Open access • Journal Article • DOI:10.1145/2543578

Accessibility Evaluation of Classroom Captions — Source link

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

Institutions: Rochester Institute of Technology, University of Rochester, Carnegie Mellon University

Published on: 01 Jan 2014 - ACM Transactions on Accessible Computing (ACM)

Topics: Closed captioning

Accessibility Evaluation of Classroom Captions

RAJA S. KUSHALNAGAR, Rochester Institute of Technology
WALTER S. LASECKI, University of Rochester
JEFFREY P. BIGHAM, Carnegie Mellon University

Real-time captioning enables deaf and hard of hearing (DHH) people to follow classroom lectures and other aural speech by converting it into visual text with less than a five second delay. Keeping the delay short allows end-users to follow and participate in conversations. This paper focuses on the fundamental problem that makes real-time captioning difficult: sequential keyboard typing is much slower than speaking. We first surveyed the audio characteristics of 240 one-hour long captioned lectures on YouTube, such as speed and duration of speaking bursts. We then analyzed how these characteristics impact caption generation and readability, considering specifically our human-powered *collaborative captioning* approach. We note that most of these characteristics are also present in more general domains. For our caption comparison evaluation, we transcribed a classroom lecture in real-time using all three captioning approaches. We recruited 48 participants (24 DHH) to watch these classroom transcripts in an eye-tracking laboratory. We presented these captions in a randomized, balanced order. We show that both hearing and DHH participants preferred and followed collaborative captions better than those generated by automatic speech recognition (ASR) or professionals due to the more consistent *flow* of the resulting captions. These results show the potential to reliably capture speech even during sudden bursts of speed, as well as for generating "enhanced" captions, unlike other human powered captioning approaches.

Categories and Subject Descriptors: H.5.2 [**Information interfaces and presentation**]: User Interfaces

General Terms: Design, Human Factors, Experimentation

Additional Key Words and Phrases: Real-time captioning, deaf, hard of hearing, crowdsourcing

ACM Reference Format:
Raja S. Kushalnagar, Walter S. Lasecki and Jeffrey P. Bigham, 2013. Accessibility Evaluation of Classroom Captions. *ACM Trans. Access. Comput.* 1, 1, Article 1 (March 2013), 24 pages.
DOI:http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Real-time captions transform aural speech into visual text in nearly real-time (less than 5-second delay) for deaf and hard of hearing (DHH) people who cannot hear or understand aural speech. Real-time captioning is usually preferred by people who have lost their hearing as adults, or DHH individuals who do not know sign language. Current human-powered approaches to generating real-time captions in the classroom are

This work is supported by grants from the National Science Foundation ("Real-Time Captioning by Groups of Non-Experts for Deaf and Hard of Hearing Students," #12070356 and #1218209, 2012), and from the National Technical Institute for the Deaf ("Multiple Viewing Perspectives," 2011) at Rochester Institute of Technology ("Accessible Viewing Device: Enhancing Classroom Access," 2012).
Author's addresses: R. S. Kushalnagar, Information and Computing Studies Department, Rochester Institute of Technology; email: rskics@rit.edu; W. S. Lasecki, Computer Science Department, University of Rochester; email: wlasecki@cs.rochester.edu; J. P. Bigham, Human-Computer Interaction Institute, Carnegie Mellon University; email: jbigham@cmu.edu.

DOI:http://dx.doi.org/10.1145/0000000.0000000

readable, but tend to be expensive and hard to obtain geographically or temporally, while current ASR is hard to read. Current real-time caption services are either expensive and hard to obtain in the case of human powered captioning, or hard to read in the case of automatic speech recognition. In this paper, we explore characteristics of classroom lectures that makes situations difficult to caption, and compare different methods in terms of readability.

1.1. Impact of Caption Generation: Cost, Quality and Availability

In terms of real-time captioning quality and availability, though the Americans with Disabilities Act was passed a generation ago in 1991, the quality of visual accommodations does not meet the needs of DHH people. One indicator of this unmet need is the fact that the graduation rate of DHH students remains abysmal at 25% [Lang 2002], in contrast to the nationwide graduation rate of 56% for all students [Aud et al. 2011]. This disparity indicates that deaf students still do not have full equal access in the classroom, even with visual translation services [Marschark et al. 2008]. These findings highlight a need to maximize the academic potential of deaf students by focusing on caption cost, accuracy, availability, and readability.

Captioning options for DHH students are limited due to cost, availability, and quality concerns. Human powered caption services are expensive, not available on demand, and their captions may not match their consumers' reading speed and abilities. While ASR is cheaper and potentially always available, it generally produces unacceptable error rates, not only in most lecture environments, but also in many unconstrained real-world environments [Kheir and Way 2007]. We discuss how collaborative captioning can help address the issues of cost, quality, and availability.

1.2. Impact of Caption Readability: Speed, Vocabulary and Flow

There is substantial literature about the readability of summarized captions that are superimposed on TV screens, but far less literature on the readability of classroom verbatim captions that are displayed on separate computer screens, which are common in higher education captioning environments. TV captions show 1-2 lines for 2 to 4 seconds, while classroom verbatim captions show 15-20 lines for up to one minute. In terms of caption reading speed and vocabulary, we find that verbatim higher education captions have substantially different characteristics than condensed TV captions. Lecture audio tends to have a faster speaking rate and present a larger set of longer and less frequently encountered words than most TV shows.

In terms of the consistency of caption rates, research on TV caption and higher education caption reading process has shown that the cognitive process of reading a real-time text representation of speech that constantly changes is different from the cognitive process of reading print, which is organized spatially and not temporally. In terms of pace, print is presented all at once, and the act of reading is controlled by the reader throughout the course of reading [Thorn and Thorn 1996; Kushalnagar et al. 2013]. Unlike print, captions force readers to read the text at the speaker's variable rhythm and pace, and the readers cannot control or predict that rhythm or pace and are bothered by excessive speed [Jensema 1998]. On the other hand, the speaking rhythm and pace is a natural and essential component [Klatt 1976] for listeners. In terms of caption accuracy, TV caption readers are most bothered by the omission of key words [Jordan et al. 2003]. Combined, these factors can aggravate the mismatch between the natural speaking and natural caption reading rates. We show that a new collaborative captioning approach used by Legion:Scribe [Lasecki et al. 2012] makes steps towards addressing each of these issues.

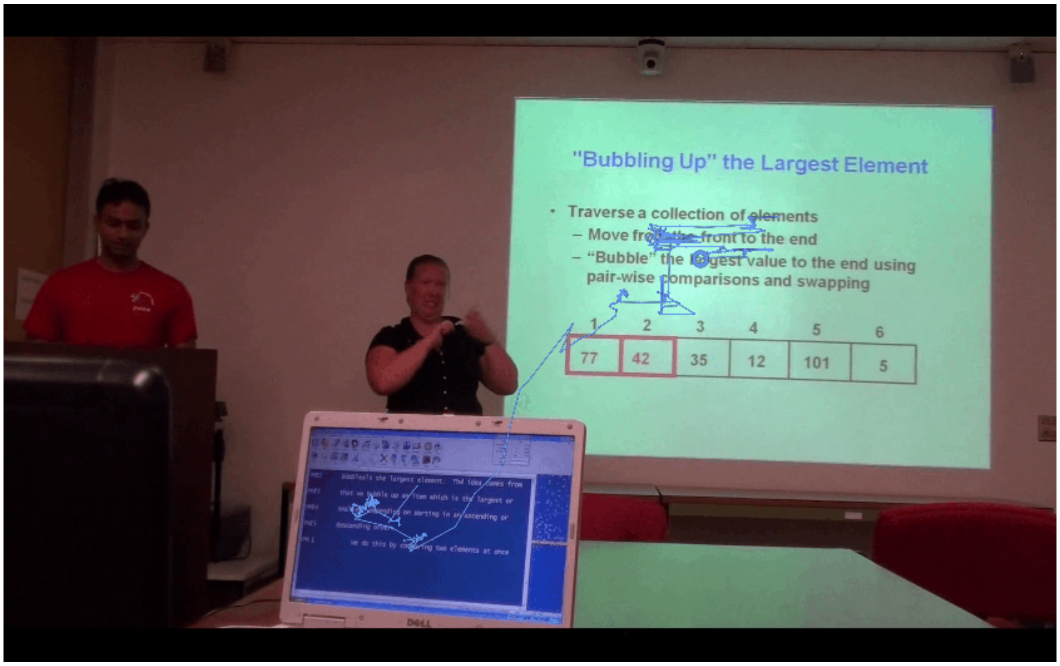

"Bubbling Up" the Largest Element
• Traverse a collection of elements
– Move from the front to the end
– "Bubble" the largest value to the end using pair-wise comparisons and swapping
1 2 3 4 5 6
77 42 35 12 101 5

Fig. 1. Eye-gaze scan path of a deaf student's reading process in a visually accessible classroom lecture that has both captions and a sign language interpreter

Figure 1 illustrates the difference in reading captions versus reading print. It shows a deaf student's reading gaze in a lecture with two visual translations of the audio. The first translation is the real-time transcription of speech onto a computer screen that is being typed by a stenographer next to the student. The second translation, which the deaf student is not watching, is a sign language interpreter who is translating the speech to signs. The eye-gaze scan path shows the difference in the student's reading gaze while looking at static text on the slide versus dynamic text on the caption screen.

1.3. Contributions

This article is an expanded version of a conference paper [Kushalnagar et al. 2012], which compared the most common real-time caption generation methods that higher education lectures use today: stenography, Automatic Speech Recognition system (ASR) and collaborative captioning that uses Legion:Scribe. We extend this comparison in three ways. First, we analyze the properties of higher education lectures and show how these impact caption generation and readability. Second, we perform a quantitative analysis of the speaking rate in a set of 240 captioned lectures on YouTube, from which we identify properties that are challenging for real-time transcription, such as complexity, average word length, and frequency. Third, we analyze how stenography currently extends the limits of human-powered captioning through parallelization of keystrokes or abbreviation of words, and show how our approach of parallelizing human typing is more scalable than either of the other two approaches. We then explore caption readability for higher education lectures through evaluations from 48 students who watched three kinds of captions generated from a stenographer, from an ASR, and from Scribe typists. We extend the findings through qualitative analysis of their eye-tracking data.

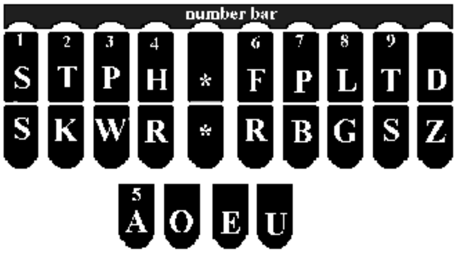


Fig. 2. Phonetic based Stenography Keyboard

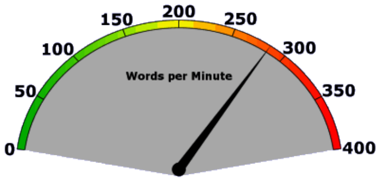


Fig. 3. Stenography Captioning speed

2. RELATED WORK

Equal access to communication is fundamental to the academic success of students, but is often taken for granted. People tend to incorrectly assume that the captioner can capture the speech accurately. They also assume that deaf students can read the verbatim text and understand the full richness of the speech. This assumption of full access is detrimental as it distracts people from addressing issues related to caption generation and readability.

2.1. Caption Generation

Our survey of 240 captioned YouTube lectures showed that the average speaking rate was around 170 words per minute (WPM). In order to transcribe an average lecture, a transcriber has to maintain a rate of at least 170 WPM. If we assume that each word has 5 letters for the purpose of transcription [Yamada 1980], the transcriber has to enter over 850 characters per minute (CPM), which is physically impossible for most individuals on a typical keyboard. Even proficient touch typists are only able to enter around 75 WPM on QWERTY keyboards [Arif and Stuerzlinger 2009], which translates into a sustainable rate of around 375 key presses per minute. Most real-time entry approaches adopt alternate data-entry methods that either parallelize key presses, or use word expansion, which enables typists to produce transcripts at a rate fast enough to keep up with natural speaking rates. We review these approaches and discuss our new approach in which we parallelize the number of typists, which is potentially much more effective than either of the two approaches mentioned above.

2.1.1. Computer-Aided Access Real Time (CART). CART is currently the most reliable transcription service, but is also the most expensive option due to the difficulty and length of the required training to simultaneously listen to speech and type in shorthand on a specialized keyboard. They type on a stenograph (short-hand typing system) keyboard, as shown in Figure 2. Certified court stenographers are required to type up to 225 WPM and faster in bursts in order to consistently transcribe all real-time speech as shown in Figure 3. This keyboard maps multiple key presses (chords), to phonemes that are expanded to verbatim full text. The phonemes of a word are combined into a chord on a steno keyboard which allows the coding of multiple phonemes at a time.

Stenographers are able to code up to one word by a simultaneous key press with up to all 10 fingers: The left keys on the keyboard are used to code the initial part of the word, the down keys code the middle part of the word, and the right keys of the keyboard code the final part of the word. For high frequency words or phrases, prefixes and suffixes, and abbreviations are used. The phonetic code of the words or the respective abbreviation is immediately translated into the corresponding long form using a dictionary.

The following example illustrates the relative efficiency of sequential presses in stenography over typing:

- A transcription of a spoken sentence on a qwerty keyboard requires 63 key presses: *Anything you can do with recursion, you can also do with a loop*
- The same transcription on a stenotype keyboard requires 12 chord presses: *TPHEUG UBG TKAO W RURGS KWBG UBG TKAO LS W A HRAOP*

The stenographer is able to keep up with natural speaking rates by parallelizing key presses. In the above example, the number of sequential presses was reduced from 63 to 12, a reduction of more than 80%. The stenographer's speed is limited mainly by his or her ability to recognize the spoken words and then to type in the corresponding chords. A stenographer averages about 5 letters per chord, based on an analysis of 30 CART-generated transcripts for an introductory programming course. In order to enter around 850 letters per minute, the stenographer only needs to type around 170 chords per minute, which is a comfortable pace for most stenographers, far below the sustained limit of 375 key presses per minute.

The main constraints for CART availability is their relative scarcity and high costs. Stenographers are relatively scarce, as few people are willing to undergo 2 to 3 years of intensive training. According to the National Court Reporters Association (NCRA) a certified stenographer must be able to achieve at least 225 words per minute (WPM) at 98% accuracy. Due to their lack of availability, it is hard for DHH students to book them especially at the last minute. For example, if an employee has to attend a meeting that was scheduled at the last minute, it would be very difficult to book a stenographer for that meeting. CART stenographers cost around $100 per hour.

In addition to these general constraints, stenographers rarely have adequate content knowledge to handle higher education lectures in specific fields, and often lack the appropriate dictionaries since they need to build up their personalized dictionary with the words used in the lecture. The dictionary is needed to translate the phonemes into words; if the captioner has to type in new words into the captions and dictionary, their transcription speed slows down considerably.

In response to the high cost of CART, non-verbatim note-taking methodologies offer benefits at a lower cost. For example, computer-based macro expansion services like C-Print or Typewell balance the trade-off between typing speed and the level of summarization by including as much information as possible. CPrint and Typewell use software to expand abbreviations to words. For example, when a CPrint typist hears 'coffee', they will type 'kfe', which is automatically expanded to 'coffee'. Since typists cannot keep up with speaking bursts, they usually summarize each sentence, which is called a meaning-for-meaning translation, and is not verbatim translation of the spoken English content. The advantage is that C-Print typists need less training, and generally charge about half the rate of CART [Elliot et al. 2001]. The disadvantage is that the CPrint typist cannot type as fast as the natural speaking rate, especially during speaking bursts. A CPrint typist has an average expansion ratio of around 2, based on a survey of 30 CPrint transcripts for an introductory programming course. In addition, CPrint typists also have to listen, understand, summarize, and type in, which is slower than typing a presented sentence [Wobbrock and Myers 2006]. In order to enter around 850 CPM, the typist would have to key in, before expansion, at a rate of 425 CPM. Even without considering the mental workload of doing meaning-for-meaning translation, this typing rate is over the upper limit of the average typing speed of 375 CPM for most proficient typists. CPrint typists charge around $50 an hour.

In either case, the captioner can only effectively convey classroom content if they understand it, which can be worse if the transcript is meaning-for-meaning translation rather than verbatim.

2.1.2. Automatic Speech Recognition (ASR). ASR systems typically use probabilistic approaches to translate speech into text. Researchers since the late 1990s, including the Liberated Learning Project, have investigated whether ASR could successfully transcribe lectures [Bain et al. 2002]. They have explored the challenges in accurately capturing modern classroom lectures that have extensive technical vocabulary, poor acoustic quality, multiple information sources, or speaker accents. Other problems with ASR include processing delays of several seconds, which lengthens as the amount of data to be analyzed gets bigger. ASR works well under ideal situations, but degrades quickly in most higher education settings. A study on untrained ASR software in lectures found that they had a 75% accuracy rate but, with training, could reach 90% under ideal single speaker conditions. However, this accuracy rate is still too low for use by deaf students [Kheir and Way 2007]. The study indicated that when the speaker has trained the ASR and wears a high-quality, noise-canceling microphone, the accuracy can be above 90%, but otherwise had far lower accuracy rates.

Additionally, the errors ASR makes can often change the meaning of the text. For instance, in Figure 7, the ASR changes ‘two fold axis’ to ‘twenty four lexus’. Human captioners on the other hand, will typically only omit words they do not understand, or make accidental spelling errors which are easier for readers to interpret. Current ASR is speaker-dependent and has difficulty recognizing domain-specific jargon [Cui et al. 2008]. This problem will also arise when the speaker’s speech is altered, for example when the speaker has a cold. ASR systems also 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. Captioners and Scribe human agents can understand most of these auditory cues and partially communicate this information by typing in conventional emoticons (such as smileys), and where to locate the simultaneous visual information, e.g., on the overhead or white board. An advantage is that these systems can be integrated with other online functionality, such as multimedia indexing. ASR services typically charge about $15-20 per hour.

2.1.3. Collaborative Captioning (Legion:Scribe). We have recently proposed a new approach [Lasecki et al. 2012]: instead of trying to reduce the number of key presses needed to type words, we propose an approach in which multiple typists collaboratively work together to achieve a higher typing rate. This approach has the advantage of being much more scalable and has potentially no upper limit. In order to keep up with the average lecture speaking rate of 170 WPM (850 CPM), we only need three people typing together, assuming that each person can type about 375 CPM.

Scribe is a crowd captioning system that combines simultaneous partial captions from multiple non-experts in real-time, as shown in Figure 4. Scribe is based on the real-time crowdsourcing approach used in Legion [Lasecki et al. 2011] to allow crowds of workers to control existing interfaces. Unlike Legion, Scribe merges responses to create a single, high-quality response instead of choosing from individual 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.

Captioners in Scribe are presented with a text input interface designed to encourage real-time answers and increase global coverage (Figure 7). A display shows workers their rewards for contributing in the form of both money and points. As workers

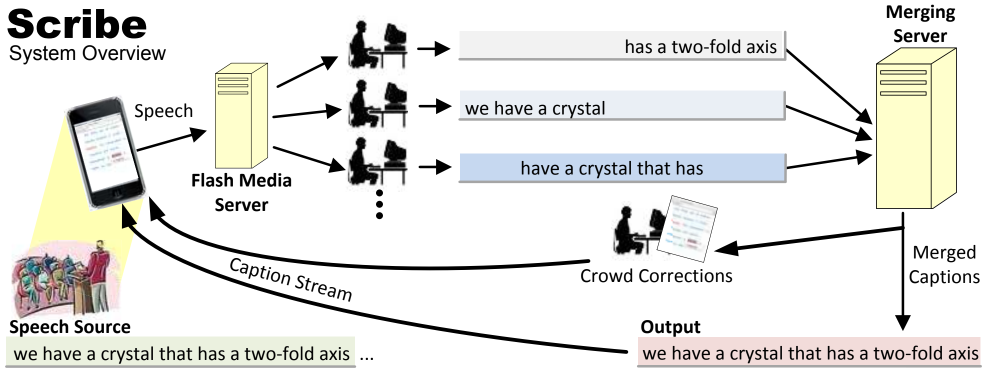


Fig. 4. Legion:Scribe allows users to caption audio on their mobile device. The audio is sent to multiple non-expert captioners in real-time who use our web-based interface to caption as much of the audio as they can. These partial captions are sent to the merging server to be merged into a final output caption stream, which is then forwarded back to the user's mobile device.

type, their input is forwarded to the online sequence alignment algorithm running on the Scribe server. In prior experiments, workers were paid $0.005 for every word the system thought was correct – resulting in a total cost of $45/hr.

By applying Scribe, it is even possible to use classmates with no formal captioning experience to generate captions in real-time. Crowd agreement mechanisms can then be utilized to vet transcript quality on-the-fly [Lasecki and Bigham 2012]. Scribe has demonstrated the ability to provide high coverage over the caption stream even with modest numbers of people [Lasecki et al. 2012].

Scribe offers several potential benefits over existing approaches:

First. Scribe is cheaper than hiring a professional captioner as non-expert captioners 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 individual typists cannot type quickly enough to keep up with natural speaking rates. However, if they are asked to type parts of the audio, they are able to keep up with these individual parts and generate partial and accurate captions. These partial captions can be combined in real-time to nicely remedy this problem. Recent work has demonstrated that small crowds can be recruited quickly on-demand (in less than 2 seconds) from such sources [Bigham et al. 2010; Bernstein et al. 2011].

Second. Scribe uses multiple audio inputs, which offers robustness in caption coverage over systems that use a single audio input, such as professional captioners or automatic speech recognition systems. Single audio input systems can only type what they hear. If they are too far away, distracted or behind, then they will not get everything and their transcript will have gaps. Recently in the US Supreme Court, the official court stenographer could not hear a remark by Justice Thomas due to laughter [USSC 2012]. The official transcript recorded four words by Thomas at this point: "Well, he did not." The next day, the captioner revised the transcript after asking other officials for clarification: "Well, there - see he did not provide good counsel." If there had been multiple typists, then at least one of them would have likely heard Justice Thomas' comment correctly and transcribed it in real-time.

2.2. Caption Presentation

The environment in which captions are generated and viewed makes the result very different from conventional print (books or online browsing).

2.2.1. Captions versus Print. In print, words are presented all together on a page or on screen. The static print enables readers to predict when and where to move the gaze according to their reading speed. By contrast, in real-time captioning, the prediction strategy is harder as there are no words to scan ahead. Instead, the text is written and read almost simultaneously, and the control of the reading speed shifts at least partly over to the speaker and the captioner. The text is not fixed in advance, but instead is being produced continuously, requiring readers to follow this word production process very closely if they want to use the real-time captions. In addition, captions roll up (or down) each time a new line of text is entered, which may result in reading disorientation, as the readers lose track of where they were looking. Because of this interaction of generation and reading in real-time, the generated text presentation must be adapted to the reader's needs. This issue will be discussed in the eye-tracking section.

2.2.2. Visual Representation of Speech. Spoken and written forms of language are transmitted and perceived through different modalities. This affects how the presented information is processed. Speech perception depends heavily on temporal markers such as rhythm and pace, while text perception depends more on spatial markers such as spacing and punctuation. That is, speech is characterized by rapidly changing patterns that do not always map well to text designed to be represented statically. In other words, reading speed is not the same as listening speed, especially for real-time moving text, as opposed to static text.

Some temporal markers like punctuation or paragraphs can be shown in text via punctuation, but many temporal markers like punch lines for jokes are not easily represented through text. As a result, reading captions can rarely be done in perfect synchrony with the audio, as the pace of speech varies according to the implied stress or emotion.

For static text, the average college student reads at around 280 WPM [Hasbrouck and Tindal 2006]. By contrast, the average caption rate for TV programs is 141 WPM [Jensema et al. 2000], while the most comfortable reading rate for hearing, hard-of-hearing, and deaf adults is around 145 WPM [Jensema 1996]. Since the speech rate is faster than the average caption reading rate, captioners tend not to include all spoken information so that readers can keep up with the transcript. Instead, they abbreviate sentences or drop words. English literacy rates among deaf and hard of hearing people is low compared to hearing peers. Captioning research has shown that both rate of text production and viewer reading ability are important [Jensema et al. 2000].

3. CAPTION GENERATION IN THE CLASSROOM

Though it may seem obvious, deaf students demonstrate a significant improvement when captions are included [Boyd and Vader 1972]. However, the technology was far too expensive for use by individual consumers in the courtroom or classroom until the 1980s [Downey 2006]. Though real-time captions continue to be expensive and not easily available, the technology has improved in terms of all of the captioning issues outlined below. If the cost comes down and captioning availability improves, then the universe of events that can be transcribed in real-time will grow exponentially and will benefit all consumers. We discuss general issues related to real-time captioning regardless of the underlying technology. In order for captions to be accessible to readers, they have to be shown at a speed that matches the reader's processing ability, and to have a low error rate so that the reader is not derailed during the reading process.

3.1. Caption Issues

There are several captioning software issues in higher education that impact their usability by both captioners and DHH students. Although most captioners are able to type with sufficient accuracy and speed in legal environments, many captioners find it challenging to transcribe in higher education environments. These caption systems issues in higher education environments include:

(1) *Vocabulary Size*
The total number of spoken words in a speech influences the captioner's ability to distinguish the words in their dictionary. It also influences the reader's ability to read and process the words quickly. Jensema et al. [Jensema 1996] analyzed a large set of captioned TV programs, which totaled around 800,000 words. There were 16,000 unique words, and over two-thirds of the transcript words consisted of 250 words. Higher education lecture transcripts have a very different profile. Our YouTube survey (discussed in section 4) has found that lectures had a much bigger vocabulary, and captioners had to enter these words into their dictionary in order to be accurate and keep up with the lecture audio.

(2) *Vocabulary Complexity*
The complexity of the words in terms of length and frequency can also influence the captioner's ability to type the right word quickly and for the reader to be able to read and understand content quickly.

(3) *Accuracy*
Research into the readability of speech recognition transcription in lectures has determined that an accuracy of at least 90% is required to make a transcript usable in the classroom [Kheir and Way 2007; Pan et al. 2010] and 98% or higher may be needed to maintain the meaning and intent of the content [Stuckless 1999].

(4) *Latency*
Transcription delay or latency occurs when captioners have to understand the speech and then type in what they have recognized. As a result, captioners tend to type the material to students with a delay of several seconds. This prevents students from effectively participating in an interactive classroom. Research findings suggest that captions need to be provided within 5 seconds so that the student can participate [Wald 2005].

3.2. Captioner Issues

Captioners face unique challenges in doing real-time transcription in higher education environments in contrast with legal environments. These challenges are:

(1) *Ease of Access*
Deaf and hard of hearing students are a low incidence population that are thinly spread. By necessity, captioners who specialize in higher education are also thinly spread. It takes time for captioners to arrive on site and transcribe, which makes it hard to book them on demand for live speech transcription or dialogue for short periods of time. In general, CART captioners usually need at least a few hours advance notice, and prefer to work in one-hour increments in order to account for their commute time. While captioners no longer need to be physically present at the event they are transcribing since captioning services are increasingly being offered remotely [Kheir and Way 2007], students still might not be able to decide at the last minute to attend a lecture or stay after class to interact with their peers and teacher.

(2) *Content Knowledge, Availability and Cost*
Real-time captioning has become a required component in the court room from the 1980s [Downey 2006], in part due to the structured legal environments that was controlled by the judge. The vocabulary also tends to be standardized and it becomes easy for captioners to start with large pre-existing dictionaries as the vocabulary was generally circumscribed with the exception of witness testimony. As such, captioners were able to build and leverage large, standardized legal and courtroom dictionaries.
In contrast, few captioners are willing to work in a higher education environment where the vocabulary varies widely from class to class, with widely varying educational content and specialization. Therefore, in the higher education environment, unlike the legal courtroom, the captioner is not able to leverage preexisting dictionaries, but instead has to develop his or her own dictionary, which takes much time and can be error prone.
Captioners are more reluctant to be generalists equipped to handle a wide range of environments and content. There are few captioners who have the needed appropriate content knowledge to handle higher education lectures that usually have specialized vocabulary. Consequently, captioners are also often simply not available for many technical fields [Fifield 2001]. Remote captioning offers the potential to recruit captioners familiar with a particular subject (e.g., organic chemistry) even if the captioner is located far away from an event, but selecting for expertise further reduces the pool of captioners, which would lead to an increase in cost. Due to high barriers to entry, and thus scarcity, wages are high and captioners are expensive relative to interpreters.

Due to the lack of availability and their cost, the number of captioners in higher education remains tiny in comparison with sign language interpreters. At the end of 2012, according to the online database managed by the Registry of Interpreters for the Deaf (RID) online database, there are 9,611 certified interpreters in the United States. According to the online database maintained by the National Court Reporters Association (NCRA), there are 316 certified CART providers offering services in higher education in the United States. The captioners on average charge more than twice as much as interpreters. Captioners are economically competitive with interpreters only because they can work alone, while interpreters work in pairs, so the total cost to consumers is similar for both.

3.3. Lecture Audio Evaluation

We evaluated 240 YouTube verbatim captioned lecture videos. Our criteria for selecting lectures in this set was as follows: first, that the lecture be produced by an educational institution in the United States and that the lecture lasts for between 50 minutes to an hour. We also looked at the lecture location and ignored those that would result in too much geographic concentration; in other words, we aimed for an even geographic distribution in order to average for a meaningful nationwide speaking rate and intelligibility according to accent. We also searched and retrieved 40 captioned transcripts of television interviews posted to YouTube, spread evenly between eastern, midwestern and western television stations. Similarly, we searched for and retrieved 40 geographically distributed captioned demonstration transcripts. Our goal was to obtain uniform geographic distribution to minimize the impact of accents and local dialects on speaking rate and vocabulary.

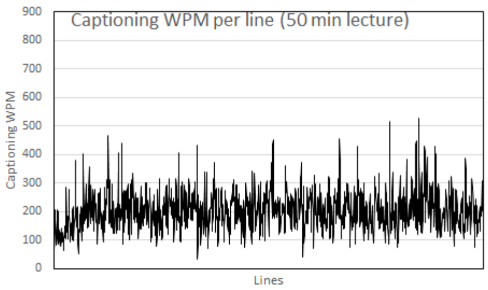


Fig. 5. Lecture: Words Per Minute (WPM)

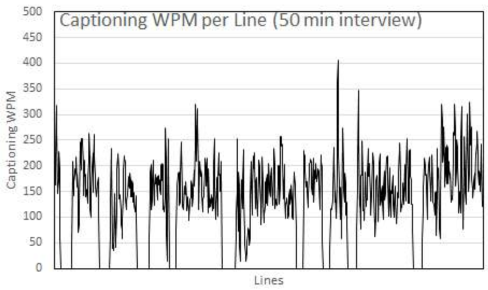


Fig. 6. Interview: Words Per Minute (WPM)

3.3.1. Speaking rate. The average speed of all of the lectures was 169.4 words per minute ($\sigma = 23.4$), which is close to 3 words per second. The average speeds for the demonstrations and interviews were very similar, at 163.8 words a minute ($\sigma = 28.4$) and 167.1 words a minute ($\sigma = 32.3$) respectively.

In hindsight, this is not so surprising, for people speak at their most comfortable speaking rate irrespective of content. The numbers suggest that the variation in speech flow was larger in the interviews and demonstrations sets than in the lecture set. In general, the lectures had longer bursts of rapid speaking in Figure 5. The interviews and demonstrations tended to have more quiet pauses than the lectures, and in these quiet pauses the speaking rate dropped close to zero, as shown in Figure 6. Every transcript had at least one speaking burst for several seconds of over 400 words a minute, and over a quarter had multiple speaker bursts of over 400 wpm.

In terms of speaking rates, the graphs show that even stenographers, let alone other typists cannot keep up their typing during peak speaking bursts, as few captioners are able to sustain typing rates faster than 300 wpm. In general though, speaking bursts do not seem to be a large hurdle, because stenographers can buffer their typing and average out the rapid and slow speaking periods. By "buffering" speaking bursts at the cost of increased delays, the stenographers eventually catch up during pauses or periods of slow speaking rates. Stenographers, and especially C-Print captioners can also condense the speech by summarization or by dropping non-essential words. However, these solutions can interfere with the comprehension of DHH readers.

3.3.2. Vocabulary size. Analysis of the lecture transcripts showed that the transcripts of the hour-long lectures had between 10,000 to 15,000 words ($\sigma = 13,810$ words). We also found that about one fifth of these words , a total of around 2,000 to 3,000 words ($\sigma = 2,421$ words) were unique, that is, were uttered only once. On the other hand, while hour long interviews and demonstrations had similar transcript sizes of around 10,000 to 15,000 words ($\sigma = 12,307$ words), the percentage of unique words in the transcript was half the number of unique words in typical lectures, at around 1,000 to 1,500 words ($\sigma = 1,241$ words). The larger dictionary size in lecture transcripts could be a challenge to process for both captioners and readers, as it is two to three times the typical size of TV program vocabulary. For classmates, who are already immersed in the classroom context, this would be less of a problem. Third, when we analyzed the vocabulary for spatial referents and pronouns, we found that these inherently visuospatial referents averaged around 12% of the total vocabulary for lectures, and around 9% for interviews and demonstrations. The higher percentage of missed non-verbal cues and information makes it more challenging for DHH students to follow the lecture, without the aid of external agents or technology.

3.3.3. Accuracy. Since all of the educational lectures had been captioned offline and disfluencies removed, nearly all of the transcripts had near perfect spellings, but not including punctuation. Running the transcripts through a dictionary program (iSpell) showed that the accuracy rate of offline captions was around 98%, as most of the YouTube transcripts were prepared with the help of stenographers. But even at 98% accuracy, a 50-minute lecture at 150 words a minute yields 7,500 words, with around 150 words in error, which is nearly three words a minute. Although not all of these errors are necessarily critical for comprehension during reading, many errors can derail the reading process, make it difficult for readers to recover from errors and ambiguousness and keep up with the flowing text. The eye-tracking qualitative observations noted that when grammatical errors were present, some readers paused at that place for up to a couple of seconds, while for syntax errors most readers did not pause at all and continued to scan forward. This variation in eye-tracking movements in the presence of caption errors shows dramatical variation by captioning error. This finding suggests that it is much harder to avoid mistakes and to instead steer the mistakes in such a way that the technical errors and cognitive load are minimized.

In summary, the lecture audio is characterized by rapid speaking rates, large vocabulary sizes, and many non-verbal cues and information. All of these features present challenges for current approaches to real-time captioning and presentation.

4. CAPTION READING IN THE CLASSROOM

We first evaluate the characteristics of higher education lecture audio and relate it to caption generation and readability. In succeeding studies, we evaluate the efficacy of collaborative captions by comparing these against Computer Aided Real-Time transcripts (CART) and Automatic Speech Recognition transcripts (ASR).

4.1. Transcript Readability Evaluation

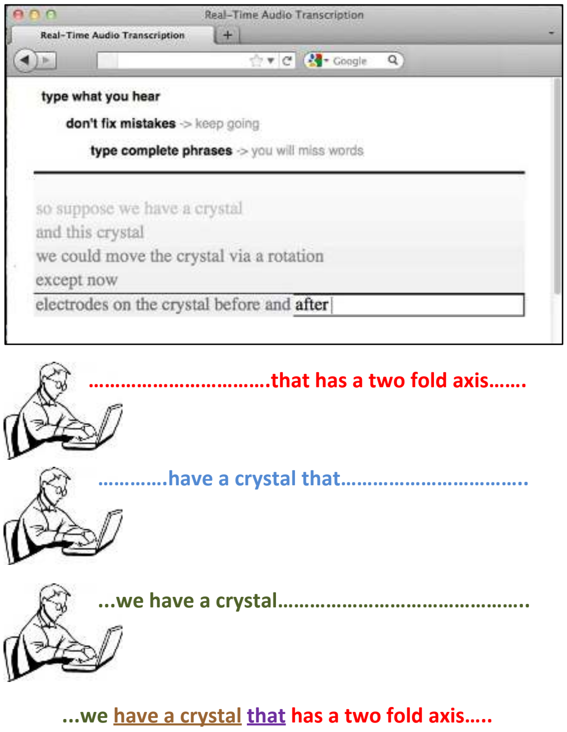
While the systems we are comparing (CART, ASR and Scribe) are able to score high on accuracy tests, this is not always a fair gauge of their utility. In this section, we report on the results of a readability evaluation with both deaf and hearing students using all three systems. Specifically, for dynamic text, the eye-gaze scan moves diagonally to both read the words on each line and to follow the lines as they move down, as each line is entered. On the other hand, for static text, the eye-gaze path is strictly horizontal for each line; after each line, the gaze returns to the start of the line and then goes down to the next line.

4.1.1. Transcript Generation. We recorded three transcriptions of an online lecture (OCW) in real-time using CART, Scribe, and ASR.

For CART, we hired an expert, professional real-time stenographer who charged $200 an hour to create a professional real-time transcript of the lecture. The captioner listened to the audio and transcribed in real-time. The mean typing speed was 181.3 WPM ($\sigma = 21.3$) with a mean latency of 4.2 seconds. We calculated latency by averaging the latency of all matched words.

For Scribe, we recruited 20 undergraduate students to act as non-expert captioners 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, as shown in Figure 7. The final transcript speed was 134.4 WPM ($\sigma = 16.7$), with a latency of 3.87 seconds.

For ASR we used Nuance Dragon Naturally Speaking 11, a high-end consumer automatic speech recognition (ASR) program. 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 soft-


Real–Time Audio Transcription
Real–Time Audio Transcription
Google

type what you hear

don't fix mistakes -> keep going

type complete phrases -> you will miss words

so suppose we have a crystal
and this crystal
we could move the crystal via a rotation
except now
electrodes on the crystal before and after

................................that has a two fold axis.......
............have a crystal that.................................
...we have a crystal...
...we have a crystal that has a two fold axis.....

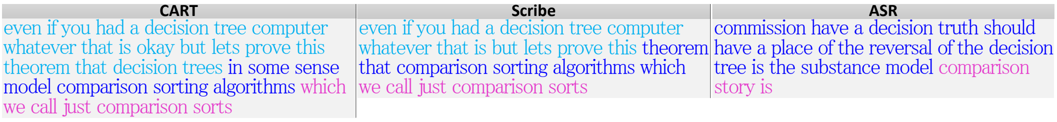
Fig. 7. The crowd captioning interface

ware loop to redirect the audio signal without resampling using SoundFlower[1], and a custom program to record the time when each word was generated by the ASR. The final transcript speed was 71.0 WPM ($\sigma = 29.7$) and had an average per-word latency of 7.9 seconds.

To evaluate the efficacy of real-time transcripts generated by crowds, we compared deaf and hearing user evaluations on their perceptions of the usability of Scribe transcripts against CART and Automatic Speech Recognition transcripts (ASR).

4.1.2. Study Demographics. We recruited 48 students for the study over two weeks to participate in the study and evenly recruited both deaf and hearing students, male and female. Twenty students were deaf, four 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 personal recruitment on campus. We asked students to contact a study assistant through email for appointments. All students were reimbursed for their participation. All deaf participants used visual accommodations for their classes.

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



CART	Scribe	ASR
even if you had a decision tree computer whatever that is okay but lets prove this theorem that decision trees in some sense model comparison sorting algorithms which we call just comparison sorts	even if you had a decision tree computer whatever that is but lets prove this theorem that comparison sorting algorithms which we call just comparison sorts	commission have a decision truth should have a place of the reversal of the decision tree is the substance model comparison story is

Fig. 8. Screenshots of Scribe, CART, and ASR output at three sequential points in time. ASR "flow," is not as smooth as the SCRIBE or CART flow. Scribe misses some content, but workers generally missed less important segments.

4.1.3. Study Methodology. Testing was conducted in a quiet room with a SensoroMotoric Instruments iView X Remote Eye-tracking Device that had an accuracy of less than 0.5 degrees. The students viewed transcripts on a 22 inch flat-screen monitor, as shown in Figure 10. Each person read a web page that explained the purpose of the study. Next, the students completed a short demographic questionnaire in order to determine eligibility for the test and gave informed consent. They then viewed a short 30-second introductory video to familiarize themselves with the process of viewing transcripts. Then the students watched 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 audio. The first video transcript contained the captions created by the stenographer. The second video transcript contained the captions created by the automatic speech recognition software. The third video transcript contained the captions created by the crowd captioning process. The total time for the study was about 15 minutes.

After the participants finished watching all three video clips of the real-time transcripts, they were asked to complete a questionnaire with three questions related to asking the participant for a Likert score rating for the captions they watched. The questions asked "How easy was it to follow the captions" for each caption they watched. In response, the participants were presented with a Likert scale that ranged from 1 through 5, with 1 being "Very hard" to 5 being "very easy". Then participants were asked to answer in their own words in response to the questions that asked participants for their thoughts about following the lecture through the transcripts. The answers were open-ended and many participants gave great feedback.

5. RESULTS

5.1. Readability Ratings

We ran a Wilcoxon signed-rank test to compare user preference ratings for collaborative captions to both CART and ASR captions. Users reported slightly greater ease in following the collaborative captions ($M = 3.15, SD = 1.06$) than for CART captions ($M = 3.09$, $SD = 1.23$), but this difference was not statistically significant ($Z = -.213$, $p = .830$). However, collaborative captions were significantly easier to follow compared to ASR captions ($M = 2.19$, $SD = 1.23$), ($Z = -3.98$, $p < .001$).

When we divided the students into deaf and hearing subgroups we found similar trends as well. There were still nonsignificant differences between collaborative and CART captions for both hearing ($Z = -.91$, $p = .41$) and deaf users ($Z = -.64$, $p = .58$). There was still a significant difference between collaborative and ASR captions for hearing users ($Z = -.91$, $p < .001$), and a marginal difference for deaf users ($Z = -.64$, $p = .06$).

In general, participants found it hard to follow ASR captions. The median rating for ASR was a 1, i.e., "Very hard". The qualitative comments indicated that many of them thought the captions were too choppy and had too much latency. In contrast to the low ratings for ASR, participants had higher median ratings of 4 for both CART and collaborative captions, indicating that they found it easier to follow them.

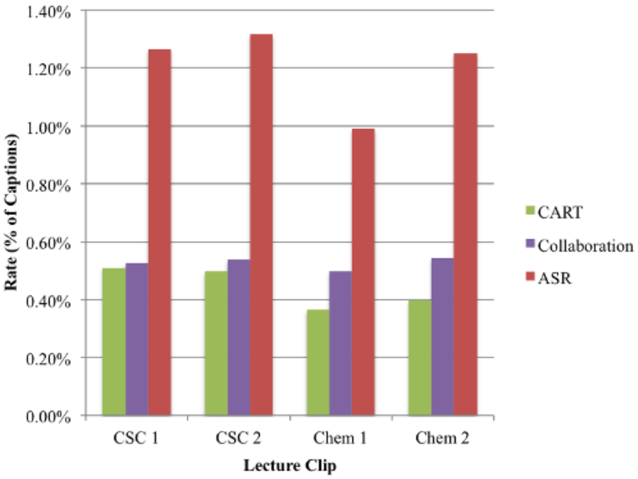


Fig. 9. A comparison of the flow for each transcript. The flow is the percentage of audio buffer that is processed and then displayed all at once. Both CART and collaborative captions exhibit smoother real-time text flow as compared to ASR transcript flow

Fig. 10. An example of a participant in our study viewing a display containing the captions generated by a stenographer.

5.2. Qualitative Comments

Qualitative comments from hearing students revealed that transcript flow (Figure 9) latency (Figure 11), and speed were significant factors in their preference ratings. For example, one hearing student (P31) commented on the CART captions as follows:

> “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 collaborative 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, a hard of hearing student (P19) commented on CART:

> “It seemed to lag quite a bit in my opinion. There were definitely some words that seemed to be typed out of context, which would cause me to pause, reread and make an educated guess as to what was being said.”

and for collaborative captions, the same hard of hearing student commented:

> “It was the most accurate and understandable. I could easily read and it definitely flowed and was much more coherent than the other two. It was accurate more than the others in my opinion, although, in any case, I’d much rather have been present to to listen to the speaker and be able to see what they were saying so I can match their speech to their lips as I lipread along to what I hear.”

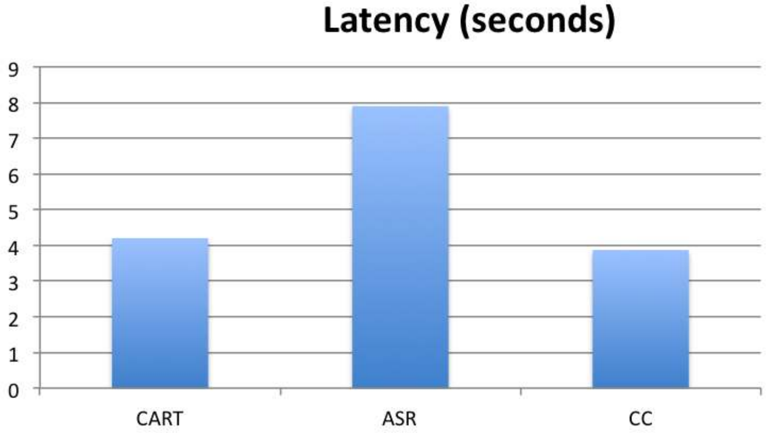


Fig. 11. A graph of the latencies for each transcript (CART, ASR and collaborative). CART and collaborative captions have reasonable latencies of less than 5 seconds, which allows students to keep up with class lectures, but not consistently participate in class questions and answers, or other interactive class discussion

Finally, a deaf student (P47) commented on CART as follows:

> “Some punctuation to show pauses and at ends of sentences would have been nice. Otherwise, it was very easy and fluid to follow.”

and for collaborative captions, the same deaf student commented:

> It was still very clear what was being said. Save for a few spelling errors and misplaced words, it was good.

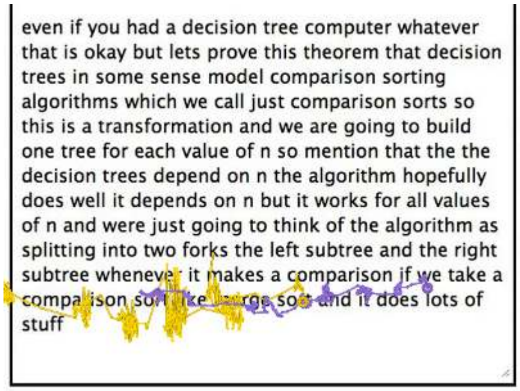
We coded user comments as being positive or negative. Although many of them exhibited advantages and disadvantages, comments were given a positive score if they conveyed overall satisfaction with the transcript. Five user comments were omitted from the CART analysis, four from ASR analyses, and seven from the collaborative captions analysis, either because the user did not make a comment, or because the user admitted that they did not have an opinion.

We compared the frequency of negative comments to positive comments using a chi-square of goodness-of-fit test. There was a significant difference between the number of negative (32) and positive comments (10) for CART captions ($\chi^2(1, N = 42)$, $p = .001$). There was also a similiar trend for ASR captions ($\chi^2(1, N = 43), p < .001$) with users speaking more negatively (38) than positively (5) regarding the captions. For the collaborative captions, there was a non-significant difference, ($\chi^2(1, N = 40)$, $p = .154$), although users had less negative evaluations (15) than positive ones (25).

5.3. Eye-Tracking Scan Path Analysis

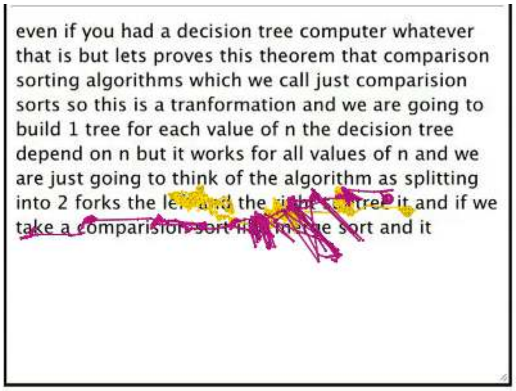
A qualitative survey of all participants’ eye-gaze scan paths as they read each kind of caption suggests that readers are most bothered by typing bursts, but are also significantly affected by unexpected words or errors, but not vocabulary or omissions.

5.3.1. CART. While watching CART captions, 16 participants repeatedly re-read the same words more than once, often many times, while all others never re-read. They fixated on the area where the next word would be typed. For those participants who re-read, the re-reading almost always occurred as soon as the transcript scrolled up. Although the active line did not shift, the scrolling of the previous (surrounding) text



even if you had a decision tree computer whatever
that is okay but lets prove this theorem that decision
trees in some sense model comparison sorting
algorithms which we call just comparison sorts so
this is a transformation and we are going to build
one tree for each value of n so mention that the the
decision trees depend on n the algorithm hopefully
does well it depends on n but it works for all values
of n and were just going to think of the algorithm as
splitting into two forks the left subtree and the right
subtree whenever it makes a comparison if we take a
comparison sort like merge sort and it does lots of
stuff

Fig. 12. Professional Captions
Purple - hearing reader
Yellow - deaf reader



even if you had a decision tree computer whatever
that is but lets proves this theorem that comparison
sorting algorithms which we call just comparision
sorts so this is a tranformation and we are going to
build 1 tree for each value of n the decision tree
depend on n but it works for all values of n and we
are just going to think of the algorithm as splitting
into 2 forks the left and the right subtree it and if we
take a comparision sort like merge sort and it

Fig. 13. Collaborative Captions
Purple - hearing reader
Yellow - deaf reader

appeared to affect participants' ability to keep track of their reading focus. Within this group, 2 participants moved their fingers in front of the screen to guide them in keeping track with their reading location, even though participants had been told to not move their body in order to maintain accuracy in the gaze measurements using the eye-tracker. If the participants are slow readers, the line they are reading will have shifted up, which disrupts their reading focus. Figures 12 and 13 show typical scan paths for both CART and collaborative captions.

5.3.2. ASR. While watching ASR captions, 44 participants repeatedly re-read the same words more than once, while the rest never re-read. They normally stayed fixated on the area where the next word would be typed. In contrast to the reading process for CART or collaborative captions, this re-reading occurred more often when there was no following text for a few seconds. No participants moved their fingers in front of the screen. This suggests that this strategy of visual guidance of the active text was not helpful when presented with "bursty" text that is often out of synchronization with the corresponding audio.

5.3.3. Collaborative Captions. While watching collaborative captions, 11 participants repeatedly re-read the same words more than once, often many times, while all others never re-read. They normally stayed fixated on the area where the next word would be typed. The re-reading pattern was similar to those who were reading CART captions in that they re-read whenever the text scrolled up. The same 2 participants who moved their fingers in front of the screen while reading the CART captions did the same for collaborative captions. This suggests that participants use the same reading strategies for both CART and collaborative captions.

5.3.4. Caption Reading Pace. The optimal temporal and spatial presentation of the text on the screen most likely differs from either the speaker's pace (170 WPM) or reader's static text pace (280 WPM) [Hasbrouck and Tindal 2006]. However, the process of reading real-time captions presented as scrolling text according to the speaker's pace, is very different from the process of reading static text controlled by the reader's reading pace. With static text, the reader knows when and where to move to the next word. On the other hand, while reading real-time captions, the sequential words appear unpredictably, and new text replaces the older text. Another shift occurs when the typed words have filled a line, and the screen scrolls up.

6. DISCUSSION

Although we did not test participants' comprehension of the lectures, we analyzed readability of the lectures through three methods: Likert ratings, qualitative comments and qualitative eye-tracking scan paths. The findings from each of the methods were consistent with each other. The low ratings, negative comments and frequent re-reading eye-gaze patterns for ASR reflect the fact that captioning flow is very important to the readers of captions. As shown in figure 9, ASR had the worst caption flow of all generated captions, and as such had the most negative ratings.

Although verbatim captions may capture all information, given the fact that spatially oriented caption text does not capture all facets of temporally oriented speech, it is not surprising that fewer participants re-read often, and gave more positive comments towards group summarized collaborative captions versus CART captions.

Overall the comments, Likert ratings and eye-tracking analysis show that the participants appeared to like the slower and more smooth flowing collaborative captions rather than the faster and less smooth CART captions, as the workers do not type in every disfluency.

When multiple typists collaborate, confusing and unexpected words are weeded out. As a result, all participants found the captions more easy to read, as shown by the smoother and less jerky eye-gaze scan path in Figure 13.

The finding that fewer participants frequently re-read collaborative captions than CART captions suggests that they had an easier time following and predicting when the captions would scroll up. This may be more important for reading real-time higher education lecture transcripts, for our survey found that higher education lecture captions had longer and less frequently encountered words than television-style interviews or demonstrations. This can aggravate the mismatch between speaking and reading rates, and increase the importance of predictability in the reading process.

In regards with the marginal difference in preference ratings between collaborative and ASR for DHH students, we attribute this to the wide range of reading abilities among DHH students in comparison with hearing students. Since DHH students on average have a wider range of reading skills [Jensema 1996], it appears that slower captions for the less proficient readers in this group do 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 as judged by the quality of their typed comments, commented that the captions were too slow. We hypothesize that these students are focusing on key-words and ignore the rest of the text. DHH students are not usually exposed to aural disfluencies and are more likely to be able to read and accept disfluencies in text. DHH students may also be more used to variable input 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. Additionally, our tests found that the errors made by ASR differ from those made by humans: ASR tends to replace words with others that are phonetically similar, but differ in meaning, while humans tend to replace words with ones that have similar meanings. This is often due to workers remembering the meaning but forgetting the exact wording as they type.

Many participants reported that the captions were hard to read when there was a lag in displaying the current text. This significant impact of captions' temporal delay reinforces the importance of minimizing translation and display delay. This is especially true for ASR captions which have both high temporal delay and jitter that worsen the temporal contiguity between the speech and captions. The participant comments support Mayer's temporal contiguity principle [Mayer et al. 2001] that states that people learn better when corresponding words and demonstration is presented simultaneously with the captions rather than successively. In other words, when they occur at the same time, the learner is better able to hold mental representations of both in working memory and build connections. Similarly, Meyer's coherence principle [Mayer et al. 2001] states that people learn better when extraneous material is excluded rather than included, i.e., the extraneous material competes for cognitive resources. Compared with ASR, we observed that collaborative captions filtered out speech disfluencies, which are extraneous to text.

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. Merging multiple caption streams into a single stream appears to smooth the flow of text as perceived by the reader, compared with the flow of text in CART captioning or ASR captions.

6.1. Types of Errors in Caption Generation

We have shown that deaf and hearing students alike prefer collaborative 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.

The student comments mentioned above indicate that collaborative captioning handles the transcription of utterances with many disfluencies well compared to the reading of prepared materials. It appears that one of the key advantages to using human captioners instead of ASR is the types of errors which are generated by the 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 collaborative captions more usable than automated systems, even in cases where there may be minimal difference in measures, such as accuracy and coverage. Figure 8 gives an example of a clip where the ASR answer provided the reader with almost no help in understanding the meaning of the words being spoken.

6.2. Collaborative Caption Advantages

We think the collaborative captioners 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 captioners are working simultaneously, it can be regarded as a group vote for the most important information. In this way, groups of non-expert captioners may be 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 WPM [Jensema 1996], coverage and flow become more important. In other words, assuming a limiting reading rate (especially for dense lecture information), the comments show that students prefer condensed material so that they can maintain reading speed/flow to keep up with the instructor.

Collaborative captions also show the potential to have more accurate technical vocabulary than either ASR or CART captions. 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. Collaborative captioning enables the software and users to effectively adapt to a variety of environments that a single captioner and dictionary cannot handle. Classmates are also likely to be more familiar with the topic being discussed, and to be used to the speaker's style. This approach can scale in terms of classmates and vocabulary. The collaborative captioning transcript, as an average of multiple streams from all captioners, is likely to be more consistent and have fewer surprises than any single captioner. As each captioner has to type less, the captions are also more likely 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; they all get a high coverage, high quality transcript that none of them could normally type on their own.

Overall, collaborative captioning is able to handle lectures that are characterized by high speaking rates with frequent pauses and rapid speed bursts, as well as by large vocabularies, and multiple speakers. First, it is scalable enough to handle speaking speed bursts. Second, it is able to slightly reduce the transcription rate through group agreement on important words and eliminating non-words. Third it is able to display a more predictable caption presentation by combining the group's data entry ability.

7. FUTURE WORK

The participants' perception of the different kinds of errors between sources of captions shows that not all errors are equally important. The comments reveal that the errors made by collaborative captioning are much easier for users to understand and adapt to than those made by ASR. They also revealed the importance of flow, and suggest that future work on presentation is warranted.

7.1. Multiple Inputs and Processing

Unlike CART, Scribe is able to handle multiple simultaneous speakers by enabling some workers to focus on transcribing one of the speakers, while other workers focus on transcribing other speakers. Any single input approach would find it almost impossible to transcribe simultaneous speakers. It may also be possible to leverage their understanding of the context that content is spoken in to generate better-then-verbatim captions. In our experiments, it was evident that skipping some errors actually results in a more readable transcript. If this can be done without removing content or easy-to-follow phrasing, then it might improve the caption reading experience for users.

7.2. Enhanced Captions

The scalability of crowd captioners opens new vistas: they can capture and convey visually all auditory information related to the lecture or event. Most non-speech markers, like accent, tone, and timbre are not shown in text. The reader has to recreate these non-speech markers while reading, and this process may not be accurate. While enhanced captions can represent some of these non-speech markers [Lee et al. 2007], there has been no attempt to do this in real-time. The scalability of Scribe offers great potential in enhancing captions in real-time.That is, they could not only show actual speech, but also non-language sounds including environmental noises or tone of the spoken words, e.g. irony or sarcasm.

7.3. Caption Flow and Sources of Errors

Spoken language often flows differently than written text. Speakers pause, change subjects, and repeat phrases, and all of these make exact transcripts difficult to read without inflection and timing information. Figure 8 shows the output of ASR, CART,

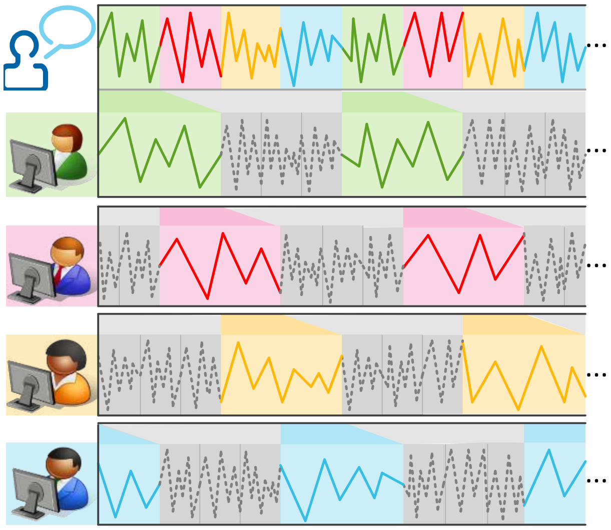


Fig. 14. *TimeWarp* forwards different "warped" versions of the audio to each worker. Workers hear a slowed down version of the content they are supposed to caption, while they hear the rest sped up. This increases the quality of workers' output by asking them to complete an easier task without losing context. The start of each segment is aligned with the original audio, allowing the crowd to collectively caption in real-time.

and Scribe over three time windows. Captioning methods such as C-Print paraphrase speech to keep up, cleaning up sentences to be easier to read, but also skip content. ASR often produces nonsensical errors. Scribe can use a language model to help correct sentences that have inconsistencies in order to make the final text more readable, and even customize these models to users.

Factors such as complex vocabulary, omission or excessive captioning speed can disrupt the reading process, as shown in Figure 12. Deaf readers are more used to ignoring or skipping disfluencies due to their exposure to reading TV captions, and sometimes classroom captions. They also don't hear or internalize the intonation or stress that can contribute to the disfluencies. As such, they tend to be able to ignore the omissions more than hearing readers, who are not as exposed to reading disfluent English.

7.4. Worker Interfaces for Improving Flow

The improved flow seen in Scribe captions is a result of using multiple workers simultaneously, reducing the potential for accrued latency. This division might also be leveraged to further improve worker's ability to provide captions with consistent flow. For example, TimeWarp [Lasecki et al. 2013] slows the audio playback rate for content that workers are asked to caption, but keeps up with real-time speech by increasing the playback speed of content workers are asked to listen to in order to maintain context but not capture as shown in Figure 14. This approach encourages workers to type words as they hear them, instead of waiting to first listen to all of the content in a segment, then quickly transcribe it [Lasecki et al. 2013]. This not only improves latency, but also can help to improve flow. One of our future goals is to develop new collective approaches such as TimeWarp and adaptive time windows for individual workers [Murphy et al. 2013] that can improve workers' ability to provide captions with better flow, while also maintaining or improving accuracy and latency.

7.5. Partial Automation

We can take advantage of the different types of errors that non-expert captioners and ASR make. Specifically, since humans generally make mistakes involving similar meaning or missed words, while ASR makes mistakes involving words that sound similar, it is likely that there is a way to combine the two to reduce the final observed error rate. It might also be possible to use one source to generate a 'quick' answer, then over the next few seconds, use the other to correct mistakes.

7.6. Ubiquitous Real-Time Captions

Prior work in crowdsourcing has shown that workers can be recruited from online crowd marketplaces, such as Amazon's Mechanical Turk, within seconds [Bigham et al. 2010; Bernstein et al. 2011], potentially enabling on-demand captioning services. We will also investigate social collaboration incentives that encourage students to type and work together in the classroom. These ubiquitous personal captions can be made available from users' smart phones, and has great potential in empowering deaf students to take control of their own learning experience in mainstreamed classrooms and beyond.

The transcription of audio information provides benefits to a wide range of users in addition to those who are deaf or hard of hearing. For example, transcripts allow audio data to be manipulated, archived, and retrieved more efficiently because text-based search is more expedient than audio-based search. Reading static text is faster for most people than listening to the auditory equivalent. Access to transcription also offers advantages to language learners, or to individuals with learning disabilities who work better with written language than spoken language.

8. CONCLUSION

A challenge in developing new approaches for real-time captioning is that it can be difficult to quantify whether the captions have been successful. As demonstrated here, accessibility of real-time captioning is dependent on much more than just measuring error rate. At minimum, we have to quantify the naturalness of errors, as well as their regularity, latency and flow. These concepts are difficult to capture automatically, which makes it hard to generate reliable comparisons between different approaches. We have analyzed characteristics of higher education lectures, namely speaking rate and word frequency. By taking into account these characteristics, we increase our ability to quantify and improve systems for real-time captioning.

We also evaluate collaborative captioning by groups of non-experts as a new source of real-time captions that improves on traditional approaches in terms of many of these factors. Our qualitative, quantitative and eye-tracking measures all show that both hearing and deaf participants preferred and followed collaborative captions better than those generated by ASR or CART due to the more consistent *flow* of the resulting captions. Combined with the crowd, this approach also holds the potential for making real-time captioning a ubiquitous accessibility service in the near future.

ACKNOWLEDGMENTS

The authors thank the participants for their time and feedback. They also thank the real-time captioner and sign language interpreter for their work in making the lecture accessible to deaf and hard of hearing students. This work was supported by NSF Awards #12070356 and #1218209, the National Technical Institute for the Deaf ("Multiple Viewing Perspectives," 2011) at Rochester Institute of Technology ("Accessible Viewing Device: Enhancing Classroom Access," 2012), and by a Microsoft Research Ph.D. Fellowship.

REFERENCES

Ahmed Sabbir Arif and Wolfgang Stuerzlinger. 2009. Analysis of text entry performance metrics. In *2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH)*, IEEE (Ed.). IEEE, Toronto, ON, 100–105. DOI:http://dx.doi.org/10.1109/TIC-STH.2009.5444533

Susan Aud, William Hussar, Grace Kena, Kevin Bianco, Lauren Frohlich, Jana Kemp, and Kim Tahan. 2011. The Condition of Education 2011. NCES 2011-033. (2011).

Keith Bain, Sara H. Basson, and Mike Wald. 2002. Speech recognition in university classrooms. In *Proceedings of the fifth international ACM conference on Assistive technologies - Assets '02*. ACM Press, New York, New York, USA, 192. DOI:http://dx.doi.org/10.1145/638249.638284

Michael S Bernstein, Joel R 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, to appear. DOI:http://dx.doi.org/10.1145/1866029.1866080

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

J Boyd and E.A. Vader. 1972. Captioned television for the deaf. *American Annals of the Deaf* 117, 1 (1972), 34–37.

Xiaodong Cui, Liang Gu, Bing Xiang, Wei Zhang, and Yuqing Gao. 2008. 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*, IEEE (Ed.). IEEE, Las Vegas, NV, 5121–5124. DOI:http://dx.doi.org/10.1109/ICASSP.2008.4518811

Greg Downey. 2006. Constructing "Computer-Compatible" Stenographers: The Transition to Real-time Transcription in Courtroom Reporting. *Technology and Culture* 47, 1 (2006), 1–26. DOI:http://dx.doi.org/10.1353/tech.2006.0068

L B Elliot, M S Stinson, B G McKee, V S Everhart, and P J Francis. 2001. College Students' Perceptions of the C-Print Speech-to-Text Transcription System. *Journal of deaf studies and deaf education* 6, 4 (Jan. 2001), 285–98. DOI:http://dx.doi.org/10.1093/deafed/6.4.285

M Bryce Fifield. 2001. Realtime Remote Online Captioning: An Effective Accommodation for Rural Schools and Colleges. In *Instructional Technology And Education of the Deaf Symposium*. NTID, Rochester, NY, 1–9.

Jan Hasbrouck and Gerald A. Tindal. 2006. Oral Reading Fluency Norms: A Valuable Assessment Tool for Reading Teachers. *The Reading Teacher* 59, 7 (April 2006), 636–644. DOI:http://dx.doi.org/10.1598/RT.59.7.3

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

Carl Jensema. 1998. Viewer reaction to different television captioning speeds. *American annals of the deaf* 143, 4 (Oct. 1998), 318–24. http://www.ncbi.nlm.nih.gov/pubmed/9842059

Carl J Jensema, Ramalinga S Danturthi, and Robert Burch. 2000. Time spent viewing captions on television programs. *American annals of the deaf* 145, 5 (Dec. 2000), 464–8. http://www.ncbi.nlm.nih.gov/pubmed/11191825

Amy B Jordan, Anne Albright, Amy Branner, and John Sullivan. 2003. *The state of closed captioning services in the United States*. Technical Report. National Captioning Institute Foundation, Washington, DC. 1–47 pages.

Richard Kheir and Thomas Way. 2007. 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 (ITiCSE '07)*. ACM Press, New York, New York, USA, 261–265. DOI:http://dx.doi.org/10.1145/1268784.1268860

Dennis H. Klatt. 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America* 59, 5 (1976), 1208. DOI:http://dx.doi.org/10.1121/1.380986

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 Press, New York, New York, USA, 71–78. DOI:http://dx.doi.org/10.1145/2384916.2384930

Raja S. Kushalnagar, Walter S. Lasecki, and Jeffrey P. Bigham. 2013. Captions Versus Transcripts for Online Video Content. In *10th International Cross-Disclipinary Conference on Web Accessibility (W4A)*, ACM (Ed.). ACM Press, Rio De Janerio, Brazil, 41–45.

Harry G Lang. 2002. Higher education for deaf students: Research priorities in the new millennium. *Journal of Deaf Studies and Deaf Education* 7, 4 (Jan. 2002), 267–280. DOI:http://dx.doi.org/10.1093/deafed/7.4.267

Walter S. Lasecki and Jeffrey P. Bigham. 2012. Online quality control for real-time crowd captioning. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility - ASSETS '12 (ASSETS '12)*. ACM Press, New York, New York, USA, 143–150. DOI:http://dx.doi.org/10.1145/2384916.2384942

Walter S. Lasecki, Christopher D. Miller, and Jeffrey P. Bigham. 2013. Warping Time for More Effective Real-Time Crowdsourcing. In *Proceedings of the 2013 annual conference on Human factors in computing systems - CHI '13 (CHI '13)*. ACM Press, New York, New York, USA, 2033–2036.

Walter S. Lasecki, Christopher D. Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja S. Kushalnagar, and Jeffrey P. Bigham. 2012. Real-time captioning by groups of non-experts. In *Proceedings of the 25th annual ACM symposium on User interface software and technology - UIST '12*. ACM Press, New York, New York, USA, 23–34. DOI:http://dx.doi.org/10.1145/2380116.2380122

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 (UIST '11)*. ACM, New York, NY, USA, To Appear.

Daniel G. Lee, Deborah I. Fels, and John Patrick Udo. 2007. Emotive captioning. *Computers in Entertainment* 5, 2 (April 2007), 11. DOI:http://dx.doi.org/10.1145/1279540.1279551

Marc Marschark, Patricia Sapere, Carol Convertino, and Jeff Pelz. 2008. Learning via direct and mediated instruction by deaf students. *Journal of Deaf Studies and Deaf Education* 13, 4 (Jan. 2008), 546–561. DOI:http://dx.doi.org/10.1093/deafed/enn014

Richard E Mayer, Julie Heiser, and Steve Lonn. 2001. Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology* 93, 1 (2001), 187–198. DOI:http://dx.doi.org/10.1037/0022-0663.93.1.187

Matthew J. Murphy, Christopher D. Miller, Walter S. Lasecki, and Jeffrey P. Bigham. 2013. Adaptive time windows for real-time crowd captioning. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems (CHI EA '13)*. ACM, New York, NY, USA, 13–18.

Yingxin Pan, Danning Jiang, Lin Yao, Michael Picheny, and Yong Qin. 2010. Effects of automated transcription quality on non-native speakers' comprehension in real-time computer-mediated communication. In *Proceedings of the 28th international conference on Human Factors in Computing Systems - CHI '10 (CHI '10)*. ACM Press, New York, New York, USA, 1725–1731. DOI:http://dx.doi.org/10.1145/1753326.1753584

Ross Stuckless. 1999. Recognition means more than just getting the words right: Beyond accuracy to readability. *Speech Technology* 1, Oct (Oct. 1999), 30–35.

Frank Thorn and Sondra Thorn. 1996. Television captions for hearing-impaired people: a study of key factors that affect reading performance. *Human factors* 38, 3 (Sept. 1996), 452–63. http://www.ncbi.nlm.nih.gov/pubmed/8865768

United States Supreme Court USSC. 2012. Boyer v. Lousiana, 11-9953, Appellants' Original Transcript. (2012).

Mike Wald. 2005. 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*. IEEE, Indianapolis, IN, 22–25. DOI:http://dx.doi.org/10.1109/FIE.2005.1612286

Jacob O. Wobbrock and Brad A. Myers. 2006. Analyzing the input stream for character- level errors in unconstrained text entry evaluations. *ACM Transactions on Computer-Human Interaction* 13, 4 (Dec. 2006), 458–489. DOI:http://dx.doi.org/10.1145/1188816.1188819

Hisao Yamada. 1980. A Historical Study of Typewriters and Typing Methods: from the Position of Planning Japanese Parallels. *Journal of information processing* 2, 4 (1980), 175–202.