WGBH Educational Foundation - GBH Archives
Institute of Museum and Library Services National Leadership Grant for Libraries
Proposal: Enhancing open source transcription tools to improve accessibility of A/V collections

Project Justification – WGBH Educational Foundation (GBH) respectfully submits this proposal to the IMLS National Leadership Grants for Libraries program in alignment with IMLS' Goal 3 "Advance Collections Stewardship and Access" and Objective 3.2 "Promote access to museum and library collections." In collaboration with Brandeis University, GBH requests $154,152 to improve two open-source tools for the creation and correction of automatic speech recognition (ASR), or speech-to-text transcripts, in order to enhance discoverability of audiovisual (A/V) collections. Creating ASR transcripts is limited to using commercial services at a cost, especially for large collections, or using one of the few open source tools (GoodFirms). In 2015, GBH was awarded an IMLS research grant to work with the Pop Up Archive to create ASR transcripts for the American Archive of Public Broadcasting (AAPB). The goal was to test whether ASR transcripts, even with inaccuracies, indexed into the archive catalog and search engines would increase discoverability. In addition, crowdsourcing was used to improve transcripts through editing via a game created by GBH and an open-source transcript editing tool initially created by NYPL. GBH installed and made slight enhancements to the NYPL tool and called it FIX IT(a platform for the public to correct transcripts). The project resulted in improved discoverability with website searches, increased traffic to individual items in the collection, and community engagement. The AAPB website has had a 625% increase in visitors over the last 5 years, and now consistently has about 40,000 visits per month. Since that project, GBH has continued to use Kaldi, an open source toolkit initially developed at Johns Hopkins University in 2009, to create the ASR transcripts of new materials added to the AAPB that are in English. The AAPB has content from 140+ public TV and radio stations, including material from Native American, Creole, Appalachian, African American, Asian American, and Latinx communities, from cities to rural locales across the country. The accuracy of the transcripts comprised of such a diverse collection with various regional accents, use of language, speech patterns, and non-English speakers, is limited. Kaldi is considered one of the most widely adopted open source ASR tools available according to Towards Data Science, but the output needs improvement. Crowdsourced editing would be more efficient and may improve volunteer motivation if less editing was required, as suggested in an article in Crowdsourcing Week. This IMLS NLG "Implementation" project will seek to address insufficiencies with Kaldi and FIX IT+.

Although commercial services through companies like Amazon, Google, Trint, and Rev exist and their output accuracy is pretty good, at volume they are often cost prohibitive to many organizations. A more accurate, simple to use, open source solution for ASR is needed for cultural heritage organizations to create transcripts for A/V materials. 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), in addition to making diverse voices more clearly represented. 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.

Easy crowdsourcing tools to improve ASR transcripts enhances the data and creates opportunities for archives and libraries to engage communities in public history and cultural heritage efforts. Other open-source tools exist for crowdsourcing transcription of archival materials, but most focus only on textual records. The Smithsonian is one of the few organizations to initiate audio crowdsourcing efforts and has used an in house developed tool specific to their platform that is neither open-source nor easily shared. Further, the tool does not utilize ASR transcripts but rather requires volunteers to copy and paste from PDF documents and add timestamps or manually type out the full transcript.

AAPB has implemented FIX IT+, a tool that has been successful at engaging volunteers and stations in correcting transcripts. After three years of use, there are several areas of improvement that GBH and users have identified that would benefit not only the AAPB, but also other organizations that use the tool. The NYPL is no longer maintaining the code, therefore enhancements such as component and code upgrades, bug fixes, ingest refactoring and dockerizing are needed. A reporting dashboard and workflow improvements would also improve usability by archivists and crowdsourcing volunteers. After five years of working with both Kaldi and FIX IT+, the GBH Archives developers and archivists have gained proficiency with both and feel it is more efficient to improve what is currently in use than start over with new tools. The linguistics expertise brought by Brandeis, along with their experience leveraging and improving open-source tools for analyzing language, will ensure the project's success. Given the persistence of these tools in the community, improving them would also benefit other organizations that have adopted them.

Project Work Plan - GBH will work with Brandeis University's Computational Linguistics department led by Dr. James Pustejovsky to improve Kaldi. While Brandeis leads the updates to the Kaldi language model, GBH will provide data for testing and will iteratively evaluate improvements. One advantage the commercial companies have to improve their machine transcription services is the large amounts of data that flow through their systems to continuously train and improve their tools. Even though the AAPB has over 120,000 items, it is still considered a small data set for training machine learning transcription tools. Since we are lacking a large amount of data to train an end-to-end ASR system like Kaldi from scratch, the goal of this project will be more focused on swapping one (but very crucial) piece of the ASR pipeline, namely the language model (predicting word sequence from predicted sound sequence), with a modern neural network-based model. While future work is needed to improve models for other languages, this project will focus on the English language only. Brandeis has computational linguistics graduate students who have experience training neural network models and are interested in solving this problem. Publicly available, pre-trained high-parameter models from big companies (BERT from Google, GPT-2 from OpenAI, and Megatron from Nvidia) will be used and tuned to produce better language models (using smaller but more relevant transcript data sets) that can be swapped into the Kaldi pipeline. The transcript outcomes will be tested and evaluated against the current Kaldi tool, and then the improved tool pipeline will be made available under an open source Apache license in GitHub. In addition, GBH plans to improve the FIX IT+ crowdsourcing tool, with input from the user community. The original code needs to be updated for the Ruby, Rails, Gulp, and Node versions, and some steps that are currently manual could be automated, making it easier for a non-developer to implement in an archive or library. 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, would be useful for administrators. GBH developers will improve FIX IT+ workflows, automating them when possible, and will implement a simple to use dashboard.

Diversity Plan - Improving Kaldi and FIX IT+ will help foster greater equity and democratization of access to historical A/V materials documenting underrepresented communities and perspectives. The project will also help address some hindrances to access for people with disabilities by improving the 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 example 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 A/V datasets. Both GBH and Brandeis are committed to hiring and retaining workers and students from underrepresented communities, supporting growth and advancement.

Project Results – The project will result immediately in 1) an improved output for ASR transcripts using Kaldi, an open source toolkit; and 2) an improved 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 will be shared to the broader library and archive communities and maintained by GBH and Brandeis for at least 3-5 years after the project. These tools can be implemented across a wide variety of media including oral histories. 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; and 3) the archives and library community will have at its fingertips an improved open-source toolkit for transcription and transcript correction, which can be further improved upon well into the future.

Budget Summary – The total funding request is $156,097. Funding includes: 1) $52,753 for GBH Salaries (Karen Cariani (CO-PD - oversee Kaldi improvements), Casey Davis Kaufman (CO-PD - oversee FIX IT+ improvements), Business Manager, Technical Project Manager and Developer Team, 2) $19,550 for Fringe Benefits at 37.06% based on GBH's federally negotiated rates, 3) $5,401 for Administration costs (occupancy, computers and phones, which are not included in GBH's Indirect Expenses); 4) $4,672 for Travel for 1 person (GBH) to disseminate the project at two national conferences (airfare, hotels, meals, ground transportation, registration fees); 5) $49,540 for Brandeis University for Professor James Pustejovsky & Senior Research Scientist Marc Verhagen, and a computational linguistics Graduate Student to improve the Kaldi tools (Includes - Salary, Fringe & Travel and their Federally approved Indirect rate of 62.5%), and 6) $24,181 for GBH's federally negotiated Indirect Expenses at 22.52%. The Fringe Benefit and Indirect Expense rates are currently being negotiated with our cognizant agency.