.

[28] R. K. Moore. Progress and prospects for speech technology: Results from three sexennial surveys. In *INTERSPEECH*, pages 1533–1536, 2011.

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, et al. The Kaldi speech recognition toolkit. In *Proc. ASRU*, Hawaii, HI; U.S.A., Dec. 2011. IEEE.

[30] A. Rousseau, P. Deléglise, and Y. Estève. Ted-lium: an automatic speech recognition dedicated corpus. In *Proc. LREC*, pages 125–129, 2012.

[31] G. Saon and J.-T. Chien. Large-vocabulary continuous speech recognition systems: A look at some recent advances. *Signal Processing Magazine, IEEE*, 29(6):18–33, Nov 2012.

[32] S. C. Shapiro. *ENCYCLOPEDIA OF ARTIFICIAL INTELLIGENCE SECOND EDITION*. New Jersey: A Wiley Interscience Publication, 1992.

[33] L. von Ahn. Human computation. In *Design Automation Conference, 2009. DAC '09. 46th ACM/IEEE*, pages 418–419, July 2009.

[34] M. Wald. Crowdsourcing correction of speech recognition captioning errors. 2011.

[35] Y.-Y. Wang, A. Acero, and C. Chelba. Is word error rate a good indicator for spoken language understanding accuracy. In *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pages 577–582. IEEE, 2003.

[36] K. Zyskowski, M. R. Morris, J. P. Bigham, M. L. Gray, and S. Kane. Accessible crowdwork? understanding the value in and challenge of microtask employment for people with disabilities. ACM, March 2015.