Caption This: Creating Efficiency in Audiovisual Accessibility Using Artificial Intelligence

2020 Catalyst Fund Final Report

Submitted by

Abigail Norris, Digital Initiatives Librarian and Assistant Professor, University of Mississippi Libraries

Michelle Emanuel, Ph.D., Head of Metadata and Digital Initiatives and Professor, University of Mississippi Libraries

This project was made possible in part by a 2020 award from the Catalyst Fund at LYRASIS

Project Goals

Like many public institutions, the University of Mississippi is committed to making all of its programs, services, and activities accessible to all students, staff, faculty, and community users. This applies especially to public-facing electronic resources, which include culturally significant digital collections in our institutional repository, eGrove. For audio and video (A/V) collections, the captioning process is time consuming and labor-intensive. It requires listening to the recording in real time, replaying the recording at different speeds to decipher difficult passages, and writing down every word, pause, and non-verbal communication with a time-stamp to indicate where in the recording the text occurred. Existing models of auto-generating caption files, such as uploading to YouTube, are known to be mediocre and do not remove the need for proofreading. Furthermore, few libraries want to keep their digital A/V collections in YouTube, because of rights and ownership issues, pop-up ads, and lack of sustainability in using personal YouTube accounts to manage A/V files in a professional setting. As more libraries recognize the need to caption both audio and video content as part of their accessibility initiatives, more work needs to be done to make this important issue easier to resolve.

The grant “Caption This: Creating Efficiency in Audiovisual Accessibility Using Artificial Intelligence” was proposed to research ways cultural heritage institutions could increase both efficiency and accuracy in captioning A/V content, with the intention of creating a toolkit that integrates artificial intelligence with library needs to hone this process. While there are resources on captioning digital A/V files using similar methods, few address the unique needs of cultural heritage content. These include factors like low or degraded audio quality, historic or niche vocabularies, and the persistent problems of underfunding and undervaluing cultural heritage initiatives that makes it hard to devote staff time to such labor-intensive work. In addition, cultural heritage institutions must consider factors such as institutional branding and donor agreements when hosting content online.

None of the identified resources offer a specific workflow for libraries to maximize efficiency in transcription, especially with limited funding and employee time. Our project tested the viability and sustainability of using pre-existing Python scripts shared on websites like GitHub to generate SubRip title (.srt) files that, when combined with corresponding A/V files in our repository, created closed captions. The Primary Investigator began by identifying scripts that were open source, operable by individuals with limited knowledge of Python, and able to handle 30+ minutes of audio. After initial work began, we also decided to include proprietary – but low-cost – solutions.

The intended audience of this toolkit is anyone working in a cultural heritage institution looking for a low-cost solution to transcribe their A/V content. Example users include an archivist at a university looking for a project for student workers, a lone librarian struggling to keep up with transcription requests, or a librarian wanting to

upload a new collection to their institutional repository but also needing to comply with university accessibility mandates.

At the same time, the toolkit aims to guide users with limited technological and/or digital skills through the process and help them gain confidence in their abilities. The toolkit guides users through using the terminal (command line) and a variety of different software platforms. in an effort to make it accessible to all librarians and archivists, regardless of digital training. The toolkit will also be helpful for individuals wanting to learn more about accessibility, its importance to libraries and archives, and the best way to make video content available to users relying on closed captioning.

We created and tested the toolkit by captioning A/V content from the University of Mississippi Libraries' digital collections. We identified 55 videos from four collections that we aimed to transcribe over the course of the grant. As we were aware that factors like audio quality and a speaker's gender or accent could affect how well the AI transcribed them, these videos featured a variety of speakers from different gender, racial, ethnic, and regional backgrounds. All of the videos were already available on the library's institutional repository, eGrove, and had not been previously captioned.

As awareness grows of the standards required to make libraries equitable and accessible to all, more institutions will need to dedicate significant amounts of time to increasing accessibility in their digital A/V collections.

Desired Outcomes:

1. Identifying an open source automatic speech recognition (ASR) script that can be modified to better identify speakers in low-quality A/V files.
2. Fine-tuning the script so that it can identify speakers from a variety of regional, racial, gender, and ethnic backgrounds, as well as audio of varying age, recording method, and quality.
3. Decreasing the time it currently takes it transcribe audio files.
4. Creating a workflow that can be applied and adapted to a variety of A/V materials at different GLAM (Gallery, Library, Archive, and Museum) institutions.
5. Captioning 55 pre-selected videos from our digital A/V collection and adding them to the library's institutional repository.
6. Making our suggested workflow available to other LYRASIS members facing the same accessibility issues with their own digital collections.

Our Process

Key Terms:

Transcript: Transcripts are used for audio-only content. A transcript is a plain text document that transcribes audio and may or may not include sound effects. Transcripts do not have time information attached to them and cannot be used as captions.

Caption: Captions are used for video content, where the sound accompanies image. Captions are time-stamped and appear on screen as the words are being spoken. They reproduce every audio element, including non-verbal audio/cues like background noise, laughs, and silence.

Automatic Speech Recognition (ASR): ASR is the technology that recognizes spoken words and translates it to text.

What We Did

Over the course of the grant, two full-time librarians (including the Primary Investigator), one Graduate Assistant (Fall '20) and four library student workers (Spring '21) worked on various aspects of the project. We began with an environmental scan of ASR tools, asking ourselves the following questions:

1. What tools exist for automatic captioning or transcription? Are they open source or proprietary?
2. How difficult is this tool to use? Could a librarian or archivist with minimal technological training use it?
3. What are the problems with automatic captioning? How does the speaker's race, gender, age, or accent affect how well they're captioned?
4. Why do the issues identified by the previous question arise, and is there a way to mitigate them?

From these questions, we compiled both a list of tools to research and a set of issues to be mindful of when analyzing and editing captions. In all, three tools were tested: Google's Speech-to-Text API, Kaldi ASR, and Rev.ai. When testing tools, the PI and GA would install any necessary software, run a specified control video, and assess the output for accuracy and amount of editing needed.

Google Speech-to-Text API

Google's automatic transcription software was initially the most promising solution and therefore the one on which the most time was spent. We identified several scripts that would use the API to transcribe A/V content and began testing them. Without any modifications, the results were disappointing. For many of the videos, the accuracy of the transcribed text was low. The API selected modern vocabulary supplements instead of decade-appropriate language (example: "app" instead of "apt"). On its own, Google Speech-to-Text will only recognize 30 seconds of audio before timing out. This was problematic for oral history interviews that lasted 30+ minutes. Finally, in the "Matthew Joseph Interviews" collection, we noted that the ASR tool accurately transcribed a much higher percentage of the white, male interviewer who spoke in a standard American accent than his interviewees, who were often Black and spoke with strong southern accents.

The PI and GA identified various modifications that could help address these issues, including an extension that would loop together chunks of 30-second audio to

create a longer output and a modification that identified silences. While the latter modification made word recognition slightly more accurate, it had to be altered for every video based on speaker speed, only adding to editing time. Overall, we were unsatisfied with the quality of the output. Through research, we determined that Google Speech-to-Text was more suited to high-quality audio or direct input from a single speaker. Despite multiple modifications, the output generally took longer to edit than it would take to manually transcribe the same video, negating the purpose of the ASR tool. Because of this, we determined that Google Speech-to-Text was not fit for our purposes.

Kaldi

Kaldi ASR was selected for testing because it is widely regarded as a high-quality, open source automatic speech recognition tool among digital scholarship practitioners. However, we quickly decided it was not a good fit for this project after installation. One of the major goals of this project was that the solution be accessible to those without much technological training. Because Kaldi required proficiency in the coding language C++, we determined that the barrier to entry was too high to continue testing the tool.

Rev.ai

When neither of these options produced the desired results, we began to consider proprietary programs. An earlier literature review suggested that Rev.ai was a low-cost solution that produced accurate results. Rev.ai is a commercial transcription service that charges $0.035/minute to transcribe audio files. While we were initially hesitant to switch to a proprietary solution, Rev offered multiple solutions that could make the time savings worth it:

1) The program promises a 95% accuracy rate. While we would still have to conduct formal tests, our experience with the program showed that it was significantly more accurate than other tested solutions.
2) Using the software was simple and could be done either through the Command Line or a website.
3) Transcriptions could be exported in multiple file formats, including SRT. This provided the added benefit of not having to time stamp or format the output, significantly increasing the time savings.

In all, the improvements Rev offered were significant enough that we decided the time savings outweighed the cost of using the program.

The late Fall 2020 and Spring 2021 semesters were spent testing and refining the caption workflow using Rev outputs, editing captions, and writing the proposed toolkit.

What Did and Did Not Work

Overall, we consider the project a success despite deviating from our original plans. The biggest challenge was not that various methods did not work, but rather that they did not work *well enough*. As discussed in more detail throughout this white paper, we successfully identified and modified a Google Speech-to-Text script. However, this ASR was not able to accurately transcribe our A/V content. While Kaldi is a good solution for institutions with personnel who can manage it, it was too technologically involved for our project goals. Automatic speech recognition improves every day, and we are eager to reassess these and other tools as time passes.

Our largest "failure" was not producing an open-source solution. We discuss this further in the "Lessons Learned" section, but want to note that we do not consider this a "failure" despite the significant change from our original plan. Everything comes at a cost, whether that is a proprietary tool or the amount of time it takes an employee to edit a Google transcript. Through the course of this project, we deemed that the time saved using Rev.ai made it more than worth the low cost of using the technology.

One perplexing problem we discovered during the course of the project is that the majority of videos from the Open Doors Collection were not recognized by any of the ASR tools. When these files were input into Google, Rev.ai, and others, the transcriptions would come back either blank or marked "error." Initial analysis did not reveal a reason for this. The file types (MOV and MP4) and digitization methods for these videos were the same as others used in the project; when the videos were viewed on eGrove or Windows Media Player, the audio was clear and did not appear to have any errors; and no encryptions, bugs, or glitches were found in the files themselves. As we continue to work on the project, we plan to consult the archivist who works with these collections to solve the issue. Because of this, only two of the ten proposed videos from this collection were captioned.

Modifications from the Original Proposal

As with every aspect of life, significant changes were made to the proposal due to the COVID-19 pandemic. Both the primary investigators and any students hired worked from home for the majority of the grant. Because of this, student workflows had to be adapted so that students could complete the work on their own devices. The primary investigators were conscious that no student lose the opportunity to work on the grant because they did not have the appropriate technology. Any student who did not have the technology to execute code on their own personal devices was able to check out a library laptop to complete their work. Ultimately, one student checked out a laptop.

Again because of the pandemic, modifications were made to the use of funds. We originally proposed purchasing new student computers and transcription aids for use in the library. When the pandemic forced us to work from home, these purchases were no longer practical. Once we decided to use Rev.ai, which costs $0.035/minute to

transcribe, we reallocated the funds originally intended to purchase equipment to pay for the transcription service.

Due to an unforeseen, last-minute change, we were unable to hire a graduate student for the Spring 2021 semester. This limited our ability to analyze statistics, as this was supposed to be one of the GA's primary duties in the spring. As we continue to work on the project, we hope to hire another GA who will audit Rev for accuracy and conduct tests on exactly how much time is saved using the platform. Currently, we do not have exact numbers for either of those.

Finally, the most significant modification was the decision to recommend a proprietary program for the automated transcription work. This is discussed in detail in the "Lessons Learned" section, but it is worth noting that we did not conduct a survey because of this change. The survey was initially intended to gauge the ease, accessibility, and usefulness of an open source solution. Because Rev is relatively straightforward to use, we did not think that conducting a survey at this point would be the most useful course of action. In the future, we may decide to conduct a survey to learn more about transcription efforts at cultural heritage institutions and whether a service like Rev.ai is preferable to other solutions.

Accomplishments

Prior to this project, video transcription in the University of Mississippi Libraries Digital Initiatives Department was cumbersome and unsustainable for our staff of three. Captions made using Google Speech-to-Text were inaccurate and took at least ten times the length of the video to edit. Captions pulled from YouTube's automatic captions were generally slightly more accurate, but still required significant edits and time-stamping. Using Rev.ai has created a more streamlined process that allows us to more efficiently provide accessible A/V content to patrons.

In all, we produced captions for 29 recordings that were uploaded to the institutional repository. Although this falls short of the 55 originally proposed, we consider this a success since the majority of these transcripts were created in the spring semester/on a shortened timeline. The 29 captions are distributed across the video collections as such:

Collection	Videos Captioned	Videos Remaining	Total:
Field School	8	7	15
Freedom Riders	10	10	20
Matthew Joseph Interviews	11	0	11
Open Doors	2	8	10
Total:	31	25	56

Please note: one video was added during the transcription process, making the final count 56.

In addition, we created a toolkit that describes proposed best practices and workflow suggestions. This document can be used by future library student workers as a training guide for creating captions. The toolkit is designed to introduce the user to key issues and ideas within the world of digital A/V accessibility before giving step-by-step instructions on how to caption and transcribe A/V content. The toolkit 1) provides background information on accessibility and its relation to library and archive digital content; 2) gives an overview of AI and ASR application within libraries; 3) contains an annotated resource list of everything needed to complete a transcription project; 4) provides a suggested workflow to transcribe audio and video content; 5) discusses open source versus third-party tools on the market and why the grant ultimately failed in its goal to provide a completely open source solution.

The toolkit gives recommendations based on what we perceive to be best practice, but users should keep in mind that their own unique needs and collections may require altered workflows. We predict that, as the technology advances and becomes more widely available, this document will need to be updated regularly.

Lessons Learned

At the beginning of this project, our intention was to present an open source, or "free," solution. However, during the tool selection process, a number of questions and issues arose that led us to recommend a paid solution, Rev.ai. During the selection process, the team considered three primary transcription options: cloud-based, open source code like Google Speech-to-Text; Rev.ai; and Kaldi. We quickly decided against Kaldi because the program was overly complicated to be installed and operated by people without coding skills and our project is intended for professionals of all technological skills levels.

Google Speech-to-Text was the solution on which we spent the most time. It is an open source solution with lots of documentation with positive feedback online. However, there were a number of problems that led to it not being the best solution. While we were able to compile a script that transcribed long audio files, the transcription output frequently contained many inaccuracies. In the majority of cases, editing these inaccuracies took longer than manually transcribing a video. Several scripts were found that improved the transcription quality, but the changes that had to be made for each video were specific to its audio quality and did not lend themselves to large-scale projects.

Simply put, the amount of time it took to edit the Google Speech-to-Text output cost significantly more in employee time than paying for a higher-quality transcription from Rev.ai. Thus, while our original intention to provide a completely open source solution was not realized, we believe that the time saved is well worth the small amount of money spent.

Next Steps

Because we were unable to hire a graduate student in the Spring ’21 semester, we had a large number of unused funds at the end of the grant. LYRASIS graciously offered to let us keep the funds to continue developing the project. In the Fall ’21 semester, we will focus on completing any transcripts unfinished by students and adding them to the institutional repository. We will also assess the Open Doors files and see if they can be transcribed or identified by various softwares. In Spring ’22, we plan to hire a Graduate Assistant with the remaining grant funds. The GA will be tasked with assessing any new or updated AI transcription tools, working with the PI to calculate statistically significant information (accuracy of Rev vs quality of file/speaker type, how long it takes to edit a Rev file vs a Google Speech-to-Text or manual transcription), completing transcription of the remaining proposed files (assisted by an additional student if funds allow), and updating the toolkit with any findings.

We recognized early on that a large barrier to improving ASR tools like Google Speech-to-Text was that these tools draw on libraries of vocabulary that we could not edit. The best way to improve ASR tools is to continue expanding and perfecting these libraries, a task which is far outside the scope of this project. Nevertheless, we believe it would be a valuable project to look into. LYRASIS can play a role in developing the project further by building a library that improves AI transcription. This library could be a community-focused effort that invites contributions from LYRASIS members, exposing the program to a variety of gender, racial, ethnic, and regional voices.

Find Out More

To download the “Caption This” Toolkit, and any future publications related to the project, please visit the grant’s site on eGrove, the University of Mississippi Libraries’ institutional repository: https://egrove.olemiss.edu/libpubs/17/

To view the videos transcribed through this project, please visit the collection pages on eGrove:

Open Doors Oral History Collection: https://egrove.olemiss.edu/opendoors/

Field School for Cultural Documentation: North Mississippi Music Project: https://egrove.olemiss.edu/fieldschool/

Matthew Joseph Interviews: https://egrove.olemiss.edu/mjinterview/

Freedom Riders Oral Histories: https://egrove.olemiss.edu/freeriders/