Institute of Museum and Library Services National Leadership Grant for Libraries
Funding Opportunity Number: NLG-Libraries-FY22
Proposal: Enhancing open source transcription tools to improve accessibility of A/V collections
March 2022/ National Digital Infrastructures and Initiatives
Narrative

Project Summary

GBH Archives is requesting $167,373 for an 18 month project to improve the transcript output of Kaldi, an open source speech-to-text toolkit, and to update and improve FIX IT+, a crowdsourcing tool to edit transcripts. This project will provide tools for libraries and archives to create affordable and accurate transcripts that will improve the online access of digital audiovisual materials in collections. This will further IMLS goals and objectives of improving the ability of libraries and archives to offer better and wider access to digital collections by providing catalogs and online search engines accurate text for users to discover items of interest.

Project Justification

Archivists and librarians managing large online audiovisual (A/V) collections often struggle to fully catalog materials and guide users to specific A/V content related to their search queries. A single video item could be an hour long documentary of which perhaps 10 minutes is useful to any particular audience. Among thousands of files, it is difficult and time consuming to pinpoint a 10 minute segment, particularly when often the only descriptive metadata available for the full-length item is a vague title. This is particularly true for hour-long news programs covering multiple topics and events during a single program.

The availability of transcripts for online A/V collections increases discoverability through catalogs and search engines and improves accessibility for finding specific content within these collections. In addition, transcripts for audiovisual materials help provide compliance with the Americans with Disabilities Act (ADA), as a secondary source allowing access to the speech content present in a sound recording of an audio or video file. Spoken words within a transcript can be indexed, exposed to search engines, and used to identify potential topics of interest. Transcripts that are time-stamped can also pinpoint where in an hour-long video or audio program a specific topic is discussed, thus getting the user directly to the content desired. Manually creating transcripts is time consuming and labor intensive. Using AI and machine learning tools to automate creation of speech-to-text transcripts could be a time saver for archives and libraries with limited resources to fully catalog media materials and thus can increase discoverability of their collections.

Transcripts have proven to be a valuable resource for scholars and researchers using audiovisual archives. Dr. Allison Perlman, Associate Professor of Film and Media Studies and History at University of California Irvine, wrote to us that “The transcripts included alongside some of the videos hosted by the American Archive of Public Broadcasting (AAPB) are a vital resource for researchers….The transcript included alongside the video was vitally important to me as I mapped the themes and implicit arguments of the film; the program utilized interviews and voice-over narration extensively, and the transcript was crucial to enabling me to analyze the text efficiently...”

Commercial services are a viable option, but it would be a significant cost to transcribe large collections, for which most cultural institutions don't have funding. Amazon's transcription services with bulk workload could be as little as $1/hour of content. For a collection the size of the AAPB, that would be about $120,000, and the collection continues to grow each year. For smaller one-off transcription needs, the cost of a commercial service may be a minor consideration, but there are benefits to supporting an open-source toolkit. Commercial services are black boxes and must be used as is, and there is no way to adapt the service to your own specific data and collection needs. Ethical, social and legal considerations must also be considered when using commercial services. The general lack of transparency regarding how resulting data is used by companies like Amazon and Google do not always reflect the needs for protecting privacy and dignity when dealing with sensitive archival materials.

However, there are few alternative open-source speech-to text, or automatic speech recognition (ASR) tools.[1] A well-trained and easy to use open-source speech-to-text solution could be less expensive for archives to implement. Commercial services benefit from their clients' large datasets by using them as training data to improve their tools, but these proprietary improvements do not necessarily return to the open-source community. Thus, developing and keeping data sets and training improvements in an open community will create opportunities for improvements across broader types of collections with diverse voices and speech. It could also encourage collaboration between AI and machine learning experts, and collection or data set owners like librarians and archivists, to continue to improve the tools, and perhaps inspire new ones to be developed and shared with the open-source and library community.

In 2015, GBH Archives was awarded an IMLS research grant to work with the Pop Up Archive to create speech-to-text transcripts for the entire American Archive of Public Broadcasting (AAPB) collection, which at the time included a diverse collection of 68,000 digitized video and audio items produced at public media stations across the United States. The goal was to test whether machine-generated transcripts -- even with inaccuracies -- that were indexed into the archive catalog and harvested by search engines would increase discoverability of the collection. As part of the project, Pop Up Archive utilized the Kaldi open-source speech-to-text toolkit, initially developed by Johns Hopkins University,[2] to create the machine transcripts.

Since then, GBH has continued to use the Kaldi toolkit to create the speech-to-text transcripts of new materials added to the AAPB, an additional 62,000 items. The AAPB has content from 150+ public media TV and radio stations, including material from Native American, Creole, Appalachian, African American, Asian American, and Latinx communities, from metropolitan cities to rural locales across the country. The quality of the transcripts from the AAPB content, consisting of such a diverse collection with various regional accents, use of language, speech patterns, musical intervals, and non-English speakers, is limited and wanting. Evaluations have shown that Kaldi-produced transcripts of recordings with good audio quality featuring white, Northern U.S. speakers can yield accuracy rates as high as 91-95%, while transcripts of recordings featuring BIPOC (Black, indigenous and people of color) speakers and others with

[1] "The Best 7 Free and Open Source Speech Recognition Software Solutions." Goodfirms, https://www.goodfirms.co/blog/best-free-open-source-speech-recognition-software.
[2] Povey, Daniel. *Kaldi ASR*, 2021, kaldi-asr.org/.

regional or cultural accents, dialects and vernacular (e.g. Cajun and Appalachian accents, etc.) can yield transcripts with accuracy rates as low as 55-57%. Feedback from users suggest significant errors in transcripts featuring people of color, including programs with well-known speakers such as James Baldwin, Cesar Chavez, and Muhammad Ali.

Kaldi is considered one of the most widely adopted open source toolkits available to create machine transcripts, according to Towards Data Science: "Kaldi is widely adopted both in Academia (400+ citations in 2015) and industry."[3] A search in Semantic Scholar conducted in March 2022 resulted in 4,060 citations. (https://www.semanticscholar.org/search?q=kaldi&sort=relevance). A search in Google Scholar returned over 5,800 citations , of which one fifth have been since 2021 (https://scholar.google.com/scholar?cites=8690069263871570191&as_sdt=2005&sciodt=2007&hl=en). In addition, Kaldi does not require GPUs to run and is therefore accessible to normal servers with conventional CPUs. This is important for libraries and archives with fewer computing resources. Known users of Kaldi are hard to determine as Kaldi consists of a toolkit containing lots of models and components. There is a need to focus on training for certain types of audiovisual assets, like TV or radio programs, which are not typically part of existing language models in Kaldi.

The AAPB transcripts are created with the Kaldi toolkit, but the output needs improvement to address challenges of handling a broad range of diverse content. Adding preprocessors to identify and isolate recognizable speech audio (e.g., speech vs. other sounds such as music, machines, and natural events) as well as English from multilingual input would increase the accuracy and efficiency of the software currently trained only for English speech. Further improvement of the Kaldi toolkit by, for example, training on different speech patterns and English accents, or adding high-performing statistical language models, would improve the output for transcripts.

A more accurate, simple to use, open source solution for speech-to-text is needed for small cultural heritage organizations with widely differing content to create transcripts from oral histories and A/V materials. Improving the output of diverse sources of audio would allow those diverse voices to be more clearly represented. Archives from all parts of the country would have better transcripts and be more likely to use an open-source tool. If their efforts are contributed back to the open collection of data, it could be used for future tool training and improvement. The ability to create transcripts for materials would greatly enhance discoverability and help ensure compliance with the ADA and the World Wide Web Consortium's Web Content Accessibility Guidelines (WCAG 2.1). Other A/V collections that could benefit from an open source speech-to-text tool are Indiana University's mMedia collection, the Paley Center, Washington University's Film and Media Archive, and any library with A/V materials in their special collections.

Machine-generated transcripts, even the most accurate, and especially the most inaccurate, need to be reviewed and validated for accuracy. Crowdsourcing this effort through online tools alleviates the burden on resource-thin archives and can help engage their communities to

[3] Ramon, Yoav. "How to Start With Kaldi and Speech Recognition." Towards Data Science, 22 Nov. 2018, https://towardsdatascience.com/how-to-start-with-kaldi-and-speech-recognition-a9b7670ffff6.

improve the quality of the transcripts. Crowdsourcing has also become a common place practice to help archives transcribe items in their collections. An article from Jan 13, 2020 in *Crowdsourcing Week* titled "The Decade in Crowdsourcing Transcription." states:

> "In 2010, crowdsourcing was considered an experiment. Now, it's a standard part of library infrastructure: seven state archives run transcription and indexing projects on FromThePage, and some (like the Library of Virginia) run additional projects on other platforms simultaneously. The British Library, Newberry, Smithsonian Institution, and Europeana all run crowdsourcing initiatives. …Library schools are assigning projects on crowdsourcing to future archivists and librarians...
>
> What's ahead for the next decade [2020s]?...
>
> …..Despite that progress, tabular documents like ledgers or census records remain hard to transcribe in a scalable way. Audio transcription seems like an obvious next step for many platforms — the Smithsonian Transcription Center already begun with TC Sound."[4]

As noted, there are limited tools and efforts of crowdsourcing the transcription of audio and A/V materials. Easy crowdsourcing tools to help fix or improve machine transcripts enhances these resources and creates opportunities for archives to engage their communities in public history and cultural heritage efforts. Correction of a better quality original transcript may take only approximately 2.5-3 times the run time, while correction of a lower quality original transcript can take up to 6-7 times the total run time. Thus, improving the speech-to-text tool will ultimately result in more transcripts corrected over time and improve volunteer satisfaction. Crowdsourced editing would be more efficient and may improve volunteer motivation and accuracy if less editing was required, as suggested in an article in *Crowdsourcing Week* (Brumfield).

Other open-source tools exist for crowdsourcing transcription of archival materials. University College London's Bentham project has utilized Media Wiki and customized it for their project, but this project focuses only on textual records. The Library of Congress' By the People crowdsourcing project has released its code under an open-source license, but again this initiative is currently limited to textual records.[5] The Smithsonian is one of the few organizations that have initiated audio crowdsourcing efforts and has used an in-house home-grown tool specific to their platform that is not open-source nor easily shared. Further, the tool does not utilize speech-to-text transcripts but rather requires volunteers to copy and paste from PDF documents and add timestamps or manually type out the full transcript.

For the 2015 grant, GBH Archives experimented with improving the transcripts through crowdsource editing. The project supported the development of an interactive game created by the GBH Digital department and the implementation of an open-source transcript editing tool initially created by the New York Public Library (NYPL) as part of its Together We Listen

---

[4] Brumfield, Ben. "The Decade in Crowdsourcing Transcription." Crowdsourcing Week, 18 Jan. 2020, crowdsourcingweek.com/blog/decade-in-crowdsourcing-transcription/.
[5] "By The People." *By the People*, Library of Congress, 1 Mar. 2021, crowd.loc.gov/about/.

project.[6] The NYPL tool provides an easy to use platform for the public to correct and edit machine created transcripts while listening to the audio file. GBH Archives installed and made slight enhancements to the NYPL tool and called it FIX IT+.[7] GBH Archives' main additions to the NYPL tool included fixes for bugs inherited from NYPL, adding functionality for converting transcripts from JSON to WebVTT (both common file formats for time-stamped transcripts), automating ingest and release workflows, and user interface (UI) enhancements.

AAPB FIX IT+ has been successful at engaging volunteers and public media stations in correcting transcripts. After four years of use, there are several key areas of improvement that we have identified that would benefit not only the AAPB, but also other users who install the FIX IT+ tool.

Massachusetts Historical Society is using the NYPL tool for their transcription project. NYPL is still using the tool for their Together We Listen project. Last year, Northeastern University reached out to WGBH to ask about the use of the FIX IT+ tool and any improvements that had been made. No doubt there are others using the tool and all institutions would benefit from code improvements and updated installation documentation. At the Library of Congress' Informal Audio Visual Summit (IVAV) event in September 2020, a panel of archivists discussed how the COVID-19 pandemic and transition to remote work enabled them to shift student and other paid workers to focus on accessibility and transcription for their digitized A/V collections while they were unable to perform their regular job duties from home.[8] When asked if paid staff would continue working on transcripts when they returned to "business as usual", the panelists were in agreement that these efforts would not continue, and that new strategies were needed to continue the important transcription work. The improvements to FIX IT+ in this project could help make crowdsourcing of transcription a more sustainable practice in libraries, cultural institutions and archives by engaging their public users and patrons.

The NYPL is no longer maintaining the original code as it has since shut down the NYPL Lab that created the tool. Enhancements such as component and code upgrades, bug fixes, refactoring, and dockerizing would better protect the application against vulnerabilities, extend the overall lifespan of the application, and make it easier for developers from other organizations to launch their own instances. Moreover, the GBH Archives developers in consultation with the Outreach Manager who led "transcript-a-thons" with users of FIX IT+, found that users would like new features such as improving how transcripts are flagged by users for more work, ways to recognize user's contributions in order to drive engagement, and improving how non-speech noise is identified in a transcript. A public and admin reporting dashboard, workflow improvements for adding and exporting transcripts, and features allowing the user to control

---

[6] Kelly, Alexandra. "Together We Listen: Make Hundreds of NYC Stories Accessible-One Word at a Time." The New York Public Library, The New York Public Library, 5 Apr. 2016, www.nypl.org/blog/2016/04/05/together-listen-oral-history-transcription.

[7] FIX IT+. American Archive of Public Broadcasting, 2021, http://fixitplus.americanarchive.org.

[8] Davis Kaufman, Casey, Laney, Michael, Magar, Jonah, Titkemeyer, Erica, and Tarr, Kimberly, panelists. Panel discussion. "Informal Audio Visual Summit." Library of Congress, *LC Labs*, 15 Sept. 2020, labs.loc.gov/static/labs/events/documents/LC-IVAV-Agenda-09042020-1.pdf.

audio playback speed and editing functionality, could greatly improve use by archivists and by those volunteering their time to correct transcripts. By significantly improving the reporting functionality of FIX IT+, archives and libraries will be able to easily view statistics regarding the number of transcriptions corrected in a time period, which will help them better forecast future work and make determinations regarding personnel and volunteers.

The former IMLS project resulted in improved discoverability with search engines, increased traffic to individual items in the collection, and community engagement through fixing the transcripts. The AAPB website has had a 584% increase in visitors over the last 5 years (with over 53% of visitors arriving from search engines) and more than 1,000 transcripts have been corrected. This project will give other archives and libraries with audiovisual collections tools to enhance their collection discoverability.

GBH Archives has chosen to continue to use the Kaldi toolkit and FIX IT+ and improve them both. After five years of working with both tools, the GBH Archives developers and archivists have gained proficiency with the tools and feel it is more efficient to improve what is currently in use than start over with a new tool. Also, given the persistence of these tools in the community, improving the tools could benefit other organizations that have adopted them.

Project Work Plan

Brandeis University's Computational Linguistics department, is currently working with the GBH Archives on an artificial intelligence (AI) pipeline for libraries and archives, called CLAMS (Computational Linguistics Applications for Multimedia Services) (https://clams.ai). The overall project aims to use state-of-the-art AI and neural network machine learning tools, such as Kaldi and other Natural Language Processing tools, to extract and create useful descriptive metadata from audiovisual files. The hope is that this will save archivists and crowdsourcing participants cataloging time to better describe content in digital audiovisual collections, thus improving access by users. Creating accurate speech-to-text transcripts from audio is one step in the pipeline.

From the work on the CLAMS project, the Brandeis team has identified weaknesses in the Kaldi toolkit that may hinder creating accurate transcripts. Kaldi is one tool in the CLAMS pipeline, but considering the availability of open-source text analysis tools, we believe that speech-to-text is one of the most crucial components of the pipeline. Improvement of Kaldi to create more accurate transcripts would certainly benefit other downstream text analysis tools in the pipeline. In addition, an improved open-source speech-to-text tool that can stand on its own would be a great benefit to the audiovisual archive community and other media collections and archives.

Neural network models are machine learning tools where a computer learns to perform a task by analyzing training examples or data, and can assign learned labels and categories to new unseen data. They have been proven successful in machine natural language processing. Brandeis has computational linguistics faculty and graduate students who have experience with training neural network models and are interested in improving the Kaldi toolkit. One advantage the larger commercial companies have to improve their automatic transcription services is the large amounts of data that flow through their systems to continuously train and improve their tools. Publicly available, pre-trained high-parameter neural network models from big companies (such

as BERT from Google, GPT from OpenAI, and Megatron from Nvidia) will be used and tuned to produce better language models that can be integrated into the existing Kaldi toolkit.

Even though the AAPB has over 120,000 items, it is still considered a small dataset for training machine learning transcription tools. Since AAPB is lacking a large amount of data to train an end-to-end speech-to-text system, the goal of this project will be more focused on swapping one very crucial piece of the Kaldi toolkit, namely the language model (predicting word sequence from predicted sound sequence), with a modern neural network-based model. With the wide diversity of voices in the AAPB collection it is a great data set to use for improving the Kaldi toolkit.

Currently, the improvements being suggested are for English speech to English transcriptions. This project budget and scope is not adequate to address other languages. GBH will work with Brandeis students to improve the output of the Kaldi tool by iteratively testing and evaluating against the current transcript output. The GBH team will test the tool improvements by comparing the new transcripts created with the updated tool against the current AAPB transcripts. Brandeis and GBH will evaluate the improvements on a variety of subsets of data that help reflect the broad range of accents present within the AAPB collection, and Brandeis will explore how finetuning the models for these sub-corpora might improve overall accuracy rates. The improved Kaldi toolkit will be made available under an open-source Apache license in a GitHub repository maintained for 3-5 years after the end of the award.

GBH developers will improve the FIX IT+ crowdsourcing tool. Workflows that are currently manual will be automated making it easier for a non-developer to implement in an archives environment. A dashboard of reporting tasks to indicate statistics such as how many transcripts are completed, how many lines have been edited, and how many people completed editing, will be created for administrators to measure crowdsourcing progress, as well as features allowing user control of the audio playback. The improved FIX IT+ will also be made available under an open-source Apache license in a GitHub repository maintained for 3-5 years after the end of the award.

The project will be accomplished through execution of two distinct timelines with milestones for both the Kaldi and FIX IT+ development and implementation.

In **Kaldi Phase 1 (Months 1-3),** the Brandeis team will install and apply pre-trained high-parameter models. In **Kaldi Phase 2 (Months 4-6)**, both the Brandeis researchers and GBH staff will identify a diverse selection of materials in the collection to use for testing to evaluate the various models and will compare the output to the current AAPB transcripts. The teams at Brandeis and GBH will identify areas of improvement. In **Kaldi Phase 3 (Months 7-15)**, using an iterative method of fine-tuning, testing and refinement of the pre-trained model, the project teams at Brandeis and GBH will evaluate the model performance after each iteration by comparing transcripts and how many words are correct vs. incorrect on the original transcripts. The final development objective for Kaldi in **Kaldi Phase 4 (Months 13-18)** will be to embed the updated tool into CLAMS pipelines and evaluate downstream performance.

While the Brandeis team takes the lead on improving Kaldi, GBH will simultaneously work on improving the **FIX IT+ crowdsourcing tool**. The original code needs crucial updates for the Ruby, Rails, Gulp, and Node versions to extend the lifespan of the application. The GBH Developer and GBH Supervising Developer will begin by reviewing the core components of the application to determine whether updating to more recent stable versions of the code libraries better serves the needs of the application versus switching to new ones. Once these updates are in place, GBH will design and implement a prototype of a user dashboard. The GBH developer will gather feedback on the new dashboard from a small pool of users who have previously corrected transcripts for the AAPB and other users of FIX IT+. The GBH Developer will also build new import and export features into the platform as well as automation scripts to allow for direct ingest of corrected transcripts to the AAPB website. Finally, new features will be added providing users with more control of audio playback functionality in the user interface. Furthermore, GBH developers will use previously created user-stories to improve the user-interface in order to better identify the status of a transcript and to improve the experience of correcting transcripts. A key metric to crowdsourcing is engagement. Through conversations, GBH developers have mapped out potential ways to recognize user contributions such as badges or leaderboards and to provide avenues for feedback from volunteers. The GBH Supervising Developer will review code produced by the GBH Developer.

The updates to the technology stack will take place during **FIX IT+ Phase 1 (Months 1-2)** of the grant. Improving the user-interface and how transcripts are flagged in FIX IT+ is planned for **FIX IT+ Phase 2 (Months 3-5)**. The next time-consuming feature is to build out reporting and dashboards, which we have slated for **FIX IT+ Phase 3 (Months 6-9)**. During **FIX IT+ Phase 4 (Months 10-12)** we plan to improve the ingest of transcripts and dockerize the application. **FIX IT+ Phase 5 (Months 13-18)** will be spent improving the workflow, gathering user feedback on the dashboard, and fixing known bugs that currently exist in FIX IT+. During this time GBH will also focus on improving user engagement, testing the application with input from crowdsourcing volunteers, and improving it through iteration.

Dissemination of the project will take place over the entire duration of the project through conference presentations, press releases and project updates via blogging and social media. In **Months 16-18,** the project teams will focus targeted energy toward dissemination of the improved Kaldi and FIX IT+ tools. During this phase, the project teams will distribute the code and conduct outreach to the library and archives communities via webinars, demonstrations, and sharing information through listservs and professional communities encouraging installation and use of the tools. The project teams will provide guidance and support to others making use of the code through demos and zoom calls, and will implement a ticketing system in Github to gather feedback from users. Although face-to-face presentations and demonstrations are effective, COVID has proven that remote conferences, webinars, and demonstrations are also efficient and allow more attendees from different locations to attend. Both remote and in-person dissemination activities will be undertaken.

Dissemination will be to both the library and archival audiovisual communities, as well as the computational linguistics community. Presentations at conferences, both live and virtual in both spheres will demonstrate the tools and explain use. The Brandeis team will propose conference sessions at the Association for Computational Linguistics, Coling, Latex, and Fantastic Futures

conferences. GBH will propose sessions to the Association of Moving Image Archivists and International Association of Sound and Audiovisual Archives. The project team will host virtual demos and info-sessions and invite users from the library, archives, public media, and computational linguistics communities.

The Project PI will oversee the project ensuring it stays within stated goals, on budget, and overseeing staff work and dissemination. They will also coordinate the communications and collaboration between the GBH team and the Brandeis team. Monthly budget and project tracking meetings will take place assessing the money spent, and the project milestones accomplished, to date. The attached schedule of completion will be the reference point for success and adjustments will be made as needed to keep the project on track overall. Both GBH and Brandeis teams have proven to be able to continue work even with COVID restrictions.

Evaluation

Success of an improved Kaldi toolkit will be measured by comparing the accuracy of transcripts created with the current Kaldi tool to transcripts created today vs. with the newly updated Kaldi tool. The team will select a set of audio and video files representing a range of content types (e.g. news program, documentary, talk show, raw interview, programs containing music or other non-speech sound), as well as programs featuring speakers with a variety of accents and speech patterns across geographic region and race/ethnic groups. The team will then create transcripts of each selection using both the original Kaldi and the improved Kaldi tools. A corrected transcript will be produced for each item. Then, the team will use a tool such as Copyscape to create a word-for-word comparison from the original and corrected transcripts. This comparison will result in a Word Error Rate, which will allow the teams to compare the rate of errors of the original Kaldi tool with the improved tool. To measure increased use of the improved Kaldi tool and FIX IT+, Brandeis and GBH will gather quantitative data on the number of times that the tools are cloned and forked in Github.

GBH will further measure improvements to the FIX IT+ tool by conducting a survey of past crowdsourcing participants in addition to a smaller virtual focus group to gather additional feedback and evaluate user experience. GBH will also survey organizations that install the tool to measure the effectiveness and quality of the administrative and workflow improvements.

Brandeis and GBH will survey organizations that install Kaldi by providing links to surveys in the Github repository and directly contacting organizations that the team is aware of installing the tool. The survey will gather information about the ease of tool documentation and installation as well as the intended uses of the tools and feedback on quality of output.

Project Results

This project will provide two affordable open source tools for archives and libraries with A/V collections that will improve discoverability. Specifically, the project will result in 1) an improved output for speech-to-text transcripts for English speaking audiovisual materials using the open-source Kaldi toolkit, 2) an improved FIX IT+ transcript editor tool that will be shared with the broader community helping libraries and archives more easily engage their communities in crowdsourcing. Documentation for use and installation of both tools will be shared to the broader library and archive communities and maintained for at least 3-5 years after the project

award. GBH will host the code on Github for at least 3-5 years or until a next generation tool or enhanced version becomes available at which point we would direct users from the Kaldi repository at Github to the improved tool repository.

Long term results will be 1) the AAPB collection will have improved transcripts to enhance the discoverability of items in the collection, particularly those from diverse local regions of the country; 2) graduate students in computational sciences will be exposed to using their skills to create tools for libraries and archives in hopes that upon graduation they consider careers other than large, for-profit companies; 3) broader use and distribution of both the Kaldi and FIX IT+ tools across the library, archives, cultural and linguistics communities.

With this project, other audiovisual collections of materials from specific regions of the country or unique dialects or accents will benefit from an improved output of Kaldi and an updated FIX IT +. Large and small audiovisual collections such those housed at Indiana University, WNYC, Northeastern University, UCLA, or the Paley Center could benefit from a more accurate open source speech to text tool. As the efficiency of Kaldi is improved - both from an updated language model to pre-processing pipelines, the more likely it will be used more broadly. In turn, greater use of the tool will draw attention to even better potential updates by professionals in the machine learning field. In addition, the accuracy of transcripts for a broader variety of English accents and speech patterns, as well as wider variances in audio quality - not studio perfect, will improve. This in turn will improve access to and discovery of more diverse content from the wide ranging regions of the country.

Both GBH and Brandeis are committed to hiring and retaining workers and students from underrepresented communities, supporting growth and advancement. Highlighting the collaboration between the GBH Archive and Brandeis University Department of Computational Linguistics will inspire other libraries and archives to reach out to experts in the fields of machine learning to develop projects of equal benefit to both. And likewise, professionals seeking various data sets to use to train tools they are creating will reach out to archives with rich and varied collections.

Improving Kaldi and FIX IT+ will help foster greater equity and democratization of access to historical A/V materials documenting underrepresented communities, perspectives and collections. The project will also help address some of the hindrances to access to A/V archives for people with disabilities by improving the open-source tools needed to create transcripts and captions from archival content. The improvements to Kaldi could help make accessibility goals among cultural heritage institutions more affordable and sustainable. Improving machine learning tools for broader use in libraries and archives, and providing examples of positive uses of these tools could encourage broader engagement from the computational professions and help archivists adapt to the growing need for skills in managing large audiovisual datasets.

Institute of Museum and Library Services National Leadership Grant for Libraries	GBH PI/PM = Red						Brandeis Team = Green											
Funding Opportunity Number: NLG-Libraries-FY22	GBH Developers = Blue																	
Proposal: Enhancing open source transcription tools to improve accessibility of A/V collections	GBH Systems Admin = Yellow																	
WGBH Educational Foundation (GBH)																		
Schedule of Completion: 09/01/2022 - 02/28/2024																		
Activity	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Kaldi Phase 1 (Months 1-3)																		
The Brandeis team will install and apply pre-trained high-parameter models.																		
Kaldi Phase 2 (Months 4-6)																		
Brandeis and GBH staff will identify a diverse selection of materials in the collection to use for testing to evaluate the various models and will compare the output to the current AAPB transcripts, identifying areas of improvement.																		
Kaldi Phase 3 (Months 7-15)																		
Using an iterative method of fine-tuning, testing and refinement of the pre-trained model, the project teams at Brandeis and GBH will evaluate the model performance after each iteration by comparing transcripts and how many words are correct vs. incorrect on the original transcripts. This work is primarily led by Brandeis with feedback and evaluation support from GBH.																		
Kaldi Phase 4 (Months 13-18)																		
The final development objective for Kaldi will be for Brandeis to embed the updated tool into CLAMS pipelines and evaluate downstream performance.																		
FIX IT+ Phase 1 (Months 1-2)																		
GBH will make needed updates to the FIX IT+ technology stack including the Ruby, Rails, Gulb and Node versions.																		
FIX IT+ Phase 2 (Months 3-5)																		
Using feedback already gathered from FIX IT+ users, GBH will make improvements to the existing user interface in FIX IT+.																		
FIX IT+ Phase 3 (Months 6-8)																		
GBH will build out reporting features and a dashboard.																		
FIX IT+ Phase 4 (Months 10-12)																		
GBH will improve the ingest of transcripts and dockerize the FIX IT+ application.																		
FIX IT+ Phase 5 (Months 13-18)																		
GBH will improve the workflow for ingest and export of transcripts, gather user feedback and fix bugs identified during testing and focus groups with crowdsourcing volunteers.																		
Kaldi/FIX IT+ Dissemination and Evaluation (Months 16-18)																		
While dissemation will take place over the full duration of the project, extra outreach effort will be spent in months 16-18 after the tools have been published. The project teams will host demos, reach out to libraries, archives and the computational linguistics community to share the tools, encourage use, and gather feedback.																		

Institute of Museum and Library Services National Leadership Grant for Libraries
Funding Opportunity Number: NLG-Libraries-FY22
Proposal: Enhancing open source transcription tools to improve accessibility of A/V collections
March 2022/ National Digital Infrastructures and Initiatives
Digital Products Plan

Type of Products to be Created

The project will result in the creation of:

- Updated software for the Kaldi toolkit - Brandeis will adapt and update language models for the automatic speech recognition (ASR) toolkit Kaldi. These models predict word sequences from sound sequences and are an essential part of ASR.
- Updated software for the FIX IT+ Transcript Editor - GBH Archives will build upon the existing transcript editor, FIX IT+, in order to update its components, improve the workflow for non-technical staff, add reporting features, streamline the user-interface for correcting transcripts, and add features to drive user engagement.
- Documentation on installing and using each of the two software packages. The documentation will be stored as .txt, html/markdown and/or PDF files and hosted in the projects' Github repositories.
- Transcripts produced as part of the testing and evaluation of new Kaldi features. The quantity of transcripts will depend on project needs and the number of iteration cycles throughout development. Resulting transcripts will be stored as text and json formats and then transformed into .vtt format when ingested into FIX IT+.

Availability

All software and source code for Kaldi will be made available at the CLAMS Project GitHub organization at https://github.com/clamsproject/ or a similar GitHub organization. Docker images will be made available at https://hub.docker.com/u/clamsproject.

The updates for FIX IT+ will be incorporated into our code available at: https://github.com/WGBH-MLA/transcript-editor.

GBH Archives and Brandeis will produce documentation for each software product. Copies of all Kaldi documentation will be made available for public access on the project's Github repository.
Copies of all FIX IT+ documentation will be preserved in the GBH Archives along with project reports and transcripts created using the software. The materials will be preserved off-line in GBH Archives' state-of-the-art vault and migrated over time in accordance with GBH Archives' digital preservation plan, as well as stored on institutional servers managed by the GBH IT department. Documentation for installing and using the software will be hosted in the project's Github repository.

GBH Archives will keep final transcripts produced by the project that will be stored on AAPB's Amazon S3 server and made available for correcting via FIX IT+ as well as for indexing and viewing on the AAPB website. This workflow has been tested and used since 2017 when AAPB staff began using Kaldi to create transcripts.

All final transcripts created as a result of this project will be indexed for searching and discovery on the AAPB website at https://americanarchive.org. The transcripts will also be searchable alongside the media player with timestamps that can be used to navigate through the media. Transcripts will be hosted on the AAPB S3 server and indexed for search and discovery on the AAPB website.. Transcripts for items available in the AAPB's Online

Reading Room will be accessible online to anyone in the United States. Transcripts for materials not available in the AAPB's Online Reading Room will be viewable on location at GBH and the Library of Congress. Staff can also provide Limited Research Access to materials not available online and for researchers in other countries via password-protected, time-restricted access.

Access

Brandeis University will create software and research data for Kaldi that will be copyrighted by Brandeis University. All Kaldi software and data will be released under non-restrictive licenses like the Apache 2.0 license or a Creative Commons license. These licenses allow redistribution and modification and are well suited to drive future development and will be bundled with the software and research data. GBH Archives will update and improve the existing FIX IT+ software that was initially developed by the New York Public Library, and originally released under an MIT license. GBH will release updates to the code and documentation under the same MIT license so that it may be redistributed and modified for future development. Any final transcripts created will be subject to the license provided by the producing organization, which GBH Archives received when the stations donated materials to the AAPB.

During the refinement of the Kaldi toolkit, GBH and Brandeis will use training data (audio and transcripts) that does not contain any sensitive or private information, selecting only among materials that have already been approved for the AAPB Online Reading Room in accordance with AAPB's access policies.

Sustainability

The team will ensure sustainability through the use of open licensing, common tools and languages in use by libraries and archives and ongoing maintenance of code for at least 3-5 years after the grant award. Brandeis uses the CLAMS (Computational Linguistics Applications for Multimedia Services) platform to deliver the models. The CLAMS platform uses open-source, publicly-available software like Python and Docker Containers to deliver models and processing pipelines. All code and documentation is delivered as GitHub repositories that can be freely copied, adapted and used, thereby guaranteeing continued availability. In some cases Kaldi processing benefits from earlier processing steps; all code for those steps, including third-party software, is also available in GitHub repositories. Kaldi will also be delivered as a standalone application for libraries and archives that are not using the CLAMS platform.

FIX IT+ is open-source software that supports import and export of the common speech-to-text transcript format WebVTT. All code is hosted on Github, and all of the changes made to FIX IT+ will be pushed back to the codebase and available to all users, as software developers at the GBH Archives have done for the past five years. The programming languages used for FIX IT+ are Ruby and Javascript. The frameworks are Ruby on Rails and Backbone. Other frontend libraries include underscore.js and jQuery. GBH Archives uses these languages and libraries because these are core to the original Transcript Editor code built by the NYPL, which GBH is proposing to enhance through this project.

All project results will be released to the general public. Theoretical findings will be published in relevant conferences and journals. Source code generated will at all times be accessible through the project's git repositories for a period of at least 3-5 years.

Institute of Museum and Library Services National Leadership Grant for Libraries
Funding Opportunity Number: NLG-Libraries-FY22
Proposal: Enhancing open source transcription tools to improve accessibility of A/V collections
March 2022/ National Digital Infrastructures and Initiatives
Organizational Profile

WGBH Educational Foundation (GBH), legally governed by the GBH Board of Trustees and led by President and CEO Jonathan C. Abbott, is America's preeminent public broadcasting producer, the source of fully one-third of PBS' prime-time lineup, along with some of public television's best-known lifestyle shows, children's programs and many public radio favorites. GBH's mission is to enrich "people's lives through programs and services that educate, inspire, and entertain, fostering citizenship and culture, the joy of learning, and the power of diverse perspectives." The mission of GBH is included in the amended and restated bylaws, approved by the Board of Trustees on June 3, 2020. Its productions include *Nova*, *Frontline*, *American Experience*, *Antiques Roadshow*, *Masterpiece*, and many children's programs such as *Arthur* on PBS. GBH is the number one producer of websites on pbs.org, the most-visited dot-org on the Internet. GBH is a pioneer in educational multimedia, as well as technologies and services that make media accessible to the 36 million Americans who rely on captioning or video description services. GBH has been recognized with hundreds of honors: Emmys, Peabodys, duPont-Columbia Awards, and even two Oscars. In 2002, GBH was honored with a special institutional Peabody Award for 50 years of excellence. For more information visit www.wgbh.org.

The GBH Archives, a department within the Legal and Business Affairs division at GBH, establishes the policies and procedures for access, acquisition, intellectual control, and preservation of GBH's physical media and digital media production and administrative resources. Today the GBH Archives' collection constitutes over 750,000 production and administrative assets including film, video, audio, computer, stills, and print media. The collection is extensively used by GBH television, radio and educational projects. The Archives maintains collection and shot level content databases, including copyright and source details, of originally produced and acquired footage and stills. The GBH Archives is open weekdays to academic scholars and students by appointment. The GBH Archives maintains a public facing website called Open Vault (http://openvault.wgbh.org) which has had more than 2 million visits by scholars, educators, journalists, filmmakers, students and lifelong learners since launching in 2007.

In August 2013, the Corporation for Public Broadcasting selected GBH in collaboration with the Library of Congress as the permanent stewards of the **American Archive of Public Broadcasting (AAPB),** an initiative that preserves and makes accessible significant historical content created by public television and radio. For more information visit http://americanarchive.org/. The AAPB team has worked with over 150 public media organizations around the country to digitize more than 130,000 public programs and more than 70,000 programs are available through an Online Reading Room. AAPB's web visits currently total more than 550,000 annually. Digitized content from the collection is preserved for future audiences at the Library of Congress. The AAPB has launched numerous projects to build national capacity for the preservation of public media, including the IMLS - funded "Improving Access to Time-Based Media through Crowdsourcing and Machine Learning" project, three Mellon grants working with the Brandeis Lab for Linguistics and Computation and to build the Archival Management System 2.0 based on the Samvera open source solution bundle Hyrax and PBCore, and the AAPB National Digital Stewardship Residency program and the AAPB Public Broadcasting Preservation Fellowship program.