9-2021

Caption This: Creating Efficiency in Audiovisual Accessibility Using Automatic Speech Recognition Toolkit

Abigail Norris-Davidson
University of Mississippi, abigailn@olemiss.edu

Michelle Emanuel
University of Mississippi, memanuel@olemiss.edu

1848

THE UNIVERSITY of MISSISSIPPI

UNIVERSITY LIBRARIES

Caption This: Creating Efficiency in Audiovisual Accessibility Using Automatic Speech Recognition Toolkit

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

Table of Contents

Caption This: Creating Efficiency in Audiovisual Accessibility Using Automatic Speech Recognition

Introduction

In the last several years, there has been a growing awareness of the need for digital accessibility in cultural heritage institutions. While initiatives to make content accessible and equitable for all patrons are vital for the continued growth and effectiveness of these institutions, they are not changes that can be made overnight. Remediating content requires time, knowledge, and effective tools. For many solo or siloed cultural heritage institutions, it can be difficult to commit the resources necessary for remediation. Nevertheless, these institutions will need to dedicate significant amounts of time to increasing accessibility in their digital collections, including audiovisual (A/V) content.

For A/V collections, the process of making material accessible to all users 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. This toolkit is intended to create an easily replicable, low-cost, efficient solution for transcribing and captioning library and archival video content, making A/V remediation feasible for institutions that lack the resources to undertake an in-depth transcription project.

Toolkit Objective

The objective of this toolkit is to create a feasible process by which cultural heritage institutions can transcribe their digital A/V content without the significant time, financial, and personnel requirements of traditional transcription projects. The toolkit aims to do this by implementing automated speech recognition (ASR) technology and open source editing tools. It integrates industry standards with workflows developed by the investigators. Users are able to follow step-by-step transcription instructions while still having the freedom to adjust workflows to fit the needs of their own collections. After completing the toolkit, users will have an increased knowledge of A/V accessibility needs, transcription workflows, and the potential of ASR and artificial intelligence (AI) to meet library and archive needs.

This toolkit does not promise to make the transcription process speedy or completely automated. Transcription still requires human editing for grammar and incorrectly transcribed words, but we believe our method significantly cuts down on the time required to transcribe A/V content. As technology advances, we anticipate the process will continue to improve. While we initially intended to release a toolkit of fully-open source programs, we ultimately decided to recommend a proprietary but low-cost transcription service, Rev.ai. The grant process revealed that the small sum incurred by using Rev.ai was ultimately worth the cost in terms of employee time spent editing and formatting output. We believe that this is the most attainable solution. For an in-depth look into open-source options considered and why we ultimately went with a paid third-party solution, please see the section entitled "Open-Source vs. Paid Solutions."

Key Terms and Concepts

Transcript: "…the process in which speech or audio is converted into a written, plain text document. Transcripts are the output of transcription, and because they are plain text there is no time information attached to it.

There are two main transcription practices: verbatim and clean read. Verbatim transcribes the audio word-for-word and includes all utterances and sound effects, great for scripted speech like a TV show, movie, or skit. Clean read transcription edits the text to read more fluidly, perfect for unscripted content like interviews and recorded speaking events."

Transcripts are used for audio-only content. This workflow recommends creating clean read transcripts, which omit unnecessary words (such as "um") while retaining the meaning of the sentence. (source)

Caption: "…a process that involves dividing transcript text into chucks, known as "caption frames," and time-coding each frame to synchronize with the audio of a video. The output of captioning are captions which are typically located at the bottom of a video screen. Captions allow viewers to follow along with the audio and video or captions interchangeably.

Closed captions should depict speech and sound effects, as well as identify various speakers. Captions must account for any sound that is not visually apparent, and assume that the viewer cannot hear the video at all. Afterall, d/Deaf and hard of hearing people often rely on captions to consume video media."

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. (source)

Web Content Accessibility Guidelines (WCAG): The guidelines have "a goal of providing a single shared standard for web content accessibility that meets the needs of individuals, organizations, and governments internationally … The WCAG documents explain how to make web content more accessible to people with disabilities." (source)

Americans with Disabilities Act (ADA): The ADA is a civil rights law that prohibits discrimination against individuals with disabilities in all areas of public life, including jobs, schools, transportation, and all public and private places that are open to the general public. (source)

Accessibility: Per the ADA, accessibility “means a person with a disability is afforded the opportunity to acquire the same information, engage in the same interactions, and enjoy the same services as a person without a disability in an equally effective and equally integrated manner, with substantially equivalent ease of use.” (source)

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

Using the Toolkit

Toolkit Design

The toolkit is designed to introduce users to key concepts 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 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.

Intended Audience

The intended audience of this toolkit is anyone who works in a cultural heritage institution and is 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 who struggles to keep up with transcription requests, or a librarian who wants to upload a new collection to their institutional repository and needs to comply with university accessibility mandates.

The toolkit aims to guide users with limited technological/digital skills through the process and help them gain confidence in their abilities. The toolkit guides users through using the command line and suggested software. The goal of the toolkit was to make it accessible to all librarians and archivists, regardless of digital training. The toolkit will also be helpful for individuals who want to learn more about accessibility, its importance to libraries and archives, and the best way to make video content available to users who rely on closed captioning.

Assumptions

The toolkit was written so that even those with limited technological training or knowledge of accessibility standards can follow it. Users who have more experience may find that they do not need to follow the instructions exactly as written due to personal preference. Ultimately, these instructions are meant to serve as suggestions.

The toolkit was also written for PC; small adjustments in programs and commands will have to be made to follow the toolkit on a Mac or Linux system.

Open-Source vs. Paid Solutions

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 project is intended for professionals of all technological skills levels, and we determined the program was overly complicated to be installed and operated by people without coding skills.

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 each video's audio quality and did not lend themselves to large-scale projects.

Rev.ai is a commercial transcription service that charges $0.035/minute to transcribe audio files. The transcriptions were significantly more accurate than Google Speech-to-Text.

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.

Suggested Caption and Transcription Workflow

Device Setup

Below is a list of recommended tools and websites needed to follow the toolkit. Notes on account and installation requirements are included. These tools were selected based on accessibility, cost, and the authors' experience. Some users may wish to replace certain programs with ones they are more comfortable with or have access to, such as substituting Panopto for CADET.

Tool/Software	Web or Desktop Application	Account/Installation Needed	Use
Command-line	Desktop	None needed	
Rev.ai	Web – command-line option	Account required; pay per use	

VLC Media Player (PC) or QuickTime Media (Mac)	Desktop	Installation may be required	
Python (language) and IDLE	Desktop	Installation required; installation instructions here	
Plain text editor of your choice – Notepad or TextEdit	Desktop	Typically, none needed	
Microsoft Word or a similar word processing software	Desktop	Typically, none needed	
CADET – Caption and Descriptive Editing Tool	Desktop – runs in browser	Installation required; installation instructions here	
Extract text only from subtitle and remove timestamps	Web	None needed	

Make sure that all of these programs are installed or accounts created before beginning the captioning process.

Save all of the A/V files you are transcribing to the same folder, located in an easily accessible area like the Desktop.

Importing to Rev.ai

This workflow covers how to upload A/V files to Rev.ai using an API. This upload method is recommended for large quantities of files and/or a more hands-off approach. It requires the use of IDLE, the coding language Python's integrated development environment, or another IDE of your choice. All required Python knowledge is explained in this workflow. Files may also be uploaded directly to Rev.ai; this method is more hands-on and is better suited to a small number of files.

1. In your Rev account, under "Access Token," generate an access token. Save this token in a secure location.
2. Open IDLE and start a new file. Copy/paste this text into the file:

```
from rev_ai import apiclient

client = apiclient.RevAiAPIClient("your access token")
```

```
job = client.submit_job_local_file("yourfilename.yourfileextension")
```

Underlined text will change based on your token and file name.

NOTE: If uploading multiple files, repeat job = client.submit_job_local_file("yourfilename.yourfileextension")with corresponding file names for as many files as you have.

3. Save this file into the same folder as the videos you are uploading.
4. Run the script by selecting Run>Run Module. You may be prompted to save the file again; do so.
5. A new window will open. Initially, you may not see any activity. This means the process is still running. You may have to wait a few minutes. The process is complete when >>> appears on the new window.
6. In your internet browser, go to rev.ai. Your transcripts will be listed under “Recent Jobs.”
7. Download your files. For this workflow, download the transcript as an SRT.
    a. NOTE: When transcripts are uploaded to Rev, their original file names are replaced by the Job ID. As you download files, make sure to rename them with their original file names. This will save time in the future.

Editing Captions

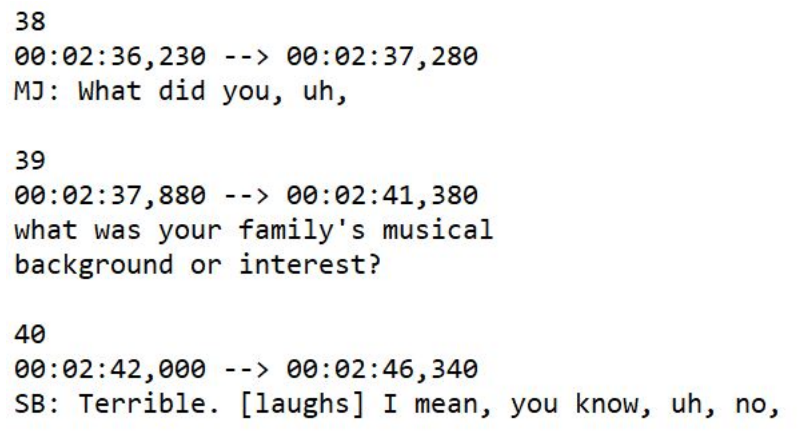
Step One: Text Editor

1. Open the SRT file in Notepad/TextEdit and the corresponding video in Windows Media Player or equivalent.
2. Play the video once through and read the captions without making any changes. This will give you an idea of the video content and how much editing needs to be done. If the video is long, you can play it at double speed or only watch a portion.
3. Start the video over again and begin editing the SRT file. As the video plays, read the text and edit any grammatical errors, incorrect words, and missing phrases. Note the speakers, separating statements made by different people into different timestamps. Only edit the text from the video; do not edit any SRT formatting.
    a. Follow the DCMP Captioning Tip Sheet and Captioning Key to determine format, non-verbal communication, speaker identification, and other important captioning elements.
    b. Maintain the SRT layout, even if it doesn’t feel intuitive to edit. Manually timestamping videos takes a long time.
    c. If the timing of the captions gets off or you have to add a significant amount of text, don’t worry about editing the timestamps. That will happen in future steps.

d. In Windows Media Player, you can slow your play speed in order to keep up with the speakers as you type. Adjust play speed as necessary to understand the speakers.

e. If there are words or phrases you're unsure about, **frame them with double asterisks.** If you can make an *informed* guess about the phrase, include it inside the asterisks. The asterisks will let the reviewer know that the section needs to be edited.

4. Play sections of the video as many times as you need to properly edit them. This is an edited SRT file:

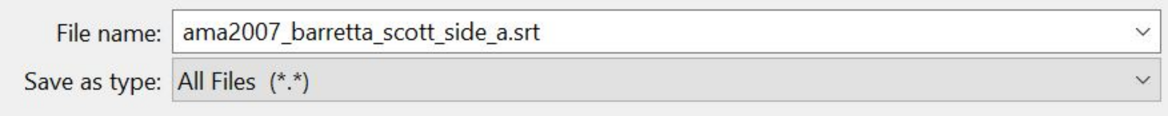


```
38
00:02:36,230 --> 00:02:37,280
MJ: What did you, uh,

39
00:02:37,880 --> 00:02:41,380
what was your family's musical
background or interest?

40
00:02:42,000 --> 00:02:46,340
SB: Terrible. [laughs] I mean, you know, uh, no,
```

5. When you are satisfied with your edits, save the file.
   a. **MAKE SURE** that you save the file as an SRT, not a TXT. To do this, under "File name:" add ".srt" after your file name; under "Save as type:" select "All files (*.*)"



File name: ama2007_barretta_scott_side_a.srt

Save as type: All Files (*.*)

Step Two: CADET

Note: Depending on the type of institution you work for, you may have access to subscription-based captioning tools such as Panopto, Adobe Premiere, or Camtasia. Adapt these guidelines to whichever tool is most convenient for you.

1. Open CADET.
2. Go to File>open Media… and open the video you're captioning.
3. Next, go to File>Import… . Under "Import type," select SRT. Navigate to the corresponding SRT file and open it.
4. Play the video, making sure the SRT file and video are aligned. Adjust timestamps as necessary.
5. If you had to add any sections to the caption, input them now. Follow the CADET Captioning Documentation to make any necessary edits.

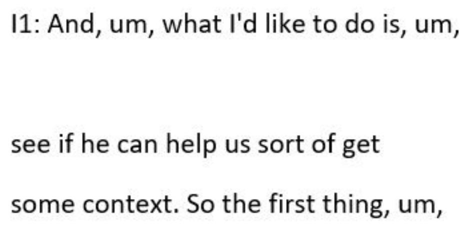
6. Each time you exit CADET, export the SRT file by going to “File>Export…” and selecting SRT as the “Export type.”
    a. We have had issues with CADET not saving changes within the platform. This extra step is to ensure no work is lost.
    b. If you continue editing the same caption the next time you work, verify that all your changes were saved. If they were not, re-import the previously exported SRT.
7. When you’ve finished your edits in CADET, save the SRT file to your project folder.
    a. Once a caption’s final version is uploaded, you may delete the video from your computer.

Reviewing Captions

1. Download the SRT caption and open the video file. In the Caption Log, change the status to “In review” and add your name as the reviewer.
2. CTRL/CMD + F to search for **double asterisks** within the document. Note that these sections will need extra review.
3. Watch the video and follow along with the SRT. Edit any mistakes you find. If there are still sections where you’re not sure what is being said, **keep or add double asterisks.** This will mark sections for review in the next step.
4. Only edit the timestamps in CADET if the timing is significantly off.
5. When you’re satisfied with your edits, save the file.
    a. **MAKE SURE** that you save the file as an SRT, not a TXT. To do this, under “File name:” add “.srt” after your file name; under “Save as type:” select “All files (*.*)”
6. The caption is complete.

Exporting an SRT to a Transcript

1. Open the Extract text only tool.
2. Copy the text from your SRT file into the box on the left side of the page.
3. Press “Do the job!” The transcript will appear in the box on the right.
4. Copy/paste the transcript into Word. The text will appear exactly as it does in the caption file, with short lines and many line breaks:

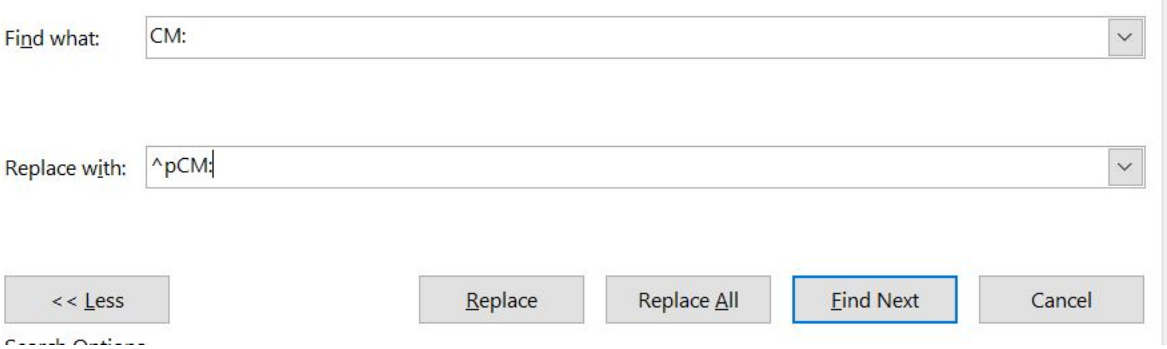


I1: And, um, what I'd like to do is, um,

see if he can help us sort of get

some context. So the first thing, um,

5. In the Toolbar, go to Find and Replace>Replace. Under “Find what,” type “^p”; under “Replace with,” type a single space “ “. Click “Replace all” - your document should now be one paragraph.
    a. If there are any accidental double spaces, you can batch edit them out using the Replace tool. Under “Find what,” type a double space “ “; under “Replace with,” type a single space “ “.
7. Next, add a line break each time someone new begins speaking. To do this, open the Replace tool again. Under “Find what,” type the initials that indicate your first speaker, followed by a colon (ex: CM:). In “Replace with,” type “^p” followed by the initials and colon (ex: ^pCM:). This will make a line break every time that individual begins speaking.



Find what: CM:

Replace with: ^pCM:

7. Repeat this step for all speakers.
8. Edit the document to follow the Transcript Template Guidelines found below.
9. Save the document as a .doc and name it according to your institution’s file naming standards.
10. The transcript is complete.

Transcript Template Guidelines

This template is designed to give your organization’s A/V transcripts consistency and appropriate metadata. Depending on the structure of your collections or organization, you may need to adapt some information. The template was created for oral histories; if you are transcribing a non-interview format item, adapt the metadata to suit. For example, instead of naming the interviewee/interviewer, you would name the speakers. Transcripts should be properly formatted and in a common, readable font such as Arial, Calibri, Times New Roman, or Georgia.

Item Title
Collection Title
Institutional Title(s)

Item status: [ex: Open to the public]

Name of interviewee: Interviewee name [IN]
Name of interviewer(s): Interviewer name [IN] (Note: If two or more speakers have the same initials, distinguish them by clarifying “IN1,” IN2,” etc.

Length of interview: hh:mm:ss
Date of interview: dd-mm-yyyy
Language of interview: [ex: English]

Name of transcriber:
Date of transcription: dd-mm-yyyy
Editorial note: [quality of recordings, background noise, gaps in recording, etc.]

Resources

Suggested Websites

Described and Captioned Media Program (DCMP)

The DCMP is “the leader for captioning and description standards. We provide not only accessible content but the standard for professionals and amateurs working to build quality, accessible media” (source). After a thorough review of captioning best practices and guides, we determined that the DCMP is the best resource. For guides on grammar, formatting, and other captioning questions, refer to the:

- DCMP Captioning Tip Sheet
- DCMP Captioning Key

CADET Documentation

This step-by-step guide provides all of the information you need on using the Caption and Descriptive Editing Tool (CADET).