Assessing the Accuracy of Automatic Speech Recognition Software on Captioning Video Objects in an Institutional Research Repository

Introduction

Florida State University (FSU) Libraries, like many other academic and research libraries, houses an institutional research repository: DigiNole. This repository hosts a plethora of research files, including dissertations, research projects, and a small percentage of video objects.

In 2020, FSU Libraries obtained a grant from the Panhandle Library Access Network to ensure the accessibility of DigiNole's audiovisual objects [1]. This grant enabled FSU Libraries to partner with the Center for Inclusive Design and Innovation at the Georgia Institute of Technology, who provided human-generated caption files (WebVTT) for videos from the research repository [1]. As the paper noted, while this approach ensured the accessibility of *those* videos, it *did not* address the need for ongoing captioning as new videos were added to the repository [1]. (This process could be repeated on a regular basis according to funding; however, there will always be periods where videos have been uploaded to the repository but have not yet been captioned.)

In order to ensure continuous audiovisual accessibility, there are two options: restrict video submissions to the research repository to only captioned videos, or caption videos right after they are uploaded to the research repository. The first option would have required the development of training materials on the captioning process and a metric for evaluating effectiveness of the training—this was out of the scope of our small research team. Instead, our team focused our efforts on the second option: how we could integrate an Automatic Speech Recognition (ASR) software into an internal captioning workflow at our library's institutional research repository.

Our research compared the Word Error Rates (WER) of four different ASR software tested on a selection of videos in our research repository: Whisper AI, Rev AI, Microsoft Stream, and AWS Transcribe. Our results showed that Whisper AI had the lowest WER score, followed somewhat closely by Rev AI. These results and methodology can help similar institutional research repositories improve the audiovisual accessibility of their content through ASR-assisted captioning.

Background and Motivation

Current research in accessibility tends to focus on visual impairments—even during 2020-2021 papers on auditory accessibility represented 17% of the research body, compared to 42% for visual accessibility [2]. In contrast, serious hearing disabilities comprise 5.7% of the population, while serious visual disabilities comprise only 4.9% [3]. Deaf and Hard-of-Hearing (DHH) people require accessible alternatives just as much Blind and Low-Vision (BLV) people, yet this area of research negatively correlates with the populations affected. Moreover, audio and video content is becoming

increasingly prevalent at institutions of higher education for communicating with their students, staff, and faculty.

There are two main ways to ensure accessibility for DHH users: providing a separate text transcript or using captions. A separate transcript is used primarily when the content is audio only (e.g. a podcast) or when video content is used by someone *other than* the video creator (e.g. a professor embedding an uncaptioned YouTube lecture into their Canvas course). For video content, captions are the preferred method, as they allow a viewer to read the captions in real time and connect the captions to the visual content of the video. Note: when "transcript" is used later in this paper, we are referring to a captions file, not to the above-mentioned transcript type.

Accurate captions are the best way to ensure a video is accessible to DHH users; however, generating those accurate captions is a time-intensive process. One potential time-saving solution is to use Automatic Speech Recognition (ASR), an Artificial Intelligence (AI) method that takes speech audio as input and returns a caption file as output. ASR may generate mostly accurate captions, but ASR-generated captions are *not sufficient* to ensure accessibility for DHH users. For instance, ASR may caption the line "Excuse me while I kiss the sky" from *Purple Haze* by Jimi Hendrix as "Excuse me while I kiss this guy." The *sounds* of "the sky" and "this guy" are almost identical, but their meaning is significantly different. Thus, human review of ASR-generated captions is necessary to ensure accuracy, though editing an ASR-generated transcript will save time compared to generating captions from scratch (for more details, please see Next Steps).

Testing

Selecting Sample Videos

Institutional research repositories consist of user-submitted content. There is no guarantee that videos submitted to the repository will have quality audio free from errors. As such, it is important that Automatic Speech Recognition (ASR) software is able to process videos with distant audio, overlapping speech, etc. We chose videos intentionally that might challenge ASR software. These are the criteria we used when selecting our sample videos [4] [5]:

- Multiple speakers
- Overlapping speech
- Background noise
- Poor audio quality
- False starts
- Strong Accents
- Use of technical terms
- Use of names and proper nouns

Selected videos represented at least one of the criteria listed, though often they represented many. All criteria were represented in at least one video.

Our research team also wanted the selected videos to be representative of the wide variety of media types that one is likely to find under the stewardship of an academic library. These categories were determined by internal experts who work closely with the research repository:

- Documentaries produced for TV broadcast and/or online streaming
- Oral history interviews
- Recorded lectures/conference presentations
- Archival film & audio recorded in the mid-20th century
- Non-professional content produced by members of university clubs

Selecting ASR Software

The ASR software we tested were limited by our resources. We had access to AWS Transcribe and Microsoft Stream because of institutional subscriptions. Rev AI's Asynchronous Text-to-Speech API was free up to the amount that we needed for our selected videos, and Whisper AI is a free, open-source tool.

ASR Software Details

Each of the software utilized was operated in a different way; however, all of them generated a WebVTT file that we reviewed. The following descriptions are from our internal report:

- Whisper AI is a free, open-source CLI application which can be run locally on a user's computer. Once installed, Whisper was run on all test items using the medium model size.
- Rev AI's Asynchronous Speech-to-Text API is a free (up to a limit) API that processes both MP3 and MP4 files using FFmpeg. A user may also submit a job to Rev AI through the command line using Curl or integrated into an application using a language like Python, JavaScript (Node), or PHP. Regardless of method chosen, a user can add custom vocabulary. To stay consistent with the other tools tested, we did not use this custom vocabulary feature.
- Microsoft Stream is a licensed video streaming application included with an Office 365 deployment. Accordingly, running the application requires no separate installation, minimal setup for a user, and incurs no separate cost. When content is uploaded to Stream (via a GUI), closed captions are automatically generated which can then be downloaded as a WebVTT file by the user.
- Transcribe is a service provided through Amazon Web Services, the cloud infrastructure that supports our web presence. Files can be batch uploaded via the AWS Management Console and return a collection of WebVTT files.

Evaluation Method

There are several methods to assess the accuracy of captions. Among them are models that evaluate the semantic difference between errors in ASR-generated captions and the accurate human-generated captions. These methods provide context that the percentage of errors alone does not; however, they are used to determine how well an *imperfect* transcript communicates the meaning of a video. Since our study measured the number of errors that would need to be fixed, the percentage of errors was the important metric.

To this end, the Word Error Rate (WER), served our purposes. The word error rate is given by the following formula:

$$WER = \frac{Substitutions + Deletions + Insertions}{Total\ Words\ in\ Transcript}$$

A Substitution error occurred when a word was replaced with another in the ASR-generated captions. A Deletion error occurred when a word was missing in the ASR-generated captions, and an Insertion error occurred when a new word was added into the ASR-generated captions.

After running ASR on our test videos, we had four WebVTT files per video. We manually reviewed each ASR-generated WebVTT file, counting the errors. This gave us the WER, which we represented as percentages. A lower percentage meant that there were fewer errors in the ASR-generated WebVTT file, which corresponds to more accurate captions.

Results

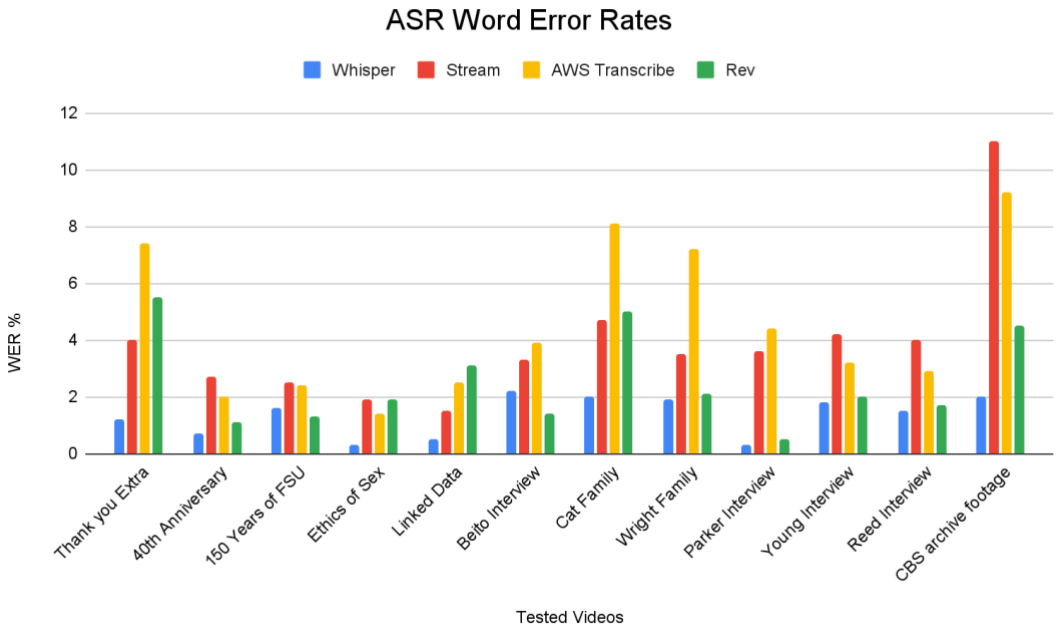

Figure 1: ASR Word Error Rates

Figure 1 shows WER percentages for each of the tested ASR software on each of the videos. All but one of the videos achieved a score under 10%, and for Whisper, all but one of the videos achieved a score under 2%. In general, Whisper tended to perform the best, followed by Rev, Stream, and Transcribe, in that order. There were variations among that order, but it is reflected by the mean error rates as shown in Figure 2 below. Additionally, the videos that gave one software more challenges tended to give more challenges to the other software, even if the specific errors were not the same.

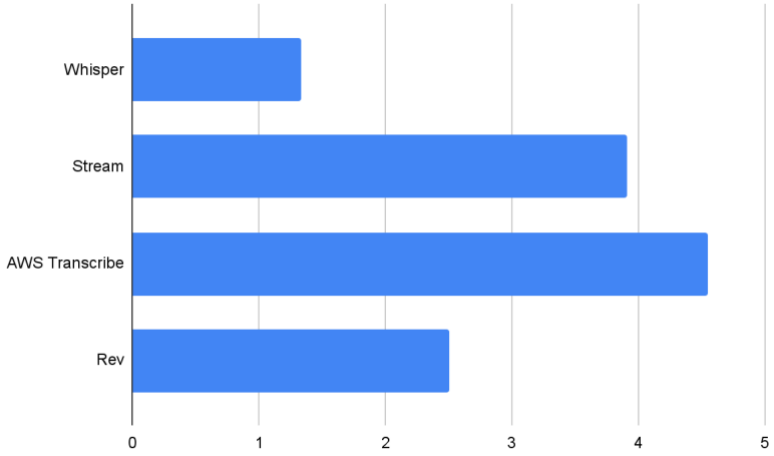

Figure 2: ASR Software Mean Error Rates

The mean values of the WER percentages were all under 5% for the videos tested; however, Whisper only had a standard deviation of 0.71%. Rev had a standard deviation of 1.64% and Stream and Transcribe both had standard deviations over 2.4%. Whisper thus had the lower WER percentage, and it *reliably* had the fewest errors.

Implications and Next Steps

Efficiency of Captions

Earlier, we mentioned that ASR-assisted captioning was inherently faster than a human captioning a video from scratch (not counting the time that the ASR software runs, as that does not require any human oversight). That is because an ASR software will generate the timestamps needed for the WebVTT file. A human transcribing from scratch will need to add those timestamps in very carefully, as an error will cause the captions to stop working. In the worst case scenario—editing a transcript with 0% accuracy—the timestamps are formatted correctly already, and since writing out the spoken words will take the same time, editing the ASR-generated transcript will be faster than writing it from scratch. (We make a minor assumption here that if the audio is not so bad that ASR could not accurately capture when someone was speaking at all.)

This is the first area for further testing: after the first time reduction from the generated timestamps, how does the accuracy of an ASR-generated transcript correlate to the time

spent editing the transcript? There may be, for instance, a significant reduction in time between a 0% accurate and 95% accurate transcript, but a negligible reduction between a 95% accurate transcript and a 98% accurate transcript. Further research would provide a quantitative answer to this question.

Expansion of Testing

Our research team tested videos within a limited scope—an institutional research repository at a large public university. The common captioning errors we used to select videos will be consistent regardless of the video source; however, the categories we used to group the videos were highly specific. We did not consider other large categories beyond the scope of an academic library, for instance: lectures, video blogs, or content produced professionally. Testing the same ASR on any of these (or other) categories would provide a complementary perspective to the results we found.

Conclusions

We have shown that Whisper ASR had the lowest Word Error Rate in our sample when tested on a representative sample of videos in an institutional research repository—less than 2% WER. Using this tool can reduce time spent on editing captions, allowing organizations to systematically ensure their audiovisual materials have accurate captions.

References

1. Dave Rodriguez, Increasing accessibility of audiovisual materials in the institutional repository at Florida State University, The Journal of Academic Librarianship, Volume 47, Issue 1, 2021 (https://doi.org/10.1016/j.acalib.2020.102291)

2. Ather Sharif, Ploypilin Pruekcharoen, Thrisha Ramesh, Ruoxi Shang, Spencer Williams, and Gary Hsieh. 2022. "What's going on in Accessibility Research?" Frequencies and Trends of Disability Categories and Research Domains in Publications at ASSETS. In Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '22). Association for Computing Machinery, New York, NY, USA, Article 46, 1–5. https://doi.org/10.1145/3517428.3550359

3. CDC, Disability Impacts All of Us (https://www.cdc.gov/ncbddd/disabilityandhealth/infographic-disability-impacts-all.html)

4. 3Play Media, Captioning Accuracy: How to Measure Error Rates (https://www.3playmedia.com/blog/captioning-accuracy-how-to-measure-error-rates/)

5. Digital Nirvana, Why caption accuracy is so important and what is WER? (https://digital-nirvana.com/blogs/why-caption-accuracy-is-so-important-and-what-is-wer/)