CLARIN Annual Conference 2018

Pisa, 8-10 October 2018
Can the Computer, and the Public, do the Metadata Work?

Karen Cariani
AAPB Project Director, WGBH
Executive Director, WGBH Media Library & Archives

Casey Davis-Kaufman
AAPB Project Manager, WGBH
Associate Director, WGBH Media Library & Archives
Agenda

• Who we are
• The project
  – Cataloguing dilemma
  – Audio cataloging
  – Speech to text results
  – Crowd source work
  – Crowd source challenges
• CL tool data output – useful?
• Additional available tools - work with CL experts
  – We have a great data set
  – You have expertise
WHO WE ARE
American Archive of Public Broadcasting

The Library of Congress
Packard Campus for Audio Visual Conservation
Who we are:
WGBH Media Library and Archives
Discover historic programs of publicly funded radio and television across America. Watch and listen.
60 years of recorded history, at your fingertips.
Watch and listen at americanarchive.org

50,000 HOURS
of historic public television and radio content are digitized and preserved

ACCESS
the entire collection at WGBH and the Library of Congress

WATCH & LISTEN
to more than 20,000 programs online at americanarchive.org

SEARCH
2.5 million additional catalog records in our database

DISCOVER
content from more than 100 organizations in 40 states and territories

EXPLORE
curated exhibits on topics such as civil rights, climate change, elections, and protesting

HELP
make the AAPB easier to search and access! Play our FIXIT game at fixit.americanarchive.org

Contact Information
aapb_notifications@wgbh.org

Facebook
facebook.com/amarchivepub

Twitter
@amarchivepub

COLLABORATORS

FUNDERS
THE PROJECT
### the situation/dilemma

<table>
<thead>
<tr>
<th>90,000 digitized television and radio programs</th>
</tr>
</thead>
<tbody>
<tr>
<td>incomplete, inaccurate metadata records</td>
</tr>
<tr>
<td>limited staff resources</td>
</tr>
<tr>
<td>what is in the collection?</td>
</tr>
<tr>
<td>users need access to the collection</td>
</tr>
<tr>
<td>continued growth of the collection (content and sparse metadata)</td>
</tr>
</tbody>
</table>
WOMAN IN SUPERMARKET: These are better than most that I’ve seen. Tomatoes in general have been so disappointing that I’ve gotten to the point where I just walk by them, I don’t buy them any more because either they’re very hard and when you eat them they have no taste at all. They’re very wary. Very often I’ve found them rotten inside and you can’t seem to tell it from the outside. They are terribly expensive, and we love them because we’re big salad eaters, you know. But they’re not the kind of tomatoes we used to get years ago.

ROBERT MacNEIL: Good evening. Most of this week we’re all going to be preoccupied with the energy crisis, but tonight we’d like to remind you that this country faces other crises than energy. And one of them is the tomato. Jim?

JIM LEHRER: It’s a crisis of confidence, Robin. Recent opinion polls ranked the tomato among the least loved of all vegetables. Turned-off tomato eaters...
Monroe Duncan

If you have more information about this item than what is given here, we want to know! Contact us, indicating the AAPB ID (cpb-eaicp/387-70zpcgn9).

<table>
<thead>
<tr>
<th>Segment</th>
<th>Monroe Duncan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contributing Organization</td>
<td>WHRO (Norfolk, Virginia)</td>
</tr>
<tr>
<td>AAPB ID</td>
<td>cpb-eaicp/387-70zpcgn9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Description</th>
<th>No description available</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asset type</td>
<td>Unedited</td>
</tr>
<tr>
<td>Genre</td>
<td>Food and Cooking</td>
</tr>
<tr>
<td>Media type</td>
<td>Moving Image</td>
</tr>
<tr>
<td>Duration</td>
<td>00:00:00</td>
</tr>
<tr>
<td>Citation</td>
<td>Chirone: &quot;Monroe Duncan,&quot; WHRO, American Archive of Public Broadcasting</td>
</tr>
<tr>
<td>Series</td>
<td>The Emily Rooney Show</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------------------------------</td>
</tr>
<tr>
<td>Title</td>
<td>WGBH Radio</td>
</tr>
<tr>
<td>Contributing</td>
<td>WGBH (Boston, Massachusetts)</td>
</tr>
<tr>
<td>Organization</td>
<td></td>
</tr>
<tr>
<td>AAPB ID</td>
<td>cpb-aacip/15-j96057dg9n</td>
</tr>
<tr>
<td>Description</td>
<td>Emily Rooney Show</td>
</tr>
<tr>
<td>Asset type</td>
<td>Program</td>
</tr>
<tr>
<td>Topics</td>
<td>Public Affairs</td>
</tr>
<tr>
<td>Publisher</td>
<td>WGBH Educational Foundation, Publisher</td>
</tr>
<tr>
<td>Media type</td>
<td>Sound</td>
</tr>
<tr>
<td>Duration</td>
<td>01:00:00</td>
</tr>
</tbody>
</table>
Wide variety

- Of content from locations across the country
- Lots of different types of speakers
- Different speech patterns and accents
- 40 states

- politicians, academics, man on street, youth, elders, artists
- north, south, mid-west, west, east
- Creole, French, Caribbean Spanish, Alaskan native language
- Music and other non-spoken sound
the potential: transforming content into data

How can we use them?
AUDIO ANALYSIS
High Performance Sound Technologies for Access and Scholarship (HiPSTAS)

generating additional programmatic data about the audiovisual material beyond the actual words being spoken in the recordings

- speaker identities
- applause
- laughter
- musical interludes

4,000 hours
• WGBH-provided training set (all audio) 373.3 hours [speakers we aren't using etc.]
• Extended corpus (all audio) 951.8 hours [collected on own, keywords, etc. might change some of it not; some included in above #]
• "Haystack" test set (all audio) 494.2 hours [gave us to label because we think they're in there; might be some duplicated so we don't use it for training; use it to run the machine. kept on separate]
• Human-labeled speaker segments (all) 103.8 hours [everything]
• UBM_600 corpus 17.9 hours [all labeled speakers they gave us; plus PennSound UBM] Universal Background Model
HiPSTAS AAPB Output

AAPB speaker identity models... so far

Hillary Clinton
Bill Clinton
James Baldwin
Malcolm X
Martin Luther King Jr.
Julia Child
Richard Nixon
Ronald Reagan
Lyndon Johnson
Gloria Steinem

Workshop
Process evaluation
Results

- We didn’t get very much identified
- Really hard to do
- Took a lot of time to identify 1 speaker
- Translating the work from the research at UT Texas and tool training to our own catalog was undefined
SPEECH TO TEXT RESULTS
The Lexicon

- A list of words with their phonetic pronunciations
- The output of the speech recognizer will only include words that are part of the lexicon
- Words can have multiple pronunciations
- The lexicon links the acoustic model and the language model

archive AA R K AY V
hello HH AX L OW
hello HH EH L OW
pop P AA P
wgbh D AH B AX L Y UW JH IY B IY EY CH
world W ER L D
up AH P
The Speech to Text Pipeline

- Task: given audio signal as output, choose the sequence of words that is the best match
- Two components to the model
  - Acoustic model (AM): which phonemes most closely match the signal
  - Language model (LM): which sequences of words are most likely
- We use Kaldi, an open source toolkit for automatic speech recognition (ASR)
- During the process of decoding the audio signal, Kaldi and recognizers store a lattice with hypotheses of the output for each utterance. Hypotheses with lower probabilities get “pruned” out
- At the end of the decoding process, the top hypothesis gets written to the output file
Transcript Creation Continues

• As part of the project, Pop Up Archive released updated language models for Kaldi.
• The code has been “Dockerized” by our University of Texas partners.
• We have incorporated transcript creation into our workflow as the collection grows.
• Github: https://github.com/WGBH/kaldi-pop-up-archive
Crowdsourcing tools

Help Preserve Public Media!

You can make historic public radio and television programs from across America easier to search and access. Help AAPB archivists by participating in any of our Citizen Archivist tools—FIX IT, FIX IT+, or Roll the Credits! Your contributions will be made available in the American Archive of Public Broadcasting.

- FIX IT: A Transcript Game
  - Play our game to correct speech-to-text transcripts
- FIX IT+:
  - Use this simple tool to correct speech-to-text transcripts
- ROLL THE CREDITS:
  - Transcribe credit information from public television programs
Zooniverse Project – “Roll the Credits”
A Conversation With James Baldwin; James Baldwin Interview

One of the significant things about the present revolution of the Negro
people in America, is maybe the fact that for the first time,
there is genuine communication between negroes and whites.
Negroes are saying out loud now
things which they have long said only to themselves.
(Rustling sound):

Possibly one of the most articulate,
passionate, and clear communicators
is my guest, James Baldwin.
James Baldwin's name is known throughout America
for carrying so passionately, and so...
Hello my name is Monica Diaz and I'm going to have a conversation with Adam Payne. ADAM PAYNE: Hi. MONICA DIAZ: Hello, for visible an Oral History Project at the New York Public Library it is is February eighteenth two thousand fifteen and this is being recorded at the the Inwood Public Library in New York. ADAM: It is. MONICA: Thank you so so much Adam. You know can we, can we start off by having you Tell me a little bit about yourself. Well, my name's Adam Payne, um I was born a long, long time ago in a galaxy far, far away. actually I was born in 1971 in Cambridge, Mass.
A production of the Mississippi Center for Educational Television. Number three series a conversation.

"text": "with program 20 30 day to direct."
"start_time": "18.97",
"end_time": "37.78",
"speaker_id": 889358
}

"id": 24387888,
"text": "You you."
"start_time": "37.78",
"end_time": "53.35",
"speaker_id": 889359
}

"id": 24387889,
"text": "MARTIN Carthy metal smith and Potter for"
"start_time": "53.35",
"end_time": "57.94",
"speaker_id": 889360
}

"id": 24387890,
"text": "Marigold Mississippi they are a"
"start_time": "57.94",
"end_time": "61.09",
"speaker_id": 889360
}

"id": 24387891,
"text": "native of the Internet"
"start_time": "63.45",
"end_time": "68.42",
"speaker_id": 889360
}

"id": 24387892,
"text": "you know. I know"
"start_time": "68.7",
"end_time": "73.2",
"speaker_id": 889360
}

"id": 24387893,
"text": "we know and what the noted European are"
"start_time": "73.68",
"end_time": "76.55",
"speaker_id": 889360
}

"id": 24387894,
A Conversation With Theora Hart

008 A conversation in the Mississippi Center for Cultural Travels. Take along your shoes, a hearty conversation.
038 A conversation with Thora Hartwell. Walking in shoes with Mississippi.
053 A similar Thora Hartwell in Los McGuire. MetalSmith and Peter More.
058 A similar conversation. The #1 of Thora Hartwell.

A native Mississippian, born in Paris, Mississippi. Internationally known.

090 A native. We, being as familiar as a name.
114 Though we know the county, the towns, sometimes is, after the three hundredamenti.
119 Our adventures, our experiences, was it ever in full bloom?
129 Windy Pointless, Master of Pointless.
138 A conversation of Thora Hartwell with record for gathering place, whereafter.
130 Extremely, far away in the distance on the hill.
140 Newspapers into doors. A day not like opening. The opening day near.
144 Remember, as the new session in Brussels.
148 In 1901 they were all in. It is not the honor.
155 Which was thoughts from Thora Hartwell's home in Gulfport, Mississippi where.
158 We are living here, experiencing, and in the Museum of Health.
209 Ask a bar, a black, a white, and a dresser by day.
205 Much of the international "hotel condition" at Munich.
213 The Marshall, the man and the "by.
218 Controversially, it was the greatest building in the international.
223 Unanswerable...
226 Please, remember, it is Mississippi. There are enough, as we know.
230 Every day through the week, but definitely approximately after one...gover.
235 This death was also covered by his own, through Roosevelt and others.
240 One is a game, one arena owner, one story.
749 What is your comment about, please have contact with a.
A production of the Mississippi Center for Educational Television. Take number three. A series conversation.

Program with Theora Hamblett. Length 20 minutes 30 seconds. Date 12/16/75. Director Seymour.

A conversation with Theora Hamblett. Talking to today with Mississippi.

Artist Theora Hamblett is Lee McCarty, MetalSmith and Potter from

Marigold Mississippi. "The Art of Theora Hamblett." "

Native Mississippian, born in Paris, Mississippi. Internationally known

Painter. You might be interested to know
Record

En Camino; 205; Special; Juan Serrano, Master Guitarist

If you have more information about this item than what is given here, we want to know! Contact us, indicating the AAPB ID (cbp-aecp/88-65566v97).

<table>
<thead>
<tr>
<th>Series</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>En Camino</td>
<td>A special interview and on stage performance by Juan Serrano, Master Guitarist, who tells about his origins, life and introduction to music and the Art of Flamenco. He began his professional life at thirteen years of age. The art of Flamenco is described as a mix of dance, guitar music, counter rhythms of hand clapping, and 'el duende' which are the deep emotions expressed by the artists.</td>
</tr>
<tr>
<td>Episode Number</td>
<td>205</td>
</tr>
<tr>
<td>Episode</td>
<td>Special; Juan Serrano, Master Guitarist</td>
</tr>
<tr>
<td>Contributing</td>
<td>KRCB (Rohnert Park, California)</td>
</tr>
<tr>
<td>Organization</td>
<td>KRCB FM, Producing Organization</td>
</tr>
<tr>
<td>AAPB ID</td>
<td>cbp-aecp/88-65566v97</td>
</tr>
</tbody>
</table>

| Created         | 1987-06-11                                                                  |
| Asset type      | Episode                                                                     |
| Topics          | Music, Performing Arts, Education, Spanish Language                         |
| Creator         | KRCB FM, Producing Organization, The San Francisco Foundation, Co-Producer, Gaona, Francisco, Co-Producer, Lukens, Jane, Co-Producer |
| Contributor     | Serrano, Juan, Interviewee, Serrano, Juan, Performer                        |
Lord thank you. Lord. You’re through. You’re. Through. Mm mm. Mm. Mm mm mm mm mm mm mm. Mm. Mm. Laura Iraq of. Glory. From. From. From. Three three. Three. Three. Three. Three.

Laura it thank. You. That was it. And those are the good other than those of the other. But other than that I'm.

But you. Know. In PSA jogging for the leg. That. Said alone but not for what they love that about them. In the U.S. I don’t really like deep down this guy John. Has seen me naked young. Youre that one or put up or go to a lie even. When we're going to be in a way that it was. A scene or so because John you and I will note that up why go to a lie. When we're going to be in before and then they run a scene will see My God. You up or something. On Monday another tour and yet. People are coming up a lot of. Joy to make us even more. Of this season might. Be made upon home with us so we're left with. The money to go before the scene was bought.
Kaldi Output for Helicopter B-Roll

A.
Leg.
Length.
Length.
Eh I'm
good.
Thank you.
You mean you think you know always going. In approximately I see that I'm being in a city that the community is Brucey putting I mean in and deal with it and they see it. Are they looking for. And COMINO is improved on so many ideas to explore that they must be in for fancier graphic thing like that Latina. Woman That is keep it down for a. Made up out of there. I mean the older you know the moment they do it and I mean no no shangri la I wound up pretty compressed. But then finally the. Most famous presenter being asked about Odama seeking impact done like a 20 year old uncle. I miss my foresight when local news reported a loss for him which was the original being with you we. Let's see it. No problem I say to you since it was your own this and not a Californian. When you guys buy up our land Mr. Merhi Nera for law enforcement you said to be sure that will cost them. Less personally and COMINO is bloviated not very effective. They must be telescreen are so fake now. But up Roamer they don't know about this the past few months I feel I am sure that. If we didn't know what it is as soon as one has been able to from my strong ally in cancer they mine. I say most I would have had lots of them in Dallas. I remember with Ahmadinejad. See that he's a see the. This was very programa I said Think end up in the Mystica. I think can assume that I would be in the family is in Gaza you know of. Just a movement that eluting mother in that race in Illinois. I mean these people are Latino and they they face. Like the Latinos in going through all of Emmelina repeat them because they see areas because it reduces Saulteaux elated.
All. Right. OK. OK. West Virginia Virginia. Plus. Brown. Girlfriend try a temporary quick. Corollary proper and I pointed to a song. Made on your apartment made I'm easy. I generally. Don't jump to a long. Record. Your point here already called an average storm either calling. Me a fearful planner for us are really powerful for one I don't live for or see me privately just 75 miles from or see. Them. President Reagan National Award. You have more screen. Generation. We have a long trip. For. A trip. Do you see music music from music. And there is a common sense for him when the. Nazi party. No no can one. Man may or. Yet you monkey have made of the tears and. Some are. Good only the Starkey blues the silly your so I look at this hour. Mr Parker knew me or started. Their. Own eyes. What I am. Not very good.
We know nothing about this item.
Your all right already. Yeah OK. For the record. Speaking to James guy Greek looked him up in the crowd doesn't mean we thought i hope I do I'm trying to talk but I think we shall I do you can and we shall not be in hall and you know I do wish I could. I will call that good old myth or the old call wouldn't let me out with a dog. Be sure they left the water good. How does how do you lead that cuckoo clock in the south have you genius. And I don't miss out on your good guy you're not familiar with what do I do when you need to whom you are now there. Dorise you I laughed and I dunno it still gets you can pick on a new record. Yeah gotta be deader than a good guy. Let's go on a con not go off on your own of kind and how can you know how God left her daughter. Julie I don't know we don't really care what you never knew that people can't go through the muck of the kind in which our not. I don't come out now if you're going to cause you need no with no trade in Kenya and it could be should be a good deal or not a number could not go home now or that not gotten bad bitches want to go Honey you do not want me one culture not clear what we've got to do what you want to try New Delhi new kind o woman in her thought didn't think enough on line to get rid of they have a machine like you.
Kaldi Output for Spoken English

• Approximately 81% word accuracy rate
  – not including punctuation errors

• Examples:
  – 95% accurate for 1960s radio program from Boston (no accents, one speaker)
  – 55% accurate for 1970s television program from Mississippi (strong Southern U.S. accent)
<table>
<thead>
<tr>
<th>Item 1</th>
<th>Item 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>The following is a production of the public affairs department of WLPB, Louisiana.</td>
<td>The following is a production of the public affairs department of We're In, Louisiana.</td>
</tr>
<tr>
<td>W will be be. Louisiana</td>
<td>The State We're In, with Public Affairs director, Beth George. Good evening. Thank you for joining us on this first edition of the Louisiana: The State</td>
</tr>
<tr>
<td>the state we're in. With public affairs director Beth George. Good evening. Thank you for joining us on this first edition of the Louisiana the state weary and Tonight we examine with our guests and on film the state prison at Angola. We have with us this evening three knowledgeable guests. We have Mr. C. Paul Phelps,</td>
<td></td>
</tr>
<tr>
<td>was secretary and head of the Department of Corrections. We also have with us Dr. Mark Carlton research director for the Public Affairs Research Council former professor of history at LSU and author of a book on the history of Louisiana's penal system. And finally we have with us this evening Representative Bobby Freeman of Plaquemine, Louisiana state representative who is one of Governor Edwards' chief leaders in the house and a member of the House Appropriations Committee, Angola has been in the news almost constantly since June 1975 when federal district judge Gordon West issued a court order declaring the prison overcrowded and the living conditions unacceptably poor.</td>
<td></td>
</tr>
<tr>
<td>« 12 words »</td>
<td>« 26 words »</td>
</tr>
<tr>
<td>« 29 words »</td>
<td>« 38 words »</td>
</tr>
<tr>
<td>« 15 words »</td>
<td>« 11 words »</td>
</tr>
<tr>
<td>« 27 words »</td>
<td>« 18 words »</td>
</tr>
</tbody>
</table>
Types of Errors Corrected by Users

- Station call letters
- Mis-transcription of words spoken in southern accents, e.g., “weary and” vs “we’re in”
- Local town names, e.g., “plaque and” vs. “Plaquemine”
- Person names, e.g., “Laurence” vs. “Lawrence”
- Numbers spelled out vs. numeric
- Adding words completely missing from original transcript
- Incorrect “corrections” by crowdsourcing participants, e.g. “achieved” vs. “achievedd” in the “corrected” transcript
JSON transcripts are stored on AAPB’s Amazon S3 account.

Transcripts are indexed for keyword searching on the AAPB website.

Transcripts are made available alongside the media on the record page.

Transcripts can play as captions within the player.

Transcripts can be harvested via an API and used as a dataset for research such as a digital humanities project.

once corrected...
A production of the Mississippi Center for Educational Television. Take number three. A series conversation program with Theora Hamblett. Length 28 minutes 30 seconds. Date 12/16/75. Director Seymour. A conversation with Theora Hamblett. Talking to today with Mississippi artist Theora Hamblett is Lee McCarty. MetalSmith and Potter from Marigold Mississippi. "the Art of Theora Hamblett" native Mississippian, born in Paris, Mississippi. Internationally known Painter. You might be interested to know though we know in this country who Theora Hamblett is. What the noted European art critic and historian, Otto Behnarmann, had to say in his book Modern Primitives - Master of Naive Paintings. And there is Theora Hamblett with one of her glaring yellow visons quite sufficient of itself to serve as flashbulbs for the photographers who would surely not miss this opening. The opening they were referring to was the world exhibition in Brussel. The pane they were referring to was...
any Palestinian element or any ultimate choice of governance. Well we may hear more of this or some variant of it. Now that it’s over the week carried New England weather beyond the conversational stage. A phenomenal week starting with the heaviest snow since weather records began, then torrential rain to turn snow banks into flooding. And then yesterday, a May day of temperatures that would have brought the tulips up if they could have found their way through the soggy snow. And to top it off, freezing again, as our weather today swings back to a January norm. Now we shiver at the news of blizzards that have paralyzed the whole middle of the country. Floundering through such a week, has blinded our interest to most other news. And preoccupation with the weather may have made it easier for

the government to hold back information about the disintegration of a
CL experts

- Improving named entity vocabularies
- Forced alignment
- Time stamp for bars and tone
- Music identification
- Foreign language identification and transcription
- OCR of text on screen (lower thirds, credits)
CL Experts and Archivists

- A larger need for more accurate output and ease of use of computational tools for audiovisual archives to create descriptive metadata and annotations.

Help Us!
Access our Dataset

• Metadata API (OAI-PMH and PBCore)
  – https://github.com/WGBH/AAPB2#api

• Transcripts API – contact us to get credentials

• Media as a Dataset (MaaD) – contact us to get digital audio/video copies for computational research
http://fixit.americanarchive.org
#FixItAAPB

americanarchive.org

Karen Cariani
Karen_Cariani@wgbh.org

@amarchivepub

Casey Davis-Kaufman
casey_davis-kaufman@wgbh.org

facebook.com/amarchivepub