INTRODUCING HIGH PERFORMANCE SOUND TECHNOLOGIES FOR ACCESS AND SCHOLARSHIP

Tanya Clement (University of Texas, Austin, USA)
David Tcheng (Illinois Informatics Institute, University of Illinois at Urbana-Champaign, USA)
Loretta Auvil (Illinois Informatics Institute, University of Illinois at Urbana-Champaign, USA)
Tony Borries

1. Introduction

In August 2010, the Council on Library and Information Resources (CLIR) and the Library of Congress (LoC) issued a report titled *The State of Recorded Sound Preservation in the United States: A National Legacy at Risk in the Digital Age*. This report explains the fact that preserving and accessing sound collections in the humanities is a complex problem. Our sound heritage continues to deteriorate on legacy formats making digitization of the utmost importance, but preservation and access cannot be solved through digitization alone. As Mark Greene and Dennis Meissner remark in their 2005 article "More Product, Less Process," "processing is not keeping up with acquisitions, and has not been for decades, resulting in massive backlogs of inaccessible collections at repositories across the country."[30] The same is true of unprocessed and therefore inaccessible archives of spoken word sound collections that hold important cultural artifacts such as poetry readings, story telling, speeches, oral histories, and other performances of the spoken word. The subject range and number of digitized and born-digital recordings like these that libraries and archives receive for processing is daunting, and it will only increase.

Observing the general dead air in this audio access and preservation soundscape, CLIR's *Survey of the State of Audio Collections in Academic Libraries* and CLIR's report with the LoC, *National Recording Preservation Plan,* cite copyright legislation reform, organizational initiatives for shared preservation networks, and improvements in the processes of discovery and cataloging as the areas where research and development for increasing access are most needed. They call for "new technologies for audio capture and automatic metadata extraction"[31] with a "focus on developing, testing, and enhancing science-based approaches to all areas that affect audio preservation"[32] to help relieve these dark backlogs of undescribed, even though digitized, audio collections. At the same time, Greene and Meissner and the 2010 CLIR report make it clear that audio preservation is contingent on audio use: if scholars and students do not use sound archives, our cultural heritage institutions will be less inclined to preserve them.[33] So too, in order to discern the minimal level of processing needed to create access, Greene and Meissner ask, "What is the least we can do to get the job done in a way that is adequate to user needs, now and in the future?"[34]

In order to identify user and infrastructure development needs and therefore increase access to (and preservation of) significant digitized spoken word sound recordings, the School of Information (iSchool) at the University of Texas at Austin (UT) and the Illinois Informatics Institute (I3) at the University of Illinois at Urbana-Champaign (UIUC) are hosting a year-long Institute in Advanced Technologies in the Digital Humanities funded by the National Endowment for the Humanities called High Performance Sound Technologies for Access and Scholarship (HiPSTAS), the first meeting of which took place in Austin in May 2013. Including

30 Mark A. Greene and Dennis Meissner, "More Product, Less Process: Revamping Traditional Archival Processing," in ***The American Archivist*** vol. 68, issue 2, 2005, 208–209.
31 A. Smith, ***et al.***, ***Survey of the State of Audio Collections in Academic Libraries***, (Washington, DC: Council on Library and Information Resources, 2004), 11.
32 Brenda Nelson-Strauss, ***et al.***, ***The Library of Congress National Recording Preservation Plan***, (Washington, DC: Library of Congress, 2012), 15.
33 Council on Library and Information Resources and the Library of Congress, ***The State of Recorded Sound Preservation in the United States: A National Legacy at Risk in the Digital Age***, (Washington DC: National Recording Preservation Board of the Library of Congress, 2010), 16.
34 Mark A. Greene and Dennis Meissner, "More Product, Less Process: Revamping Traditional Archival Processing," 240.

twenty humanities junior and senior faculty and advanced graduate students as well as librarians and archivists from across the U.S. interested in analyzing large collections of spoken word audio collections, the first meeting comprised expert panels and workshops for the participants. The year-long work will include support for use case studies, limited software development, and a final meeting to assess implementation needs. The objectives of the Institute are threefold: (1) to perform an assessment of user requirements for large scale computational analysis of spoken word collections of keen interest to the humanities; (2) to complete an assessment of infrastructure needed for short term (sandbox) and long term (sustainable) access and deployment of supercomputing resources for visualizing and mining large audio collections for humanities users; and (3) to produce preliminary results using these supercomputing resources within an example dataset of interest to the participants.

In particular, this paper will give an introduction to the HiPSTAS project by discussing the impetus for the HiPSTAS project, the ARLO (Adaptive Recognition with Layered Optimization) software we are using for access and analysis, and the basic user requirements that were identified from introducing the participants to ARLO.

2. Background and related work

Current processing workflows that depend on the constant presence of a person who must listen to each recording one-by-one in order to provide discoverable metadata require human resources that are simply impractical. Even though we have digitized hundreds of thousands of hours of culturally significant audio artifacts and have developed increasingly sophisticated systems for digitization, management, and delivery of sound, there is little provision for the kinds of analysis that let one discover, for instance, how prosodic features change over time and space or how tones differ between groups of individuals and types of speech, or how one poet or storyteller's cadence might be influenced by or reflected in another's. There are few means by which a librarian or archivist can discern the genre, the composition, or the quality of an audio file on a large scale. Currently, with analog recordings, a librarian or an archivist must rely on unconfirmed legacy labels on old, dusty boxes and must listen, in real time, to each recording to confirm or describe its contents in descriptive metadata. Listening to an analog recording might destroy it, but listening to each digital recording in order to create metadata is still prohibitive since it requires precious human resources. The same is true of born-digital recordings "made in the field," which can be rushed or produced in such volumes that accompanying metadata is lacking. Yet, to date, there are no widely used technologies specifically designed to augment metadata creation for spoken word collections. Finally, there are no opportunities for those interested in spoken word texts such as speeches, stories, and poetry to use or to understand how to use high performance technologies for accessing and analyzing large collections of sound.

At this time, there are a few free open-source audio and video content management systems that enhance access for end users to audio and video in well-designed environments that work well with repository infrastructures. While Murkurtu has been built with indigenous communities to manage and share digital cultural heritage, the Avalon Media System at Indiana and Northwestern and the Oral History Metadata Synchronization project (OHMS) out of the University of Kentucky are both open source systems specifically designed for managing large collections of digital audio and video files that enable users to curate, distribute and provide online access to their collections for purposes of teaching, learning, and research. There are still other systems that allow users to segment or to create clips and playlists for audio organization such as the Stories Matter Project at Concordia. Finally, there are other professional tools such as Kairos, a lightweight content management framework specifically designed for editing metadata, transcoding (or ripping) and creating derivatives, and synchronizing changes across audio and video files; and GLIFOS (used at the Briscoe Center for American History and the Benson Latin American Collection), which uses a rich-media wiki designed much like Avalon and OHMS to automate the production, cataloging, digital preservation, and access, as well as the delivery of rich-media over diverse data transport platforms and presentation devices. With these tools, archivists and librarians can

link together diverse materials that all relate to a single event, but these are collections that have already been processed, even minimally, with metadata.

This is the day and age when computer performance—in terms of speed, storage capacity, and advancements in machine learning—has increased, however, to the point where it is now possible to automate some aspects of how we discover and catalog large audio collections. The very popular Digging into Data challenge is a testament to the wide array of perspectives and methodologies digital projects can encompass. In particular, the first (2009) and second (2011) rounds of awards include projects that are using machine learning and visualization to provide new methods of discovery. Some analyze image files ("Digging into Image Data to Answer Authorship Related Questions") and the word ("Mapping the Republic of Letters" and "Using Zotero and TAPoR on the Old Bailey Proceedings: Data Mining with Criminal Intent"). Others provide new methods for discovery with audio files by analyzing large amounts of music information—the "Structural Analysis of Large Amounts of Music" and "the Electronic Locator of Vertical Interval Successions (ELVIS)" project—and large scale data analysis of audio, specifically the spoken word (the "Mining a Year of Speech" and the "Harvesting Speech Datasets for Linguistic Research on the Web" projects).

The HiPSTAS project is about leveraging this research and development in the sciences and social sciences and with music and natural language in order to apply it to the problem of creating descriptive metadata for large and important spoken word collections in the humanities such as poetry, folklore, and oral history collections. For example, analysis of the spoken word is related to the work that scholars have done for decades on features of music and bird song that include pitch, tempo, and accent. J. Stephen Downie and Michael Welge developed a system for comparing different music retrieval systems (MIR) for which they received funding from the National Science Foundation. The I3 group also collaborated on NEMA (Networked Environment for Music Analysis), which brings together the collective projects and the associated tools of world leaders in the domains of music information retrieval, computational musicology, and digital humanities research. Examples of how NEMA could be used for music analysis helped in understanding how ARLO could be used in the HiPSTAS Institute for analyzing spoken word audio. The NEMA system, for instance, can be used for genre and mood classification as well as composer identification (corresponding to identifying genre, mood, and author in spoken word audio); for similarity retrieval where similarity is measured on prosodic features of pitch, tempo, and accent or the key or tone of music; and structural segmentation evaluation that identifies the key structural sections in music such as a change in verse, movement, or the addition of a chorus (which can correspond with segmenting stanzas in a poem or sections of spoken audio that contain a story).

3. ARLO (adaptive recognition with layered optimization) software

A significant part of the HiPSTAS Institute includes introducing participants, all of whom had never used advanced machine learning technologies for accessing and analyzing sound, to the ARLO (Adaptive Recognition with Layered Optimization) software. Developed for classifying bird calls and using spectral visualizations to help scholars classify pollen grains, ARLO has the ability to extract basic prosodic features such as pitch, rhythm and timbre for matching, discovery (clustering), and automated classification (prediction or supervised learning), as well as visualizations for spectral matching. The original implementation of ARLO for modeling ran in parallel on systems at the National Center for Supercomputing Applications (NCSA) at the University of Illinois, Urbana-Champaign. As part of HiPSTAS, the I3 team has implemented ARLO on the Texas Advanced Computing Center's Stampede system, one of the largest computing systems in the world for open science research. Subsequently, the I3 team has developed ARLO to take advantage of parallel processing on the super computing system and developed ARLO's interface (and documentation) to allow the HiPSTAS participants to test the machine learning system in real time.

The machine-learning algorithm ARLO uses to find events in audio is called "instance based learning" (IBL). In IBL, the machine memorizes a number of examples and matches them

against new examples to predict events. To find a match in an audio stream is to find it in a certain number of positions per second, which is part of the supervised discovery parameters chosen by the user. ARLO finds matches by taking each known example and "sliding" it across new audio files looking for good matches. The spectrograms used in ARLO show the amount of energy in different frequency bands over time. For example, the spectrograms shown in the figures below are based on a two dimensional matrix of energy which is represented by numerical values. Each row of pixels is a frequency band presented across an X-access of time. The frequency band functions much like an inner hair cell in the human ear as it responds to sound waves. The color of each pixel represents the numerical value of energy of a particular frequency for that point in time or how much the hair tremors. The energy value represents the total energy a hair cell (tuning fork) has at a given time, which represents the sum of potential energy (the deflection of the fork or hair) and kinetic energy (based on the speed of the movement). The spectrogram shows a map of that energy relationship at any point in time with a heat-based color scheme. The lowest values are black (cool), and then blue, green, red, yellow, and the points with the highest or most intense energy values are white.

Figure 2 shows three voices saying the same three-word phrase—"some such thing"—reading from Gertrude Stein's novel *The Making of Americans*. From left to right is a computer-generated voice, Gertrude Stein, and then Gregory Laynor, an experimental artist who sings the novel from start to finish. It is easy to see through the visualization the differences between the regulated computer voice on the left, Stein's more dramatic emphasis (on "some") and de-emphasis (on "such") of the words, and Laynor's sung version in which the words tend to bleed into each other. Figure 3 shows an excerpt from Ezra Pound's Canto XLV recorded in Washington, D.C., 1958 (top) and at Harvard University, Boston, Massachusetts, 1939 (bottom) from the PennSound archive. The recordings produced almost twenty years apart sound almost identical but similar words (in this case "usura") appear differently in the spectrogram due to differences too subtle for the human ear. Figure 4 shows an example of Pound's choice to use the word "design" in the first reading, and "delight" in the second reading of the same poem. Figure 5 shows the once hidden recording of Robert Frost reading "Stopping by Woods on a Snowy Evening" among other poems on Side B of folklorist William A. Owens' recordings within the Oral History of the Texas Oil Industry Records at the Dolph Briscoe Center for American History at the University of Texas, Austin. Currently a user can only discover the recording because a diligent archivist included that fact in the metadata. Figure 6 shows an excerpt from a 2007 interview with Larry Aitken, the tribal historian from the Leech Lake Band of Ojibwe conducted by Dr. Tim Powell when he was director of the Center for Native American studies at the University of Pennsylvania. The second image has been tagged by a user to show the different ways that English, spoken Ojibwe, drum-beats, and chanting are visualized. With enough such examples, ARLO can be trained to automatically identify these different genres.

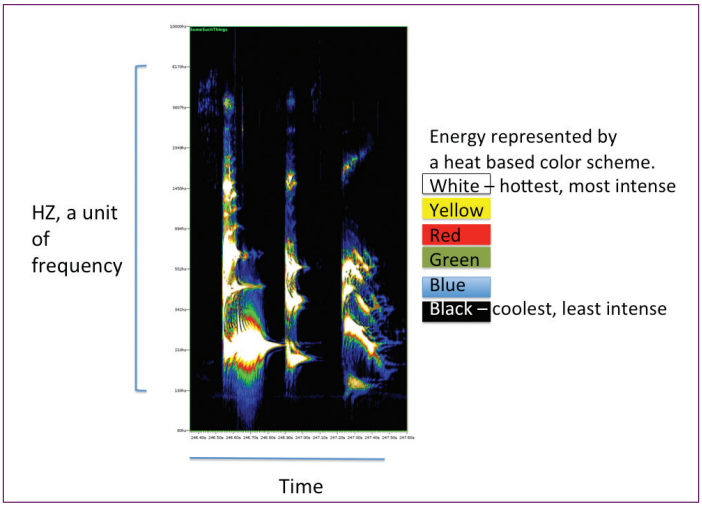

Figure 1: In this visualization of a woman speaking, each row of pixels is a frequency band presented across an X-access of time.

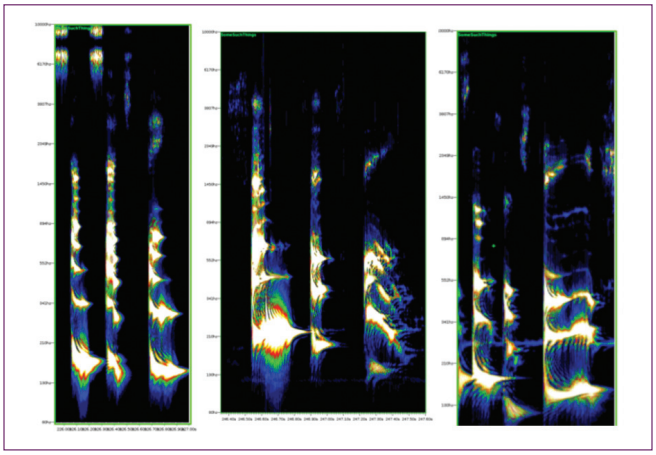

Figure 2: Spectogram created with ARLO of three voices speaking the same words "some such thing," reading from Gertrude Stein's novel The Making of Americans. From left to right: computer-generated voice, Gertrude Stein, and Gregory Laynor, experimental artist.

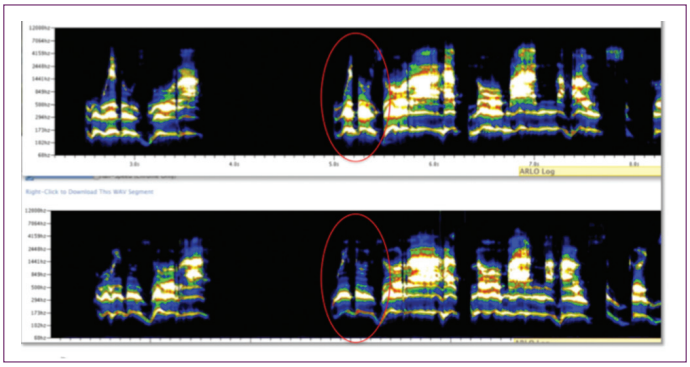

Figure 3: From the PennSound Collection, Ezra Pound's Canto XLV recorded in DC, 1958 (top) and Harvard, 1939 (bottom). The recordings sound the same but similar words appear differently in the spectrogram.

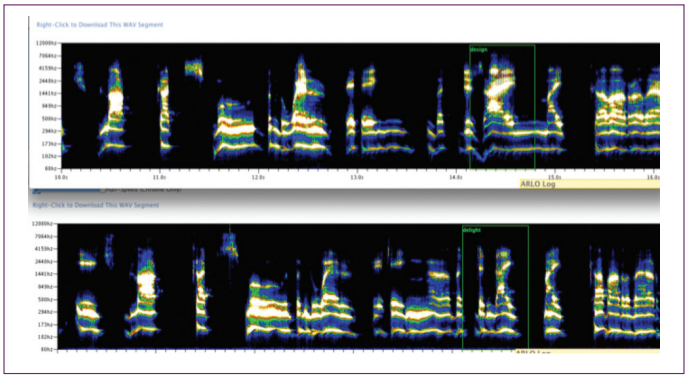

Figure 4: From the PennSound Collection, Ezra Pound's Canto XLV recorded in DC, 1958 (top) and Harvard, 1939 (bottom). This example shows Pound's choice to use the word "design" in the first reading, and "delight" in the second reading of the same poem.

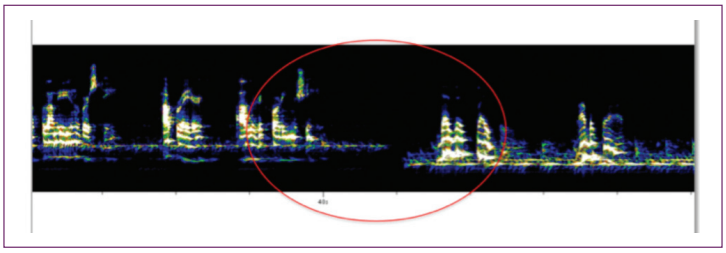

Figure 5: From the William A. Owens Collection, Dolph Briscoe Center for American History, University of Texas at Austin. July 1939 Iowa City, Iowa. Robert Frost reading "Desert Places" poems on Side B. of William A. Owens' folklore recordings. Shifts in speakers and in genre such as shown here are easily discernible by ARLO.

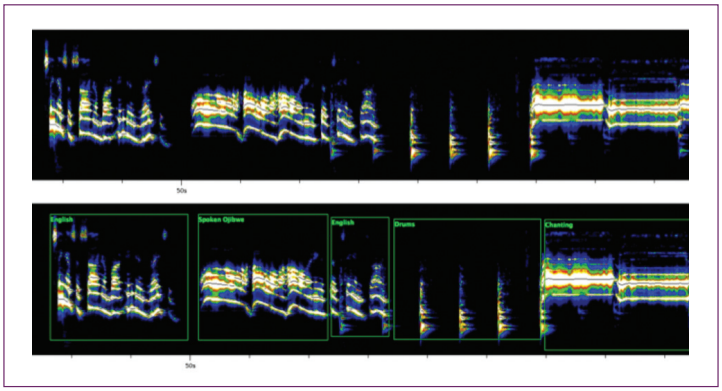

Figure 6: The above two figures show an excerpt from a 2007 interview with Larry Aitken, the tribal historian from the Leech Lake Band of Ojibwe conducted by Dr. Tim Powell when he was director of the Center for Native American studies at the University of Pennsylvania. The second image has been tagged by a user to show the different ways that English, spoken Ojibwe, drum beats, and chanting are visualized.

The development work on ARLO for HiPSTAS has included limited interface development for humanities users, such as the ability to analyze longer files, adding short keys for play, stop, and fast-forward, as well as infrastructure development that allows multiple users to use multiple collections and to perform exploratory discovery (clustering) and automated classification (prediction or supervised learning) processes as well as visualizations on collections of interest to them. This initial development was essential for the May 2013 Institute meeting for identifying user and technological infrastructure needs.

4. User needs

HiPSTAS participants were professionals and graduate students accustomed to working daily with large audio collections such as PennSound's poetry archive, which includes approximately 30,000 files of recordings; the 600,000 digital collections objects and growing in the American Folklife Center (AFC) of the Library of Congress; the 30,000 hours of oral histories in the StoryCorps collection; and the 3,000 hours in the American Philosophical Society's Native American Collection, which includes recordings from more than fifty tribes across Native America. Other collections of interest to the participants included collections of speeches of such luminaries as Ralph David Abernathy, Jesse Jackson, and Martin Luther King, Jr. in the Southern Christian Leadership Conference recordings currently archived at Emory University; 700 recorded readings and lectures in the Elliston Poetry at the University of Cincinnati; and thirty-six interviews in the Dust, Drought and

Dreams Gone Dry: Oklahoma Women and the Dust Bowl (WDB) Oral History Project out of the Oklahoma State Libraries. The participants had never, before the HiPSTAS Institute, had access to data mining, visualization, and supercomputing resources for analyzing these collections.

Already, as a result of the first meeting, we have a base understanding of minimal and more complex user needs. For analysis, users want to tag or note laughter, silence, emotions, applause, pauses, and feedback noises, as well as shifting speakers, languages, and dialects. They want to use ARLO to understand how these sometimes-subjective aspects of speech map to patterns of tempo and rhythm, pitch, tone or timbre, and sound dynamics. For access purposes, users want to use ARLO to automatically mark (a) information about speakers including the number of different speakers on a track and when the speaker changes, the genre of the speech (such as a monologue, poetry, an interview, elicitation, and the presence of music), as well as, in some cases, to train the software to identify the speaker and the language being spoken; and (b) information about the performance or recording including breaks in tracks between interviews or performances, the location of dead space or "drop offs," as well as areas of the recording that are damaged or seem to contain unintelligible "noise."

A final project result is greater literacy in sound analysis, visualizations, and infrastructure development for the humanities. Increased visibility for sound collections and scholarship that incorporates sound artifacts means that more scholars and students will, as Charles Bernstein suggests, encourage treating sound as texts and data for study. Participants want to learn more about how to use the parameters of digital sound that affect sound visualization (to which ARLO can give them access) such as damping ratios, gain, frequencies, spectra, and pitch energy. Understanding what the users of sound collections want to do and can do with software like ARLO is only a first step. Johanna Drucker cautions that "[s]oftware and hardware only put into effect the models structured into their design" and advises that if humanities scholars want digital humanities tools with "the subjective, inflected, and annotated process central to humanistic inquiry, [humanists] must be committed to designing the digital systems and tools of our future work."[35] Accordingly, the HiPSTAS institute has two primary outcomes: (1) to produce new modes of access to and analysis with large scale audio collections using advanced technologies such as classification, clustering, and visualizations; and (2) a broadened engagement in the work of digital infrastructure development through contributing to recommendations for the implementation of a suite of tools for collecting institutions interested in supporting advanced digital scholarship in sound.

5. Conclusion: future work

This initial year of HiPSTAS is the first step in our attempt to create open source software that may be used by any institution interested in using advanced technologies for accessing and analyzing spoken word sound collections. Taking advantage of the advanced computing resources that are part of Stampede is essential for assuring the kind of powerful infrastructure that high performance computing on sound files requires. At the end of this first phase of the project, there will be an initial public release of ARLO for use with the collections in the project including PennSound, the Folklore Center Archives at the Briscoe American History Center, and other collections that the participants will have the opportunity to make publicly available. This release will include documentation for users and developers. Eventually, with more funding and additional research, we will release an open access software bundle that other libraries and archives can use to setup their own systems, to share computational resources, and to share data across collections and institutions. We will also make available through the project website use cases developed by project participants who shepherd and use the sound collections that are part of the project as well as any resulting scholarship from documented usability studies that reflect how users interact with sound collections in the ARLO environment.

35 John Drucker, "Blind Spots," in *The Chronicle of Higher Education*, April 3, 2009, 5.

References:

Bernstein, C. *Attack of the Difficult Poems: Essays and Inventions*. University of Chicago Press, 2011.

---. *Close Listening: Poetry and the Performed Word*. New York: Oxford University Press, 1998.

Council on Library and Information Resources and the Library of Congress, *The State of Recorded Sound Preservation in the United States: A National Legacy at Risk in the Digital Age*. Washington DC: National Recording Preservation Board of the Library of Congress, 2010.

Drucker, J. "Blind Spots". In *The Chronicle of Higher Education*, April 3, 2009.

Greene, M. A. and Meissner, D. "More Product, Less Process: Revamping Traditional Archival Processing". In *The American Archivist* vol. 68, issue 2, 2005, pp. 208–263.

Nelson-Strauss, B. *et al. The Library of Congress National Recording Preservation Plan*. Washington, DC: Library of Congress, 2012.

Pond, M. Personal correspondence. February 2012.

Smith, A. *et al. Survey of the State of Audio Collections in Academic Libraries*. Washington, DC: Council on Library and Information Resources, 2004.