Automated speech recognition: tool evaluation and possible workflows for enhancing accessibility of A/V materials

Presented by Florida State University Libraries:
Ruben Aleman, Digital Media & Accessibility Specialist
Bryan Brown, Digital Repository Developer
Dave Rodriguez, Digital Services Librarian
Scope of presentation

- Overview of ASR technology and challenges/opportunities (Dave)
- Introduction to A/V media accessibility (pre-recorded Ruben)
- Introduction to the caption formats (pre-recorded Bryan)
- FSU Libraries’ research with ASR tools (Dave)
- Possible future applications + Q&A (Dave + Bryan via Zoom)
Out of scope of presentation

- Low-level, “nuts n’ bolts” of AI or algorithm mechanics for ASR
- Deep dive into accessibility standards and evaluation
- Non-English language transcription
An important thing to keep in mind...

There is no “out-of-the-box,” 100% accurate means of generating captions for AV media using only machines. All machine-generated transcripts will require some level of human editing/correction/intervention. The goal of this presentation is to discuss which tools provide the best starting place if you need to create captions in-house using ASR.

For now, the only way to create 100% accessible captions is to involve humans in the process.

Additional resource: National Deaf Center - Why ASR is Not the Answer (yet) (2020)
Historically, the core problem with ASR...
What makes for quality captions?

**Accurate**
Errorless captions are the goal for each production.

**Consistent**
Uniformity in style and presentation of all captioning features is crucial for viewer understanding.

**Clear**
A complete textual representation of the audio, including speaker identification and non-speech information, provides clarity.

**Readable**
Captions are displayed with enough time to be read completely, are in synchronization with the audio, and are not obscured by (nor do they obscure) the visual content.

**Equal**
Equal access requires that the meaning and intention of the material is completely preserved.
What we can focus on with ASR evaluation...

**Accurate**
Errorless captions are the goal for each production.

**Equal**
Equal access requires that the meaning and intention of the material is completely preserved.

**Consistent** (formatting decisions based on best practices like [Captioning Key](#))
Uniformity in style and presentation of all captioning features is crucial for viewer understanding.

**Clear** (will always require human intervention until AI reaches scary levels of awareness)
A complete textual representation of the audio, including speaker identification and non-speech information, provides clarity.

**Readable** (text position adjusted with [CSS in the VTT file](#); duration set with cue timing)
Captions are displayed with enough time to be read completely, are in synchronization with the audio, and are not obscured by (nor do they obscure) the visual content.
Introduction to A/V Accessibility (Ruben)
Introduction

- Prevalence of video/audio
- Important that everyone can access that content
Why Captions—Disability

- Users who are deaf or hard of hearing (DHH) - 5.7%
- Users with cognition impairments - 10.9%
- Up to 16.6% of the total population who may not fully perceive audio
  - Without audio alternatives, video content becomes entirely inaccessible
- Legal and moral obligation to include users of all ability
Why Captions—Beyond Disability

- Bad audio quality
- Thick accents or poor pronunciation
- SEO/discoverability
- Reinforces understanding
- Additional context
How Captions

- Where do captions live?
- How do captions appear on the screen?
Introduction to caption formats (Bryan)
Caption file formats: Which one should I use?

- Whichever one your web application supports
- Over 25 caption file formats
- Most are XML (not friendly to human readers)
- Most not used in modern web applications
- 2 modern choices: SubRip or WebVTT
  - Very similar and widely used
  - WebVTT is a superset of SubRip
    - More features
    - Better documentation
    - Better specification
SubRip

- Name comes from **SubRip** DVD subtitle ripping software
- .srt file extension (SubRip Text)
- Cue timing format = hours:minutes:seconds,milliseconds
- First human readable plaintext format
- Good enough for basic needs
Anatomy of an SRT file

1
00:00:00,000 --> 00:00:05,000
This is the first section of an SRT file.

2
00:00:06,000 --> 00:00:10,000
And this is the second.

3
00:00:11,000 --> 00:00:15,000
Here's some <b>bolded</b>, <u>underlined</u> and <i>italicized</i> text.

4
00:00:16,000 --> 00:00:20,000
You can even do <font color="#ff0000">this</font>.
Anatomy of an SRT file

1. 00:00:00,000 --> 00:00:05,000
   This is the first section of an SRT file.

2. 00:00:06,000 --> 00:00:10,000
   And this is the second.

3. 00:00:11,000 --> 00:00:15,000
   Here's some <b>bolded</b>, <u>underlined</u> and <i>italicized</i> text.

4. 00:00:16,000 --> 00:00:20,000
   You can even do <font color="#ff0000">this</font>.
Anatomy of an SRT file

1
00:00:00,000 --> 00:00:05,000
This is the first section of an SRT file.

2
00:00:06,000 --> 00:00:10,000
And this is the second.

3
00:00:11,000 --> 00:00:15,000
Here's some <b>bolded</b>, <u>underlined</u> and <i>italicized</i> text.

4
00:00:16,000 --> 00:00:20,000
You can even do <font color="#ff0000">this</font>. 
Anatomy of an SRT file

1
00:00:00,000 --> 00:00:05,000
This is the first section of an SRT file.

2
00:00:06,000 --> 00:00:10,000
And this is the second.

3
00:00:11,000 --> 00:00:15,000
Here's some <b>bolded</b>, <u>underlined</u> and <i>italicized</i> text.

4
00:00:16,000 --> 00:00:20,000
You can even do <font color="#ff0000">this</font>.
Anatomy of an SRT file

<table>
<thead>
<tr>
<th>Line</th>
<th>Time</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>00:00:00,000 --&gt; 00:00:05,000</td>
<td>This is the first section of an SRT file.</td>
</tr>
<tr>
<td>2</td>
<td>00:00:06,000 --&gt; 00:00:10,000</td>
<td>And this is the second.</td>
</tr>
<tr>
<td>3</td>
<td>00:00:11,000 --&gt; 00:00:15,000</td>
<td>Here's some <strong>bolded</strong> text, <em>italicized</em> text.</td>
</tr>
<tr>
<td>4</td>
<td>00:00:16,000 --&gt; 00:00:20,000</td>
<td>You can even do <strong>this</strong> text.</td>
</tr>
</tbody>
</table>
WebVTT

- VTT = “Video Text Tracks”
- .vtt file extension
- Cue timing format = hours:minutes:seconds.milliseconds
- Created by W3C for HTML5 <track> element
  - Originally WebSRT (subtitle resource tracks) but changed name to WebVTT to avoid confusion
  - Has an official specification ([https://www.w3.org/TR/webvtt1/](https://www.w3.org/TR/webvtt1/))
- Has additional features over SRT
  - Header metadata
  - Font styling
  - Comments
Features of a WebVTT file

WEBVTT
Kind: subtitles
Language: en

STYLE
::cue(b) {
  color: red;
}

NOTE This is a comment

1
00:00:00.000 --> 00:00:05.000
This is the first section
of a WebVTT file.

2
00:00:06.000 --> 00:00:10.000 align:right
And this is the second.

3
00:00:11.000 --> 00:00:15.000
<b>This will appear bolded AND red.</b>
Features of a WebVTT file

WEBVTT
Kind: subtitles
Language: en

STYLE
::<cue(b) {
    color: red;
}

NOTE This is a comment

1
00:00:00.000 --> 00:00:05.000
This is the first section of a WebVTT file.

2
00:00:06.000 --> 00:00:10.000 align:right
And this is the second.

3
00:00:11.000 --> 00:00:15.000
<b>This will appear bolded AND red.</b>
Features of a WebVTT file

WEBVTT

Kind: subtitles
Language: en

STYLE
::cue(b) {
  color: red;
}

NOTE This is a comment

1
00:00:00.000 --> 00:00:05.000
This is the first section
of a WebVTT file.

2
00:00:06.000 --> 00:00:10.000 align:right
And this is the second.

3
00:00:11.000 --> 00:00:15.000
<b>This will appear bolded AND red.</b>
Features of a WebVTT file

WEBVTT
Kind: subtitles
Language: en

STYLE
::cue(b) {
  color: red;
}

NOTE This is a comment

1
00:00:00.000 --> 00:00:05.000
This is the first section of a WebVTT file.

2
00:00:06.000 --> 00:00:10.000
align:right
And this is the second.

3
00:00:11.000 --> 00:00:15.000
<b>This will appear bolded AND red.</b>
Features of a WebVTT file

WEBVTT
Kind: subtitles
Language: en

STYLE
::cue(b) {
  color: red;
}

NOTE This is a comment

1
00:00:00.000 --> 00:00:05.000
This is the first section
of a WebVTT file.

2
00:00:06.000 --> 00:00:10.000 align:right
And this is the second.

3
00:00:11.000 --> 00:00:15.000
<b>This will appear bolded AND red.</b>
Back to Dave
ASR evaluation methodology @ FSU Libraries

- Sourced sample set of 12 AV items from DigiNole, FSU’s digital library and institutional repository
- Strived for sample set to reflect a wide array of content that contains different features which may present issues for ASR (e.g. 1-to-many speakers, accents, sound quality issues, jargon, etc).
- Identified a set of ASR tools we could readily test
  - Whisper AI
  - Microsoft Stream
  - AWS Transcribe
  - Rev API
- Record cost, resource consumption, accuracy, and other important features of each tool against each item in the sample set
  - Accuracy measured using WER (Word Error Rate)
WER (word error rate) analysis

- A common metric for assessing word accuracy in captions/transcriptions

\[ WER = \frac{S + D + I}{N} \]

- Resulting value (%) indicates overall level of errors in a given document

\[ S = \text{Substitution errors} \\
D = \text{Deletion errors} \\
I = \text{Insertion errors} \\
N = \text{Total number of words in the caption/transcript} \]
P.O.V.: you’re conducting WER analysis

don January 24, 2008, our students, alumni, faculty, and friends marked FSU’s 150th anniversary.
with a grand celebration.

This event took place on Westcott Plaza, a location that I’m proud to say marks the oldest
continuous site of higher education in the state of Florida.

The video you are about to watch details Florida State University’s first 150 years.

As you will see, we have crossed many frontiers since our earliest days when
Thomas Jefferson’s grandson, Francis Epps, worked to bring this institution to Tallahassee.

Over the years, we’ve educated Pulitzer Prize winners, NASA astronauts, elite athletes,
great scientists, and some of the world’s greatest actors, dancers, and musicians.

Outstanding liberal arts program, combined with the colleges of business, law, engineering,
Results

Whisper, MS Stream, AWS and Rev API

<table>
<thead>
<tr>
<th>AV item</th>
<th>Whisper</th>
<th>MS Stream</th>
<th>AWS</th>
<th>Rev API</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thank you E...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40th Ann EPC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>150 years of...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethics of Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linked Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beito Interview</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parker Interview</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Young Interview</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reed Interview</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBS Archive</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cat Family R...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wright Famil...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results (cont.)

Average WER results (%)

<table>
<thead>
<tr>
<th>ASR tool</th>
<th>Error rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whisper</td>
<td>1.10</td>
</tr>
<tr>
<td>MS Stream</td>
<td>3.90</td>
</tr>
<tr>
<td>AWS</td>
<td>4.40</td>
</tr>
<tr>
<td>Rev API</td>
<td>2.40</td>
</tr>
</tbody>
</table>
2 of 4 tools were completely free-to-use for the library:
- Whisper = command-line utility openly available via GitHub
- MS Stream = enterprise app provided by FSU ITS

1 tool was free but with strict limitations
- Rev API = free-tier up to 45 min per account
- Regular pricing after limit is $0.02/minute
- Enterprise pricing also available

1 tool was free for a term-limited period
- AWS Transcribe = free-tier is 60 minutes per month for 12 months
- After 12 months, pricing changes depending on usage
  - For 1st 250,000 minutes (~4,1667 hrs) - $0.02400/minute
Resource consumption - 🕒 ⌚️

- **Whisper AI**
  - average ~48 minutes* for files that ranged in duration from 00:02:47 to 00:17:00 (HH:MM:SS)

- **MS Stream**
  - average <5 minutes to create captions once uploaded

- **AWS Transcribe**
  - 1 - 5 minutes once Transcribe function called

- **Rev API**
  - average <2 minutes once API called

* Commands were run on a 2019 MacBook Pro with a 2.6 GHz 6-Core Intel Core i7 processor and 16 GB RAM. Different computing environments would significantly affect run-time speeds.
UI considerations for “workflowization”

- **MS Stream**
  - Pro: YouTube-esque GUI w/ drag n’ drop upload
  - Con: requires manual uploading of titles and retrieval of VTT outputs

- **Whisper**
  - Pro: stand-alone, customizable, portable, and programmable
  - Con: requires CLI and/or developer knowledge

- **AWS Transcribe**
  - Pro: easy to access within AWS controls
  - Con: requires access to AWS controls (usually tightly controlled)

- **Rev API**
  - Pro: batchable
  - Con: requires knowledge of making API calls
Doctor Howard headed an organization.

He was our own son, see.

Leadership,
Reminder: always review your ASR outputs!

12:06.000 --> 12:10.000
If he came back the third time, you got to f*** him.

02:10.480 --> 02:16.040
We absolutely love playing kickball and eating carrot cake, but we just want to let you know that you guys are so

02:16.040 --> 02:19.480
f***ing great and appreciative, and we love you. Bye!
Recommendation

**Whisper AI**
- Best WER results
- Runs as a stand-alone, open-source CLI application
- Does not require agreements with or payment to vendors

Check out the "Show and Tell" section on Whisper’s GitHub for more information on other implementations (e.g. GUI front-ends, etc).
Next steps & possible “workflowization”

Technical side:
- Continue exploring Whisper configurations and customizations
- Experiment in different hardware environments
- Overall, optimize our use of the tool

Administrative side:
- Develop best practices for local use in editing/creating captions
- Seek out funding for OPS workers to be trained and paid for editing work
- Possibly launch a pilot initiative focusing on creating captions for select group of works in DigiNole
Thank you! Questions?

Ruben Aleman: raleman2@fsu.edu

Bryan Brown: bjbrown@fsu.edu

Dave Rodriguez: dwrodriguez@fsu.edu