In D. Reitter & F. E. Ritter (Eds.), Proceedings of the 14th International Conference on Cognitive Modeling (ICCM 2016). University Park, PA: Penn State.

An Update on Automatic Transcription vs. Manual Transcription

Frank E. Ritter (frank.ritter@psu.edu)
Catherine Bouyat (cmb6187@gmail.com)
Kaitlyn Ekdahl (kaitlyn.ekdahl@gmail.com)
Daniel Guzek (dan.guzek@psu.edu)
College of Information Sciences and Technology, Penn State University
University Park, PA 16802 USA

Abstract

Cognitive modelers have long used verbal protocol analysis to gather data to test their models. Recently developed tools offer support for automatic transcription of audio. In this short paper, we compare the time it takes to transcribe a video done (a) exclusively with Google's subtitle tool, (b) corrected from Google's subtitle tool, and (c) done completely by hand. We found that using the subtitle tool alone can yield too high an error rate, correcting Google subtitles took about 2.5x the video length, and transcribing completely by hand took approximately 11x the video length. We can thus recommend using Google subtitles as a starting point for verbal transcription as it offers a useful speed up in transcription. In addition, we can recommend when and how to use Google subtitles, and the use of headphones and automatic spelling correction in text editors.

Keywords: Transcription, Google closed captioning, YouTube, Verbal protocol analysis

Introduction

Cognitive modelers have long used verbal protocols (Larkin, McDermott, Simon, & Simon, 1980; Newell & Simon, 1972). The transcription of these protocols can be problematic, often taking 10x longer to transcribe than record (Ericsson & Simon, 1993; Ritter & Larkin, 1994). Thus, there is interest in automatic protocol transcription systems.

In this paper we examine a new tool to perform automatic verbal protocol transcription: YouTube's automated closed captioning service. We start to explore how to use the service more efficiently by examining how long it takes to transcribe an example video and note some lessons about how to assist transcription.

Method

In this pilot study, we took an example 47-minute video and used three approaches to transcribe its content: we copied the transcript created by YouTube's closed captioning (CC) service directly from YouTube; we copied YouTube's transcript and had a human coder manually correct it while listening to the video; and we manually transcribed a portion of the video from scratch. The human coder stopped out of frustration after approximately three hours of manual transcription, having successfully transcribed only 15 minutes of content, so we use that amount in our analyses.

Apparatus and Material

We used Google's CC service on YouTube to transcribe a video of a seminar presentation (https://www.youtube.com/watch?v=RcZU-fb0Q10), chosen somewhat arbitrarily as an example of naturalistic, public speech rather than a strictly concurrent verbal task protocol.

The coders each accessed the video on a PC and Mac laptop. They used either a Google Chrome (Mac) or Mozilla Firefox (PC) browser to view the video, and Microsoft Word (PC) or TextEdit (Mac) to hold and edit the transcripts.

Participants/Coders

The two coders were undergraduates in the College of IST at Penn State, whose first language is English. They have taken multiple courses in HCI and worked as research assistants for at least six months.

Design and Procedure

Both coders were each given the link and asked to transcribe the audio. Coder 1 (PC, arbitrarily chosen) worked first with the YouTube CC-generated transcription and edited the transcription as they watched the video. Coder 2 (Mac, arbitrarily chosen), transcribed by hand.

During the transcription process, the coders each used two windows—a browser and a text editor—on their own laptops. Headphones were also used, and both transcriptions were done in quiet spaces.

Results and Discussions

Table 1 shows an example of the automatic transcription for both speakers involved in the example video.

Table 1. Example of the unedited automatic transcription.

Example of unedited transcription from Speaker 1

0:03 ok once we get started my neighbors call for a research professor I have been

0:16 here for various events if you are out for those of you that don't have a lot

Example of unedited transcription from Speaker 2

3:17 this point it becomes how could I ever that confused but I was very green at

3:20 that point and so I'm just gonna talk a little bit about who's who in the zoo is

Table 2 presents the time to transcribe the video, including that the transcription time using the YouTube CC editing is nearly instantaneous and would be constant across length.

The time to correct the video's transcript takes about 4x the video's length. The time to manually transcribe, however, is about 11.4x the video's length, which is consistent with previous estimates.

Table 2. Time to transcribe the first 15 min. of the sample video.

Transcription Method	Time (min.)	Ratio to video length	Ratio to Manual
YouTube CC Transcript	2*	0.14	0.01
YouTube CC Edited Transcript	67	4.4	0.39
Manual Transcription	171**	11.4	1.0

* Does not include time to upload the video (32 min.) on ~6 Mbps link. Time to provide an automatic transcription is variable and may take several hours or a day.

** Only the first 15 min. were transcribed.

The YouTube CC transcript does not include the time needed to correct most grammatical errors (as YouTube's closed captions lack punctuation).

We also noticed that audio quality impacted error rates in the generated transcription. Fixing the generated transcript of the video's first speaker (0:00-2:13) took 27 min. The correction rate of 12.2 min./1 min. of audio is roughly on par with previous rates. This audio portion was less structured, consisted of more informal dialogue, and was articulated less clearly because the podium microphone was farther from the speaker.

Fixing the generated transcript of the second speaker (2:14-15:03) took only 46 min., or 3.6 min./min. This speaker wore a clip microphone, with clearer audio. Additionally, his speech was more practiced and steadily paced, which contributed to more correct grammar in the generated transcription.

Discussion and Conclusions

Based on this process, we can make several suggestions for how to do further verbal protocol transcriptions. (a) It can be useful to run a block of audio through YouTube's closed captioning tool to generate an automatic transcript. This may be possible using YouTube's private settings or using publicly available video. Editing the automatic transcript appears to save a lot of time if the audio is as clear as in our single example.

(b) We recommend using headphones to provide better quality sound. Headphones allow transcribers to better understand the speaker in the video.

(c) YouTube's video settings allow the user to adjust the speed of the video. By adjusting the speed manually, the audio can be slowed down to better recognize the words being spoken.

(d) Modern text editors can assist transcription efforts because they can autocorrect many typos. Coder 2 found that TextEdit was superior to Word and that the tradeoff between corrections vs. over-corrections (or lack of corrections) was worthwhile.

There may be a few limitations to this approach. Autocorrection in text editors could be a drawback in some cases if non-standard speech is being analyzed. Also, the video we analyzed might not be representative of verbal protocols. Future work should test more naturalistic verbal protocol material.

In addition, we can note a second, even more automatic method for obtaining a transcript from videos on YouTube. This method uses a Python-based command-line utility, youtube-dl, which works on Unix and Windows systems as long as a Python interpreter is present. The utility when passed the argument "--write-auto-sub" will download the video file in .mkv format and the automated captions in vtt-format. The .vtt file provides resolution to millisecond precision about when YouTube should highlight each word.

While the first method provided less timestamp information, it sufficed for our purposes. The .vtt file would have needed parsed to extract the transcript and discard all text coloring metadata, a task for which no such tool publicly exists at this time.

It appears now possible to automatically transcribe verbal protocols, at least approximately, using YouTube's closed captioning with a few errors, or with about a 4x cost to correct errors. The ability to use verbal protocols seems to have become easier, particularly where the audio is clear.

Acknowledgments

This work was funded by ONR (N00014-15-1-2275). We thank Ysabelle Coutu for ways to improve this paper.

References

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (2nd ed.). Cambridge, MA: MIT Press.

Larkin, J. H., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science, 208*, 1335-1342.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Ritter, F. E., & Larkin, J. H. (1994). Developing process models as summaries of HCI action sequences. *Human-Computer Interaction, 9*(3&4), 345-383.