Real-Time Captioning by Non-Experts with Legion Scribe

Walter S. Lasecki[1], Christopher D. Miller[1], Raja Kushalnagar[2], and Jeffrey P. Bigham[3,1]

Computer Science, ROC HCI[1]
University of Rochester
{wslasecki,cmill32}@cs.rochester.edu

Computer Science, NTID[2]
Rochester Institute of Technology
rskics@rit.edu

HCII[3]
Carnegie Mellon University
jbigham@cmu.edu

ABSTRACT

Real-time captioning provides people who are deaf or hard of hearing access to speech in settings such as classrooms and live events. The most reliable approach to provide these captions is to recruit an expert stenographer who is able to type at natural speaking rates, but they charge more than $100 USD per hour and must be scheduled in advance. We introduce Legion Scribe (Scribe), a system that allows 3-5 ordinary people who can hear and type to jointly caption speech in real-time. Each person is unable to type at natural speaking rates, and so is asked only to type part of what they hear. **S**cribe automatically stitches all of the partial captions together to form a complete caption stream. We have shown that the accuracy of **S**cribe captions approaches that of a professional stenographer, while its latency and cost is dramatically lower.

Categories and Subject Descriptors

H.5.2 [Information interfaces and presentation]: User Interfaces. - Graphical user interfaces.

General Terms

Design, Human Factors

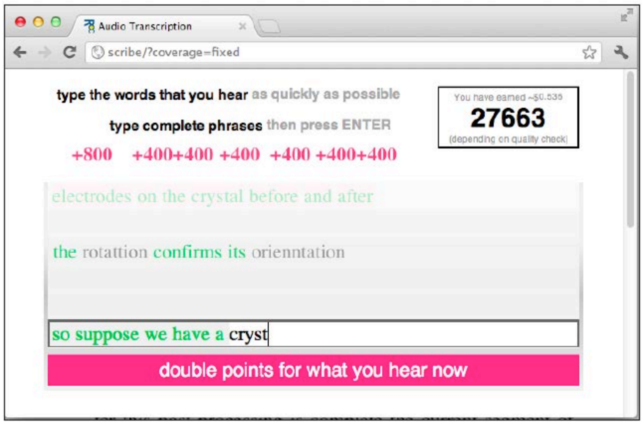
Keywords

Captioning, Real-time human computation, deaf, hard of hearing, crowdsourcing

1. LEGION SCRIBE

Real-time captioning provides a verbatim copy of spoken content with a target latency of less than 5 seconds, providing access to live events for deaf and hard of hearing people (e.g. mainstream lectures), an immediate transcript in the courtroom, and federally mandated access to live TV.

Currently, the most reliable solution is to hire professional stenographers who provide captions for $100-300/hour, depending on their skill level. In addition to their high cost, stenographers must be arranged several days in advance and can only be scheduled in one-hour blocks.



type the words that you hear as quickly as possible

type complete phrases then press ENTER

+800 +400+400 +400 +400 +400+400

You have earned ~$0.535
27663
(depending on quality check)

electrodes on the crystal before and after

the rotattion confirms its orienntation

so suppose we have a cryst

Figure 1: The worker interface encourages workers to type audio by locking in words soon after they are typed. To encourage typing specific segments, visual and audio cues are given, and the volume of the audio is reduced during off periods, while rewards are increased for on periods.

Real-time captioning is difficult because even fast typists cannot type at the 250 words per minute (wpm) necessary to keep up with natural speech (stenographers train to use special keyboards and input phonemes instead of individual letters). Meanwhile, automatic speech recognition (ASR) only captures about 40% of the speech in real settings and confuses readers by making seemingly random errors [1].

Legion Scribe allows a group of 3-5 people who can hear and type at an ordinary rate to collectively caption speech in real-time (Figure 1), for 20-30% the cost of an expert stenographer. These non-expert captionists can be recruited on-demand for as long or as short as the user needs.

Scribe uses multiple people typing what they hear (using the interface shown in Figure 1) and then stitches the pieces back together using an algorithm based on Multiple Sequence Alignment. This approach allows as few as 3 average typists to match the performance of an expert. Anyone who can listen and type can contribute without needing special training. For instance, work-study students (paid around $10 / hour) could fill this role in many classrooms. In our experiments, we also recruited workers cheaply and on-demand from crowdsourcing marketplaces like Amazon Mechanical Turk, Mobile Works, and oDesk.

ASSETS '13, October 21 - 23 2013, Bellevue, WA, USA
Copyright 2013 ACM 978-1-4503-2405-2/13/10…$15.00.

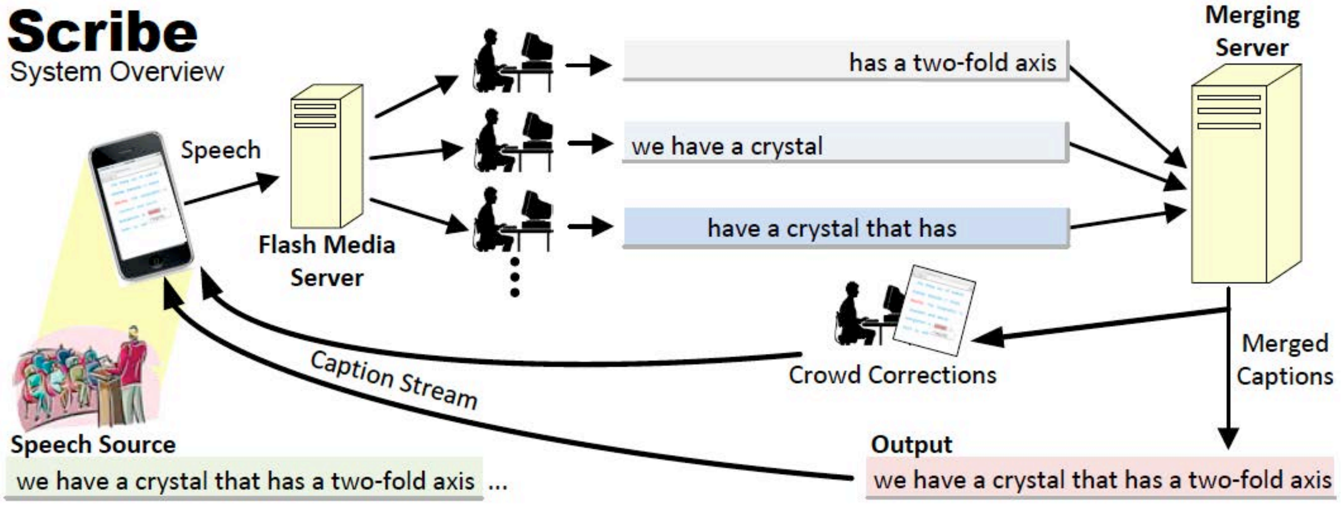
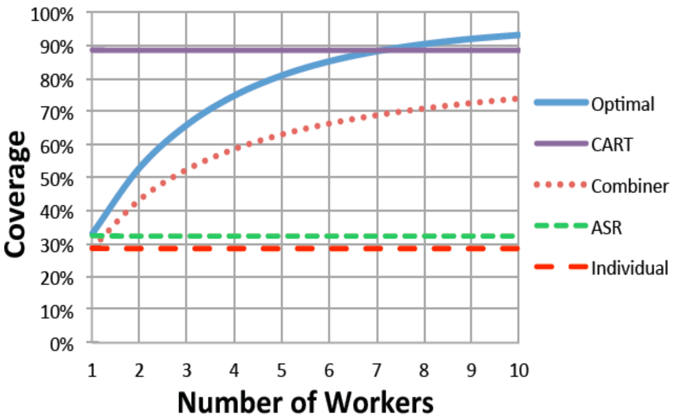


Figure 2: SCRIBE allows users to receive speech captions on their phone or mobile device. The audio is sent to multiple non-expert captionists who use the SCRIBE web-based interface to caption as much of the audio as they can in real-time. Partial captions are then merged into a final output stream, which is in turn forwarded back to the user.

We have also shown that by systematically slowing down and speeding up the audio for individual workers we can improve both precision and recall by more than 10% [2]. This is the *TimeWarp* approach to real-time human computation. Our interface (Figure 2) coordinates different workers so they type different portions of the streaming audio while maintaining the context of all of the speech.

2. PERFORMANCE

Scribe currently comes close to the performance of stenographers in terms of *coverage,* how many of the words in the ground truth appear in the final output stream, and *precision*, how many of the output words are correct. Coverage is shown in the graph below. Tests have shown Scribe is currently able to reach a precision of 84.8% of that of a professional. We expect that over time Scribe will become even more competitive and might be able to even surpass the performance of stenographers in terms of both coverage and precision. Additionally, TimeWarp also showed almost 20% improvement in latency, indicating multiple workers might also be able to outperform a single expert even in terms of speed, given the right workflow.



3. POST-PROCESSING

Once the captions are forwarded back to the users, they have the ability to fix any mistakes that can be corrected using context, via the user interface. This means the results shown here can be further improved upon. Because every user in a given session can contribute, this phase is also a collaborative process, but with a different set of workers.

The last component is the manner in which we present results to workers. By using multiple captionists, it is less likely that the system falls very far behind the speaker. This means that compared to a single stenographer, the flow of the text is more consistent. This likely helps users read content more easily, and is one reason why the captions produced by Scribe, while not perfect, can actually be preferred to professionals.

4. CONCLUSION

We have presented **S**cribe, a reliable human-powered approach for providing on-demand real-time captioning at low cost. Scribe uses multiple individual captionists, each of whom type a piece of what they hear, and then stitches the partial captions back together automatically. Scribe performs competitively with current approaches, but for a fraction of the cost to operate.

5. REFERENCES

[1] W. S. Lasecki, C. D. Miller, A. Sadilek, A. Abumoussa, D. Borrello, R. Kushalnagar, and J. P. Bigham. Real-time captioning by groups of non-experts. In *Proceedings of UIST 2012*. p23-33.

[2] W. S. Lasecki, C. D. Miller and J. P. Bigham. Warping Time for More Effective Real-Time Crowdsourcing. In *Proceedings of CHI 2013*. p2033-2036.

[3] R. Kushalnagar, W.S. Lasecki, J.P. Bigham. A Readability Evaluation of Real-Time Crowd Captions in the Classroom. In *Proceedings of ASSETS 2012*. p71-78.