Evaluating Automatic Captioning Systems for Improving the Accessibility of Library Video Content

PRESENTER:
Matt Vaughn, Indiana University

Introduction

Scholarly content like conference poster presentations is being recorded now more than ever. As librarians endeavor to preserve and disseminate this content in an accessible way, one of the most important things we can do is add captions and transcripts. Automatic speech recognition (ASR) platforms can facilitate the creation of accurate transcripts, but it is important to understand the strengths and limitations of these tools. A recent study by Rodriguez and Brown (2023) demonstrated that in comparison to Amazon Transcribe and Microsoft Stream, the Whisper AI ASR tool had by far the lowest Word Error Rate (WER) in transcribing audio. The accuracy of Whisper is such that it could save librarians considerable time in creating transcripts for recordings. This study attempts to replicate and extend their work by comparing Whisper's audio transcription capacity to that of YouTube, Kaltura, and Zoom.

Methodology

Three video recordings of poster presentations were selected from Indiana University's audiovisual repository, Media Collections Online. In order to test the effectiveness of these tools in transcribing different types of audio, I selected one poster presentation with clear, high-quality audio, one with low quality audio, and one with variable quality audio and multiple presenters. Each presentation had similar word counts and was about three minutes long. First, I created an accurate, manual transcript of the three videos. I then used each of the four ASR tools to automatically generate transcripts for the videos. Finally, I used the Amberscript WER tool and the GoTranscript text comparison tool to compare the automatic transcripts to my manual transcripts and, thereby, determine the WER and the number of corrections needed for each transcript (see Figures 1 and 3).

Whisper AI **outperforms** YouTube, Kaltura, and Zoom **in generating accurate transcriptions** of poster presentation video recordings.

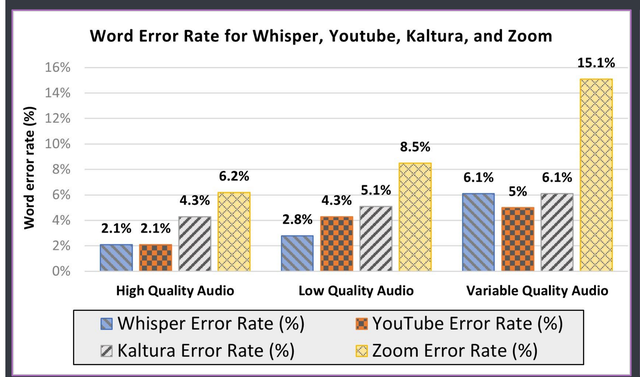

Figure 1: Whisper and YouTube had the lowest WER of the four ASR tools tested. Unlike the other tools, Zoom must transcribe videos in real time – they are not uploaded. This likely accounts for Zoom's high WER.

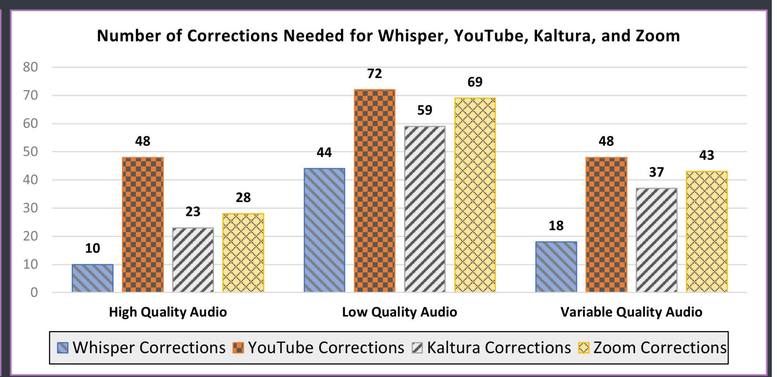

Figure 3: WER accounts for word accuracy alone and ignores elements like punctuation and capitalization. When these additional elements are tracked, Whisper's transcripts still require fewer corrections – while Zoom's performance improves over YouTube's.

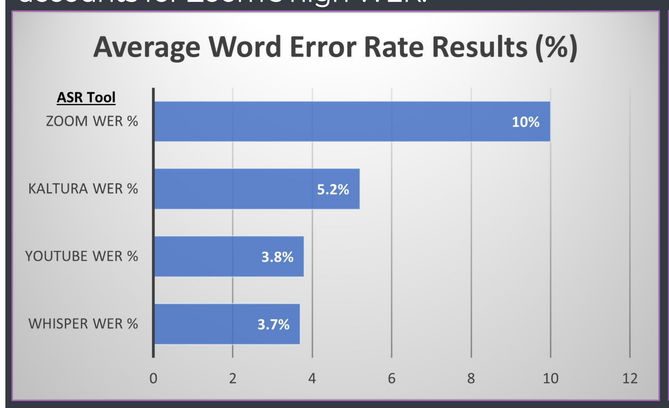

Figure 2: WER accounts for incorrect words, missing words, and inserted words that are not actually spoken in the video. For the automatic transcriptions of the 3 videos, Whisper had the lowest average WER % – just under YouTube's %.

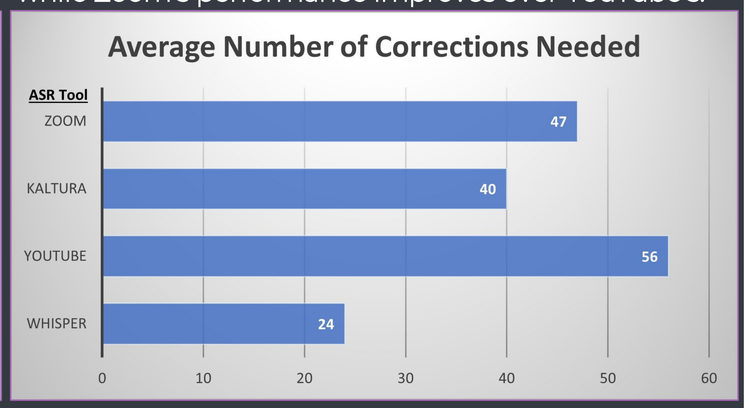

Figure 4: Whisper transcripts required the lowest average number of corrections needed. In addition, Zoom outperformed YouTube when punctuation and capitalization were considered. While YouTube excels at word accuracy, it struggles with editorial style.

Results

- Whisper and YouTube had the lowest average Word Error Rates of the four ASR tools tested (Figures 1 and 2).
- In contrast to the other tools, Zoom must transcribe recorded audio as is it is played on a local computer in real time – videos are not uploaded directly for transcription. This likely accounts for Zoom's high WER %.
- When punctuation and capitalization are tracked in addition to word errors, Whisper's transcripts still required the fewest corrections. (Figures 3 and 4).
- Zoom outperformed YouTube when punctuation and capitalization were considered (Figures 3 and 4).
- While YouTube excels at transcribing words accurately, it struggles with punctuation and capitalization.

Conclusion

This study supports previous work by Rodriguez and Brown demonstrating Whisper's overall effectiveness. While Whisper achieved the best results, each of these tools can be useful in creating captions and transcripts for your library video and audio content. The other three tools, for instance, automatically generate downloadable caption files for your content in one or both of the most commonly used subtitle formats, VTT and SRT. In addition to the relative effectiveness of these platforms, your choice of tool will also likely depend on your familiarity with it as well as its ease of use and availability to you. As scholarly video content proliferates, we need better tools and efficient workflows to make this content more accessible for everyone. ASR Tools like these can help us to achieve our accessibility goals.

References

- Rodriguez, D., & Brown, B. J. (2023). Comparative analysis of automated speech recognition technologies for enhanced audiovisual accessibility. *Code4Lib Journal*.
- Amberscript WER tool: https://www.amberscript.com/en/resources/wer-tool/
- GoTranscript text comparison tool: https://gotranscript.com/text-compare

Inclusion, Diversity, Equity, and Accessibility in Libraries & Archives
TORONTO 24
Sustainable Resistance and Restoration in Global Communities
INDIANA UNIVERSITY