A Readability Evaluation of Real-Time Crowd Captions in the Classroom

Raja S. Kushalnagar, Walter S. Lasecki†, Jeffrey P. Bigham†

Department of Information and Computing Studies
Rochester Institute of Technology
1 Lomb Memorial Dr, Rochester, NY 14623
rskics@rit.edu

†Department of Computer Science
University of Rochester
160 Trustee Rd, Rochester, NY 14627
{wlasecki, jbigham}@cs.rochester.edu

ABSTRACT

Deaf and hard of hearing individuals need accommodations that transform aural to visual information, such as transcripts generated in real-time to enhance their access to spoken information in lectures and other live events. Professional captionists's transcripts work well in general events such as community, administrative or legal meetings, but is often perceived as not readable enough in specialized content events such as higher education classrooms. Professional captionists with experience in specialized content areas are scarce and expensive. Commercial automatic speech recognition (ASR) software transcripts are far cheaper, but is often perceived as unreadable due to ASR's sensitivity to accents, background noise and slow response time. We evaluate the readability of a new crowd captioning approach in which captions are typed collaboratively by classmates into a system that aligns and merges the multiple incomplete caption streams into a single, comprehensive real-time transcript. Our study asked 48 deaf and hearing readers to evaluate transcripts produced by a professional captionist, automatic speech recognition software and crowd captioning software respectively and found the readers preferred crowd captions over professional captions and ASR.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems; K.4.2 [**Social Issues**]: Assistive technologies for persons with disabilities

General Terms

Human Factors, Design, Experimentation

Keywords

Accessible Technology, Educational Technology, Deaf and Hard of Hearing Users

ASSETS'12, October 22–24, 2012, Boulder, Colorado, USA.

1. INTRODUCTION

Deaf and hard of hearing (DHH) individuals typically cannot understand audio alone, and access to the audio through accommodations that translate the auditory information to visual information. The most common accommodations are real-time transcription or sign language translation of the audio.

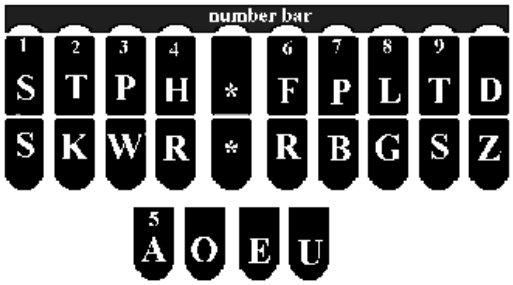
As a low incidence disability, deaf and hard of hearing individuals are evenly and thinly spread [18]. As a result, many DHH individuals tend to be located far from major population centers and find it hard to obtain accommodation providers, especially those who can handle situations that require specialized content knowledge. These providers prefer to live in close to areas where they can obtain enough demand to provide services. If there is not enough demand for providers in the area, there is a catch-22 for the DHH students and institutions. Therefore, for many institutions in terms of content knowledge, availability and cost, it is best to use accommodation services centered on the student such as classmates or on-demand remote workers.

This paper analyzes the readability of a new student-centered approach to real-time captioning in which multiple classmates simultaneously caption speech in real-time. Although classmates cannot type as quickly as the natural speaking rate of most speakers, we have found that they can provide accurate partial captions. We align and merge the multiple incomplete caption streams into a single, comprehensive real-time transcript. We compare deaf and hearing students' evaluation of the effectiveness and usability of this crowd-sourced real-time transcript against transcripts produced by professional captionists and automatic speech recognition software respectively.

2. BACKGROUND

Equal access to communication is fundamental to students' academic success, but is often taken for granted. In mainstream environments where deaf, hard-of-hearing, and hearing students study and attend classes together, people tend to assume that captioners or interpreters enable full communication between deaf and hearing people in the class. This assumption is especially detrimental as it does not address other information accessibility issues such as translation delays that impact interaction and readability that impacts comprehension.

There are two popular approaches to generating real-time captions that attempt to convey every spoken word in the classroom: professional captioning and automatic speech



(a) A stenograph keyboard that shows its phonetic-based keys.

(b) A stenographer's typical Words Per Minute (WPM) limit and range.

Figure 1: Professional Real-Time Captioning using a stenograph

recognition (ASR). Both professional captioning and ASR provide a real-time word-for-word display of what is said in class, as well as options for saving the text after class for study. We discuss the readability of these approaches and a new approach, which utilizes crowd sourcing to generate real-time captions.

2.1 Professional Captioning

The most widely used approach, Communications Access Real Time (CART), is generated by professional captionists who use shorthand software to generate captions can keep up with natural speaking rates. Although popular, professional captioners undergo years of training, which results in professional captioning services being expensive. Furthermore, captionists usually have inadequate content knowledge and dictionaries to handle higher education lectures in specific fields. is the most reliable transcription service, but is also the most expensive one. Trained stenographers type in shorthand on a stenographic (short hand writing system) keyboard as shown in Figure 1. This keyboard maps multiple key presses to phonemes that are expanded to verbatim full text. Stenography requires 2-3 years of training to achieve at least 225 words per minute (WPM) and up to 300 WPM that is needed to consistently transcribe all real-time speech, which helps to explain the current cost of more than $100 an hour. CART stenographers need only to recognize and type in the phonemes to create the transcript, which enables them to type fast enough to keep up with the natural speaking rate. But the software translation of phonemes to words requires a dictionary that already contains the words used in the lecture; typing in new words into the dictionary slows down the transcription speed considerably. The stenographer can transcribe speech even if the words or phonemes do not make sense to them, e.g., if the speech words appear to violate rules of grammar, pronunciation, or logic. If the captioner cannot understand the phoneme or word at all, then they cannot transcribe it.

In response to the high costs of CART, computer-based macro expansion services like C-Print were developed and introduced. C-Print is a type of nearly-realtime transcription that was developed at the National Technical Institute for the Deaf. The captionist balances the tradeoff between typing speed and summarization, by including as much information as possible, generally providing a meaning-for-meaning but not verbatim translation of the spoken English content. This system enables operators who are trained in academic situations to consolidate and better organize the text with the goal of creating an end result more like class notes that may be more conducive to for learning. C-Print captionists need less training, and generally charge around $60 an hour. As the captionist normally cannot type as fast as the natural speaking rate, they are not able to produce a verbatim real-time transcript. Also, the captionist can only effectively convey classroom content if they understand that content themselves. The advantage is that the C-Print transcript accuracy and readability is high [21], but the disadvantage of this approach is that the transcript shows the summary that is based on the captionist's understanding of the material, which may be different from the speaker or reader's understanding of the material.

There are several captioning challenges in higher education. The first challenge is content knowledge - lecture information is dense and contains specialized vocabulary. This makes it hard to identify and schedule captionists who are both skilled in typing and have the appropriate content knowledge. Another captioning issue involves transcription delay, which occurs when captionists have to understand the phonemes or words and then type in what they have recognized. As a result, captionists tend to type the material to students with a delay of several seconds. This prevents students from effectively participating in an interactive classroom. Another challenge is speaker identification, in which captionist are unfamiliar with participants and are challenged to properly identify the current speaker. They can simplify this by recognizing the speaker by name, or asking the speaker to pause before beginning until the captionist has caught up and had an opportunity to identify the new speaker. In terms of availability, captionists typically are not available to transcribe live speech or dialogue for short periods or on-demand. Professional captionists usually need at least a few hours advance notice, and prefer to work in 1-hour increments so as to account for their commute times. As a result, students cannot easily decide at the last minute to attend a lecture or after class interactions with peers and teacher. Captionists used to need to be physically present at the event they were transcribing, but captioning services are increasingly being offered remotely [12, 1]. Captionists often are simply not available for many technical fields [21, 8]. Remote captioning offers the potential to recruit captionists familiar with a particular subject (e.g., organic chemistry) even if the captionist is located far away from an event. Selecting for expertise further reduces the pool of captionists. A final challenge is their cost - professional captionists are highly trained to keep up with speech with low errors rates, and so are highly paid. Experienced verbatim captionists' pay can exceed $200 an hour, and newly trained summarization captionists can go as low as $60 an hour [21].

2.2 Automatic Speech Recognition

ASR platforms typically use probabilistic approaches to translate speech to text. These platforms face challenges in accurately capturing modern classroom lectures that can have one or more of the following challenges: extensive technical vocabulary, poor acoustic quality, multiple information sources, speaker accents, or other problems. They also impose a processing delay of several seconds and the delay lengthens as the amount of data to be analyzed gets bigger. In other words, ASR works well under ideal situations, but degrades quickly in many real settings. Kheir et al. [12] found that untrained ASR software had 75% accuracy rate, but with training, could go to 90% under ideal single speaker, but this accuracy rate was still too low for use by deaf students. In the best possible case, in which the speaker has trained the ASR and wears a high-quality, noise-canceling microphone, the accuracy can be above 90%. When recording a speaker using a standard microphone on ASR not trained for the speaker, accuracy rates plummet to far below 50%. Additionally, the errors made by ASR often change the meaning of the text, whereas we have found non-expert captionists are much more likely to simply omit words or make spelling errors. In Figure 2 for instance, the ASR changes 'two fold axis' to 'twenty four lexus', whereas the c typists typically omit words they do not understand or make spelling errors. Current ASR is speaker-dependent, has difficulty recognizing domain-specific jargon, and adapts poorly to vocal changes, such as when the speaker is sick [6, 7]. ASR systems generally need substantial computing power and high-quality audio to work well, which means systems can be difficult to transport. They are also ill-equipped to recognize and convey tone, attitudes, interest and emphasis, and to refer to visual information such as slides or demonstrations. ASR services charge about $15-20 an hour. However, these systems are more easily integrated with other functions such as multimedia indexing.

2.3 Crowd Captions in the Classroom

Deaf and hard of hearing students have had a long history of enhancing their classroom accessibility by collaborating with classmates. For example, they often arrange to copy notes from a classmate and share it with their study group. Crowdsourcing has been applied to offline transcription with great success [2], but has just recently been used for real-time transcription [15]. Applying a collaborative captioning approach among classmates enables real-time transcription from multiple non-experts, and crowd agreement mechanisms can be utilized to vet transcript quality [14].

We imagine a deaf or hard of hearing person eventually being able to capture aural speech with her cellphone anywhere and have captions returned to her with a few seconds latency. She may use this to follow along in a lecture for which a professional captionist was not requested, to participate in informal conversation with peers after class, or enjoy a movie or other live event that lacks closed captioning. These use cases currently beyond the scope of ASR, and their serendipitous nature precludes pre-arranging a professional captionist. Lasecki et al. have demonstrated that a modest number of people can provide reasonably high coverage over the caption stream, and introduces an algorithm that uses overlapping portions of the sequences to align and merge them using the Legion:Scribe system [15]. Scribe is based on the Legion [13] framework, which uses crowds of

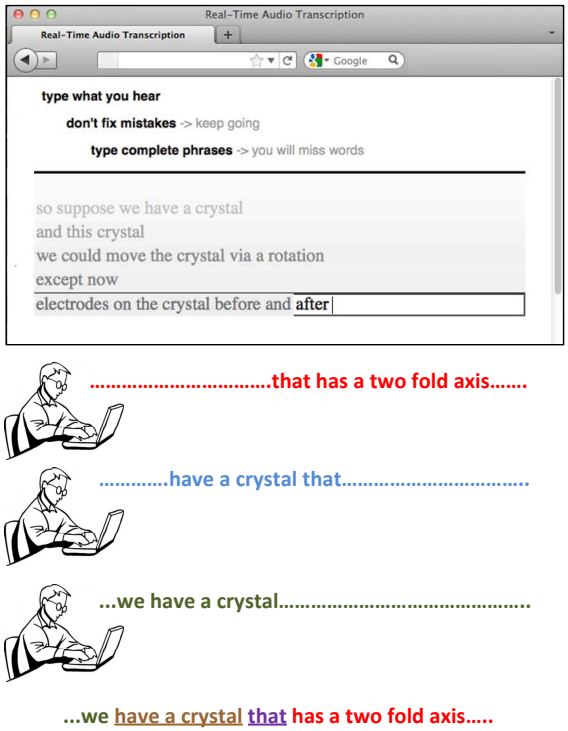


Figure 2: The crowd captioning interface. The interface provides a text input box at the bottom, and shifts text up as users type (either when the text hits the end of the box, or when the user presses the enter key). To encourage users to continue typing even when making mistakes, editing of text is disabled word by word. Partial captions are forwarded to the server in real-time, which uses overlapping segments and the order in segments are received to align and merge them.

workers to accomplish tasks in real-time. Unlike Legion, Scribe merges responses to create a single, better, response instead of selecting from inputs to select the best sequence. This merger is done using an online multiple sequence alignment algorithm that aligns worker input to both reconstruct the final stream and correct errors (such as spelling mistakes) made by individual workers.

Crowd captioning offers several potential benefits over existing approaches. First, it is potentially much cheaper than hiring a professional captionist because non-expert captionists do not need extensive training to acquire a specific skill set, and thus may be drawn from a variety of sources, e.g. classmates, audience members, microtask marketplaces, volunteers, or affordable and readily available employees. Our workforce can be very large because, for people who can hear, speech recognition is relatively easy and most people can type accurately. The problem is that individually they cannot type quickly enough to keep up with natural speaking rates, and crowd captioning nicely remedies this problem. Recent work has demonstrated that small crowds can be recruited quickly on-demand (in less than 2 seconds) from such sources[4, 3]. Scribe4Me enabled DHH users to

receive a transcript of a short sound sequence in a few minutes, but is not able to produce verbatim captions over long periods of time [17].

In previous work, we developed a crowd captioning system that accepts realtime transcription from multiple non-experts as shown in Figure 2. While non-experts cannot type as quickly as the natural speaking rate, we have found that they can provide accurate partial captions. Our system recruits fellow students with no training and compensates for slower typing speed and lower accuracy by combining the efforts of multiple captionists simultaneously and merges these partial captions in real-time. We have shown that groups of non-experts can achieve more timely captions than a professional captionist, that we can encourage them to focus on specific portions of the speech to improve global coverage, and that it is possible to recombine partial captions and effectively tradeoff coverage and precision [15].

2.4 Real-time text reading versus listening

Most people only see real-time text on TV at the bar or gym in the form of closed captions, which tend to have noticeable errors. However, those programs are captioned by live captionists or stenographers. To reduce errors, these real-time transcripts are often corrected and made into a permanent part of the video file by off-line captionists who prepare captions from pre-recorded videotapes and thoroughly review the work for errors before airing.

The translation of speech to text is not direct, but rather is interpreted and changed in the course of each utterance. Markers like accent, tone, and timbre are stripped out and represented by standardized written words and symbols. Then the reader interprets these words and flow to make meanings for themselves. Captionists tend not to include all spoken information so that readers can keep up with the transcript. Captionists are encouraged to alter the original transcription to provide time for the readers to completely read the caption and to synchronize with the audio. This is needed because, for a non-orthographic language like English, the length of a spoken utterance is not necessarily proportional to the length of a spelled word. In other words, reading speed is not the same as listening speed, especially for real-time scrolling text, as opposed to static pre-prepared text. For static text, reading speed has been measured at 291 wpm [19]. By contrast the average caption rate for TV programs is 141 wpm [11], while the most comfortable reading rate for hearing, hard-of-hearing, and deaf adults is around 145 wpm [10]. The reason is that the task of viewing real-time captions involved different processing demands in visual location and tracking of moving text on a dynamic background. English literacy rates among deaf and hard of hearing people who is low compared to hearing peers. Captioning research has shown that both rate and text reduction and viewer reading ability are important factors, and that captions need to be provided within 5 seconds so that the reader can participate [20].

The number of spoken words and their complexity can also influence the captioning decision on the amount of words to transcribe and degree of summarization to include so as to reduce the reader's total cognitive load. Jensema et al. [10] analyzed a large sample of captioned TV programs and found that the total set had around 800K words consisting of 16,000 unique words. Furthermore, over two-thirds of the transcript words consisted of 250 words. Higher education lecture transcripts have a very different profile. For comparison purposes, we selected a 50 minute long clip from the MIT Open CourseWare (OCW) website[1]. The audio sample was picked from a lecture segment in which the speech was relatively clear.We chose this lecture because it combined both technical and non-technical components. We found that the lecture had 9137 words, of which 1428 were unique, at 182.7 wpm. Furthermore, over two thirds of the transcript consisted of around 500 words, which is double the size of the captioned TV word set.

3. EVALUATION

To evaluate the efficacy of crowd-sourced real-time transcripts, we compared deaf and hearing user evaluations on their perceptions of the usability of crowd-sourced real-time transcripts against Computer Aided Real-Time transcripts (CART) and Automatic Speech Recognition transcripts (ASR).

3.1 Design Criteria

Based on prior work as well our own observations and experiences, we have developed the following design criteria for effective real-time transcript presentation for deaf and hard of hearing students:

1. The transcript must have enough information to be understood by the viewer.
2. The transcript must not be too fast or too slow so that it can be comfortably read.
3. Reading must not require substantial backtracking.

3.2 Transcript Generation

We obtained three transcriptions of an OCW lecture using crowd captioners, professional captioner and automatic speech recognition software and generated three transcripts of the lecture.

A professional real-time stenographer captionist who charged $200 an hour to create a professional real-time transcript of the lecture. The captioner listened to the audio and transcribed in real-time. The mean typing speed was about 180 wpm with a latency of 4.2 seconds. We calculated latency by averaging the latency of all matched words.

We recruited 20 undergraduate students to act as non-expert captionists for our crowd captioning system. These students had no special training or previous formal experience transcribing audio. Participants then provided partial captions for the lecture audio. The final transcript speed was about 130 WPM, with a latency of 3.87 seconds.

In addition to the these two transcripts, we generated a transcript using an automatic speech recognition *ASR* using Nuance Dragon Naturally Speaking 11 software. We used an untrained profile to simulate our target context of students transcribing speech from new or multiple speakers. To conduct this test, the audio files were played, and redirected to Dragon. We used a software loop to redirect the audio signal without resampling using SoundFlower[2], and a custom program to record the time when each word was generated by the ASR. The ASR transcript speed was 71.0 wpm (SD=23.7) with a latency of 7.9 seconds.

3.3 Transcript Evaluation

[1]http://ocw.mit.edu/

[2]http://code.google.com/p/soundflower/

Figure 3: The transcript viewing experience.

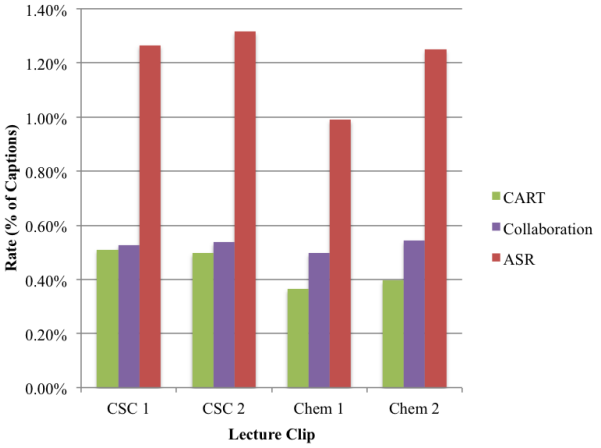


Figure 4: A comparison of the flow for each transcript. Both CART and crowd captions exhibit a relatively smooth real-time text flow. Students prefer this flow over the more choppy ASR transcript flow.

We recruited 48 students for the study over two weeks to participate in the study and evenly recruited both deaf and hearing students, male amd female. Twenty-one of the of them were deaf, four of them were hard of hearing and the remainder, twenty-four, were hearing. There were 21 females and 27 males, which reflects the gender balance on campus. Their ages ranged from 18 to 29 and all were students at RIT, ranging from first year undergraduates to graduate students. We recruited through flyers and word of mouth on the campus. We asked students to contact and schedule through email appointment. All students were reimbursed for their participation. All deaf participants were asked if they used visual accommodations for their classes, and all of them answered affirmatively.

Testing was conducted in a quiet room with a 22 inch flat-screen monitor as shown in Figure 3. Each person was directed to an online web page that explained the purpose of the study. Next, the students were asked to complete a short demographic questionnaire in order to determine eligibility for the test and asked for informed consent. Then they were asked to view a short 30 second introductory video to familiarize themselves with the process of viewing transcripts. Then the students were asked to watch a series of transcripts on the same lecture, each lasting two minutes. Each clip was labeled Transcript 1, 2 and 3, and were presented in a randomized order without any accompanying audio. The total time for the study was about 15 minutes.

After the participant completed watching all three video clips of the real-time transcripts, they were asked to complete a questionnaire. The questionnaire asked three questions. The first question asked "How easy was it to follow transcript 1?". In response to the question, the participants were presented with a a Likert scale that ranged from 1 through 5, with 1 being "Very hard" to 5 being "very easy". The second question asked "How easy was it to follow transcript 2?". In response to this question, participants were prompted to answer using a similar Likert scale response as in question 1. The third question was "How easy was it to follow transcript 3?". In response to this question, participants were promoted with a similar, corresponding Likert scale response to question 1 and 2. Then participants were asked to answer in their own words to three questions that asked participants for their thoughts about following the lecture through the transcripts; the first video transcript contained the captions created by the stenographer. The answers were open ended and many participants gave wonderful feedback. The second video transcript contained the captions created by the automatic speech recognition software, in this case, Dragon Naturally Speaking v. 11. The third and final video transcript contained the captions created by the crowd captioning process.

4. DISCUSSION

For the user preference questions, there was a significant difference between the Likert score distribution between Transcripts 1 and 2 or 2 and 3. In general, participants found it hard to follow Transcript 2 (automatic speech recognition); the median rating for it was a 1, i.e.,"Very hard". The qualitative comments indicated that many of them thought the transcript was too choppy and had too much latency. In contrast, participants found it easier to follow either Transcript 1 (professional captions) or 3 (crowd captions). Overall both deaf and hearing students had similar preference ratings for both crowd captions and professional captions (CART), in the absence of audio. While the overall responses for crowd captions was slightly higher at 3.15 (SD=1.06) than for professional captions (CART) at 3.08 (SD=1.24), the differences were not statistically significant ($\chi^2 = 32.52$, $p < 0.001$). There was a greater variation in preference ratings for professional captions than for crowd captions. When we divided the students into deaf and hearing subgroups and

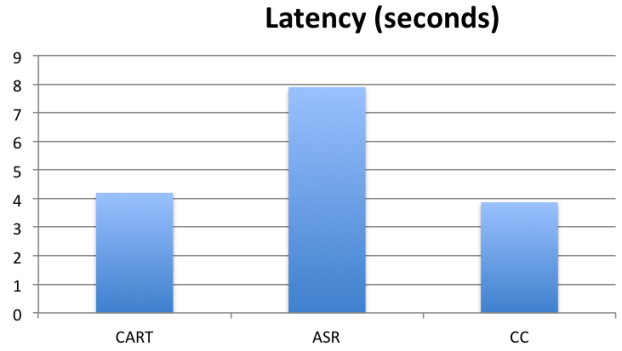


Figure 5: A graph of the latencies for each transcript (professional, automatic speech recognition and crowd). CART and Crowd Captions have reasonable latencies of less than 5 seconds, which allows students to keep up with class lectures, but not consistently participate in class questions and answers, or other interactive class discussion.

looked at their Likert preference ratings, there was no significant difference between crowd captions and professional captions for deaf students ($\chi^2 = 25.44$, $p < 0.001$). Hearing students as a whole showed significant difference between crowd captions and professional captions ($\chi^2 = 19.56$, $p = 0.07$).

The qualitative comments from hearing students revealed that transcript flow as shown in Figure 4, latency as shown in Figure 5 and speed were significant factors in their preference ratings. For example, one hearing student had the following comment for professional captioned real-time transcript: *"The words did not always seem to form coherent sentences and the topics seemed to change suddenly as if there was no transition from one topic to the next. This made it hard to understand so I had to try and reread it quickly"*. In contrast, for crowd captioning, the same student commented : *"I feel this was simpler to read mainly because the words even though some not spelled correctly or grammatically correct in English were fairly simple to follow. I was able to read the sentences about there being two sub-trees, the left and the right and that there are two halves of the algorithm attempted to be explained. The word order was more logical to me so I didn't need to try and reread it"*. On the other hand for the professional captions, a deaf student commented: *"It was typing slowly so I get distracted and I looked repeatedly from the beginning*"; and for crowd captions, the deaf student commented: *"It can be confusing so slow respsone on typing, so i get distracted on other paragraphs just to keep myself focused"*.

Overall, hearing participants appeared to like the slower and more smooth flowing crowd transcript rather than the faster and less smooth captions. Deaf participants appear to accept all transcripts. It may be that the deaf students are more used to bad and distorted inpurt and more easily skip or tolerate errors by picking out key words, but this or any other explanation requires further research. These considerations would seem to be particularly important in educational contexts where material may be captioned with the intention of making curriculum-based information available to learners.

A review of the literature on captioning comprehension and readability shows this result is consistent with findings from Burnham et al. [5], who found that there was no reduction in comprehension of text reduction for deaf adults, whether good or poor at reading. The same study also found that slower caption rates tended to assist comprehension of more proficient readers, but this was not the case for less proficient readers. This may explain why hearing students significantly preferred crowd captions over professional captions, whereas deaf students did not show any significant preference for crowd captions over professional captions. Since deaf students on average have a wider range of reading skills, it appears slower captions for the less proficient readers in this group does not help. Based on the qualitative comments, it appears that these students preferred to have a smoother word flow and to keep latency low rather than to slow down the real-time text. In fact, many of the less proficient readers commented that the captions were too slow. We hypothesize that these students, who tend to use interpreters rather than real-time captions, are focusing on key-words and ignore the rest of the text.

5. CONCLUSIONS

Likert ratings showed that hearing students rated crowd captions at or higher than professional captions, while deaf students rated both equally. A summary of qualitative comments on crowd captions suggests that these transcripts are presented at a readable pace, phrasing and vocabulary made more sense and that captioning flow is better than professional captioning or Automatic Speech Recognition.

We hypothesize that this finding is attributable to two factors. The first factor is that the speaking rate typically varies from 175-275 wpm [19], which is faster than the reading rate for captions of around 100-150 wpm, especially for dense lectures material. The second factor is that the timing for listening to spoken language is different from the timing for reading written text. Speakers often pause, change rhythm or repeat themselves. The end-result is that the captioning flow is as important as traditional captioning metrics such as coverage, accuracy and speed, if not more. The averaging of multiple caption streams into an aggregate stream appears to smooth the flow of text as perceived by the reader, as compared with the flow of text in professional captioning or ASR captions.

We think the crowd captionists are are typing the most important information to them, in other words, dropping the unimportant bits and this happens to better match the reading rate. As the captionists are working simultaneously, it can be regarded as a group vote for the most important information. A group of non-expert captionists appear to better able to collectively catch, understand and summarize as well as a single expert captioner. The constraint of the maximum average reading real-time transcript word flow reduces the need for making a trade off between coverage and speed; beyond a speed of about 140 words per minute [10], coverage and flow becomes more important. In other words, assuming a limiting reading rate (especially for dense lecture information), the comments show that students prefer to condensed material so that they can maintain reading speed/flow to keep up with the instructor.

One of the key advantages to using human captionists instead of ASR is the types of errors which are generated system when it fails to correctly identify a word. Instead of random text, humans are capable of inferring meaning, and selecting from possible words which make sense in the

context of the speech. We anticipate this will make quickCaption more usable than automated systems even in cases where there may be minimal difference in measures such as accuracy and coverage.

We propose a new crowd captioning approach that recruits classmates and others to transcribe and share classroom lectures. Classmates are likely to be more familiar with the topic being discussed, and to be used to the speaker's style. We show that readers prefer this approach. This approach is less expensive and is more inclusive, scalable, flexible and easier to deploy than traditional captioning, especially when used with mobile devices. This approach can scale in terms of classmates and vocabulary, and can enable efficient retrieval and viewing on a wide range of devices. The crowd captioning transcript, as an average of multiple streams from all captionists, is likely to be more consistent and have less surprise than any single captionist, and have less delay, all of which reduce the likelihood of information loss by the reader. This approach can be viewed as a parallel note-taking that benefits all students who get an high coverage, high quality reviewable transcript that none of them could normally type on their own.

We have introduced the idea of real-time non-expert captioning, and demonstrated through coverage experiments that this is a promising direction for future research. We show that deaf and hearing students alike prefer crowd captions over ASR because the students find the errors easier to backtrack on and correct in real-time. Most people cannot tolerate an error rate of 10% or more as errors can completely change the meaning of the text. Human operators who correct the errors on-the-fly make these systems more viable, opening the field to operators with far less expertise and the ability to format, add punctuation, and indicate speaker changes. Until the time ASR becomes a mature technology that can handle all kinds of speech and environments, human assistance in captioning will continue to be an essential ingredient in speech transcription.

We also notice that crowd captions appear to have more accurate technical vocabulary than either ASR or professional captions. Crowd captioning outperforms ASR in many real settings. Non-expert real-time captioning has not yet, and might not ever, replace professional captionists or ASR, but it shows lot of promise. The reason is that a single captioner cannot optimize their dictionary fully, as they have to to adapt to various teachers, lecture content and their context. Classmates are much better positioned to adapt to all of these, and fully optimize their typing, spelling, and flow. Crowd captioning enables the software and users to effectively adapt to a variety of environments that a single captionist and dictionary cannot handle.

One common thread among the feedback comments revealed that deaf participants are not homogenous, and there there is no neat unifying learning style abstraction. Lesson complexity, learning curves, expectations, anxiety, trust and suspicions can all can affect learning experiences and indirectly the satisfaction and rating of transcripts.

6. FUTURE WORK

From the perspective of a reader viewing a real-time transcript, not all errors are equally important, and human perceptual errors of the dialog is much easier for users to understand and adapt to than ASR errors. Also unlike ASR, crowd captioning can handle poor dialog audio or untrained speech, e.g. multiple speakers, meetings, panels, audience questions. Using this knowledge, we hope to be able to encourage crowd captioning workers to leverage their understanding of the context that content is spoken in to capture the segments with the highest information content.

Non-expert captionists and ASR make different types of errors. Specifically, humans generally type words that actually appear in the audio, but miss many words. Automatic speech recognition often misunderstands which word was spoken, but generally gets then number of words spoken nearly correct. One approach may be to use ASR as a stable underlying signal for real-time transcription, and use non-expert transcription to replace incorrect words. This may be particularly useful when transcribing speech that contains jargon terms. A non-expert captionist could type as many of these terms as possible, and could fit them into the transcription provided by ASR where appropriate.

ASR usually cannot provide a reliable confidence level of their own accuracy. On the other hand, the crowd usually has a better sense of their own accuracy. One approach to leverage this would be to provide an indication of the confidence the system has in recognition accuracy. This could be done in many ways, for example through colors. This would enable the users to pick their own confidence threshold.

It would be useful to add automatic speech recognition as a complementary source of captions because its errors are generally independent of non-expert captionists. This difference means that matching captions input by captionists and ASR can likely be used with high confidence, even in the absence of many layers of redundant captionists or ASR systems. Future work also seeks to integrate multiple sources of evidence, such as N-gram frequency data, into a probabilistic framework for transcription and ordering. Estimates of worker latency or quality can also be used to weight the input of multiple contributors in order to reduce the amount of erroneous input from lazy or malicious contributors, while not penalizing good ones. This is especially important if crowd services such as Amazon's Mechanical Turk are to be used to support these systems in the future. The models currently used to align and merge sets of partial captions from contributors are in their infancy, and will improve as more work is done in this area. As crowd captioning improves, students can begin to rely more on readable captions being made available at any time for any speaker.

The benefits of captioning by local or remote workers presented in this paper aims to further motivate the use of crowd captioning. We imagine a deaf or hard of hearing person eventually being able to capture speech with her cellphone anywhere and have captions returned to her within a few seconds latency. She may use this to follow along in a lecture for which a professional captionist was not requested, to participate in informal conversation with peers after class, or enjoy a movie or other live event that lacks closed captioning. These use cases currently beyond the scope of ASR, and their serendipitous nature precludes prearranging a professional captionist. Moreover, ASR and professional captioning systems do not have a consistent way of adding appropriate punctuation from lecture speech in real-time, resulting in captions that are very difficult to read and understand [9, 16].

A challenge in developing new methods for real-time captioning is that it can be difficult to quantify whether the captions have been successful. As demonstrated here, us-

ability and readability of real-time captioning is dependent on much more than just Word Error Rate, involving at a minimum naturalness of errors, regularity, latency and flow. These concepts are difficult to capture automatically, which makes it difficult to make reliable comparisons across different approaches. Designing metrics that can be universally applied will improve our ability to make progress in systems for real-time captioning.

7. ACKNOWLEDGMENTS

We thank our participants for their time and feedback in evaluating the captions, and the real-time captionists for their work in making the lecture accessible to deaf and hard of hearing students.

8. REFERENCES

[1] Faq about cart (real-time captioning), 2011. http://www.ccacaptioning.org/articles-resources/faq.

[2] 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.

[3] M. S. Bernstein, J. R. Brandt, R. C. Miller, and D. R. Karger. 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, page to appear, New York, NY, USA, 2011. ACM.

[4] 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.

[5] D. Burnham, G. Leigh, W. Noble, C. Jones, M. Tyler, L. Grebennikov, and A. Varley. Parameters in television captioning for deaf and hard-of-hearing adults: Effects of caption rate versus text reduction on comprehension. *Journal of Deaf Studies and Deaf Education*, 13(3):391–404, 2008.

[6] X. Cui, L. Gu, B. Xiang, W. Zhang, and Y. Gao. Developing high performance asr in the ibm multilingual speech-to-speech translation system. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 5121 –5124, 31 2008-april 4 2008.

[7] L. B. Elliot, M. S. Stinson, D. Easton, and J. Bourgeois. College Students Learning With C-Print's Education Software and Automatic Speech Recognition. In *American Educational Research Association Annual Meeting*, New York, NY, 2008.

[8] M. B. Fifield. Realtime remote online captioning: An effective accommodation for rural schools and colleges. In *Instructional Technology And Education of the Deaf Symposium*, 2001.

[9] A. Gravano, M. Jansche, and M. Bacchiani. Restoring punctuation and capitalization in transcribed speech. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4741 –4744, april 2009.

[10] C. Jensema. Closed-captioned television presentation speed and vocabulary. *American Annals of the Deaf*, 141(4):284 – 292, 1996.

[11] C. J. Jensema, R. Danturthi, and R. Burch. Time spent viewing captions on television programs. *American Annals of the Deaf*, 145(5):464–468, 2000.

[12] R. Kheir and T. Way. Inclusion of deaf students in computer science classes using real-time speech transcription. In *Proceedings of the 12th annual SIGCSE conference on Innovation and technology in computer science education*, ITiCSE '07, pages 261–265, New York, NY, USA, 2007. ACM.

[13] W. Lasecki, K. Murray, S. White, R. C. Miller, and J. P. Bigham. Real-time crowd control of existing interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11, page To Appear, New York, NY, USA, 2011. ACM.

[14] W. S. Lasecki and J. P. Bigham. Online quality control for real-time captioning. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '12, 2012.

[15] W. S. Lasecki, C. Miller, A. Sadilek, A. Abumoussa, D. Borrello, R. Kushalnagar, and J. P. Bigham. Realtime captioning by groups of non experts. In *Proceedings of the 25th ACM UIST Symposium*, UIST '12, 2012.

[16] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1526 –1540, sept. 2006.

[17] T. Matthews, S. Carter, C. Pai, J. Fong, and J. Mankoff. In *Proceeding of the 8th International Conference on Ubiquitous Computing*, pages 159–176, Berlin, 2006. Springer-Verlag.

[18] R. E. Mitchell. How many deaf people are there in the United States? Estimates from the Survey of Income and Program Participation. *Journal of deaf studies and deaf education*, 11(1):112–9, Jan. 2006.

[19] S. J. Samuels and P. R. Dahl. Establishing appropriate purpose for reading and its effect on flexibility of reading rate. *Journal of Educational Psychology*, 67(1):38 – 43, 1975.

[20] M. Wald. Using automatic speech recognition to enhance education for all students: Turning a vision into reality. In *Frontiers in Education, 2005. FIE '05. Proceedings 35th Annual Conference*, page S3G, oct. 2005.

[21] M. Wald. Creating accessible educational multimedia through editing automatic speech recognition captioning in real time. *Interactive Technology and Smart Education*, 3(2):131–141, 2006.