Teachable AI for the Archival Professions – Module 5: **AI/ML for Processing Audiovisual Records in Archives**

Nataliya Radke[1] and Richard Arias-Hernandez[2]

This educational module is part of a series of learning materials developed by InterPARES Trust AI[3] researchers and educators to train archival professionals and students to effectively leverage artificial intelligence in their archival work. The current version was completed on March 27th, 2026.

This learning module has a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International BY-NC-SA 4.0 license, which requires that reusers give credit to the creators. It allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, for noncommercial purposes only. If others modify or adapt the material, they must license the modified material under identical terms.[4]

[1] InterPARES Trust AI Graduate Academic Assistant, University of British Columbia.

[2] Associate Professor of Teaching, School of Information, University of British Columbia and InterPARES AI Trust Researcher. richard.arias@ubc.ca

[3] This work is an outcome of InterPARES Trust AI, an international research partnership led by Drs. Luciana Duranti and Muhammad Abdul-Mageed, University of British Columbia, and funded by the Social Sciences and Humanities Research Council of Canada (SSHRC). https://interparestrustai.org/

[4] Teachable AI for the Archival Professions – Module 5: AI/ML for Processing Audiovisual Records in Archives © 2025 by Radke, Nataliya and Arias-Hernandez, Richard is licensed under CC BY-NC-SA 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by-nc-sa/4.0/

Module 5: AI/ML for Processing Audiovisual Records in Archives



READ FOR THIS MODULE

Required:

- Anderson, D. & Rowe, S. (2025). What 400 Hours of AI Transcription Taught the Wildenstein Plattner Institute. Archival *Outlook, 6-7. https://mydigitalpublication.com/publication/?i=858398*

- Magnus, B., Priem, M., Vanderperren, N., Berghe, P. V., Keer, E. V., & Vissers, R. (2025). Metadata creation and enrichment using artificial intelligence at meemoo. Journal of Digital Media Management (London), 13(2), 110-123. https://doi.org/10.69554/NGFF5280

- Sullivan, P. and Sengsavang, E. (2024). UNESCO Audio Archives: AI for Metadata Enrichment. In: Duranti, L. and Rogers, C. (Eds.). (2024). Artificial intelligence and documentary heritage. SCEaR newsletter, Special issue 2024. pp. 27-33. PURL: https://unesdoc.unesco.org/ark:/48223/pf0000389844

Recommended:

- Cameron, S., Franks, P., & Hamidzadeh, B. (2023). Positioning Paradata: A Conceptual Frame for AI Processual Documentation in Archives and Recordkeeping Contexts. *Journal on Computing and Cultural Heritage, 16*(4), 1-19. https://doi.org/10.1145/3594728

- Colavizza, G. & Jaillant, L. (2026). AI Preparedness Guidelines for Archivists. Archives & Records Association (UK & Ireland). https://static1.squarespace.com/static/60773266d31a1f2f300e02ef/t/69789fd03543a1698d645315/1769512912257/AI-Preparedness-Guidelines_February_2026.pdf

- Gupta, V. & Boulianne, G. (2020). Automatic Transcription Challenges for Inuktitut, a Low-Resource Polysynthetic Language. *Proceedings of the 12th Conference on Language Resources and Evaluation*, 2521-2527. https://aclanthology.org/2020.lrec-1.307.pdf

- Jansen, A. and Cruz, M. (2024). iTrust AI Public Symposium. Using AI To Interrogate Archives - Enhancing Access to Video Using Lox Code/No Code AI Services (Conference Presentation). iTrust AI Public Symposium, Honolulu, Hawai'i. February 23, 2024. https://www.youtube.com/watch?v=hJjifghl6AQ&t=7074s&ab_channel=Hawai%CA%BBiStateArchives

- Koenecke, A., Nam, A., Lake, E., Nudell, J., Quarry, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 117(14), 7684-7689. doi: 10.1073/pnas.1915768117.



OVERVIEW

This module provides an overview of the past and ongoing use cases of artificial intelligence (AI) and machine learning (ML) tools in archival institutions for the processing of audiovisual records. This module discusses the history of tools such as automatic speech recognition (ASR), how ASR works, how ASR is being used in archival contexts today, as well as the challenges and limitations of ASR regarding privacy, copyright, and cost. The module looks at how AI tools are being used for processing video records in archival and records management contexts while also highlighting challenges that are unique to video records. Finally, it also outlines effective strategies to implement AI tools for the processing of audiovisual records, such as adopting a human-in-the-loop approach and identifying project-specific use cases.



LEARNING OBJECTIVES

By the end of this lesson, students will be able to:

- Explain current and potential uses of Artificial Intelligence (AI) and Machine Learning (ML) for the processing of audiovisual records
- Use AI/ML tools for processing of audiovisual records
- Critically analyse the challenges of current AI/ML applications for processing of audiovisual records (e.g., criminal/illegal, copyrights or individual rights, biases, commercialization, etc.)

History and Applications of ASR Technology

A Brief History of AI for Audio/Speech Processing

Automated Speech Recognition (ASR) technology began in 1952 when researchers at Bell Laboratories developed the first machine capable of recognizing speech. The Automatic Digit Recognition machine — nicknamed 'Audrey' — could recognize any number between zero and nine to facilitate voice dialling (Moskvitch, 2017). When listening to its creators, Audrey had an accuracy rate of 70-90%, although it relied on a room full of highly specialized and costly equipment to do so (Moskvitch, 2017). Still, Audrey was a precursor for what was to come, as ten years later, IBM released 'Shoebox' (IBM, n.d.). Named for its compact size, the machine was capable of recognizing nine numbers as well as six words to perform basic mathematical equations (IBM, n.d.). Around the same time, researchers in the Soviet Union invented the dynamic time warping algorithm to better handle variations in speaking speed, resulting in the first machine to have a 200-word vocabulary (Moskvitch, 2017). However, ASR would not be able to

recognize entire sentences until the US Department of Defense funded the five-year Speech Understanding Research program (SUR) between 1971-1976 (Moskvitch, 2017). In collaboration with IBM, Carnegie Mellon University (CMU), and Stanford Research Institute, SUR resulted in a machine that had a vocabulary of 1,011 words, which was named 'Harpy' (Moskvitch, 2017).

Harpy set the stage for the release of IBM's voice-activated typewriter in the mid-1980s, Tangora. Intended to help streamline office correspondence, Tangora boasted a vocabulary of 20,000 words thanks to advancements in digital signal processing and predictive technologies based on hidden Markov models (HMM) and n-gram language models (Juang & Rabiner, 2004).

Improvements in computer processors, pattern recognition technology, and artificial neural records in the 1980s made ASR technologies accessible to the broader public for the first time with interactive voice recognition systems such as phone tree systems, which are still in use today (Juang & Rabiner, 2004). In 1990, Dragon Systems released the first ever consumer speech recognition product for PCs, Dragon Dictate (Moskvitch, 2017). With a 5,000-word vocabulary, Dragon Dictate allowed users to control a PC using only voice commands and, as a result, found considerable success as an assistive technology for users with accessibility concerns (Dragon Medical Transcription, n.d.). For example, for users who could not manipulate a mouse or press keys, Dragon Dictate gave them access to most computer functions, such as navigating web pages and typing (Leib, n.d.). Individuals with visual impairments found similar success with the software as it was capable of "executing commands faster than screen reading and magnification programs" (Leib, n.d.). The application of ASR to assistive technologies would remain a driving force in its innovation and development throughout the 1990s and continues to this day (Hux et al., 2000).

However, alongside a hefty price tag of $9,000 USD (roughly $22,000 USD when adjusted for inflation today), Dragon Dictate also required users to

pause after each word to ensure accurate transcription, called discrete speech (Moskvitch, 2017). This resulted in slower transcription speeds and limited the wider market appeal of Dragon Dictate (Dragon Medical Transcription, n.d.). Nonetheless, the innovation and success of Dragon Dictate caught the attention of the US Department of Defense, who signed a contract with Dragon Systems in 1993 to develop continuous speech and voice recognition systems (Dragon Medical Transcription, n.d.). The result of this collaboration came in 1997 with Dragon NaturallySpeaking, the first continuous speech and voice recognition product with a vocabulary of 23,000 words (Dragon Medical Transcription, n.d.). No longer limited to discrete speech, Dragon NaturallySpeaking recognized speech at roughly 100 words per minute and, for the first time, made it practical to use voice and speech recognition for document creation (Moskvitch, 2017). Although Dragon NaturallySpeaking found widespread success among professionals, academics, and users with accessibility needs, it was particularly impactful for healthcare professionals who used it to dictate patient notes and streamline the otherwise time-consuming charting process (Parente, Kock, & Sonsini, 2004). In fact, the software is still in use today under the name of Dragon Medical and has become a critical component in the creation of electronic health records with over 550,000 users (Nuance, 2023).

Still, Dragon NaturallySpeaking and similar technologies of the 1990s — such as IBM's ViaVoice — had their limitations. For example, Dragon NaturallySpeaking was speaker-dependent, meaning it had to be trained to reliably and accurately recognize a speaker's voice (Juang & Rabiner, 2004). This made it ineffective for spontaneously transcribing multiple speakers, such as interviews, and limited its use to single-user transcription. Additionally, despite boasting accuracy rates as high as 90% and being able to accommodate different accents, dialects, and speech variations, Dragon NaturallySpeaking and IBM's ViaVoice struggled to reliably transcribe non-native English speakers (Coniam, 1999). This resulted in a bias towards native English speakers that dominated ASR technology throughout the

1990s and 2000s and even continues to this day. For example, contemporary ASR systems are disproportionately trained on English data sets, restricting their availability and efficacy for non-native speakers. Recent efforts are being made to make the technology to be more inclusive, which is discussed later in this module (Sullivan, Shibano, & Abdul-Mageed, 2023).

Following the success of the 1990s, the next major breakthrough in ASR technology came in 2008, when Google Voice Search was released as an app for the iPhone, allowing users to search the internet using voice commands (Elmore, 2025). By offloading processing requirements to datacentres, Google was able to implement complex deep neural networks and machine learning algorithms to continuously improve their ASR models from users' search data (Elmore, 2025). This integration of AI and ML with ASR paved the way for the introduction of voice-based virtual assistants like Apple's Siri in 2010 or Amazon's Alexa in 2014. With accuracy rates ranging from 93-95%, ASR technologies are now present in everyday life. For example, voice-activated navigation and search systems in cars are used to improve driver safety while healthcare professionals rely on dictation software like Dragon Medical to automate charting and maximize doctor-patient interactions (IBM, 2025).

History of ASR Technology in Archives

Although existing since the 1950s, ASR did not become more accessible to the wider public until the 2000s, limiting its use in archival contexts until that point despite having potential to expedite the time-consuming processing of audio and video records. This was due to a number of reasons, some of which were discussed earlier. For one, ASR technologies in the 1990s and 2000s were prohibitively expensive. With systems such as Dragon Dictate retailing for $22,000 US for a single user license, they simply were not a realistic option (Moskvitch, 2017). And while ASR technology became more accessible and affordable beginning in the 2000s, opportunities

for implementation were severely limited by accuracy rates of just 30-50% (Munteanu et al., 2006). Additionally, audio recordings with multiple or non-recurring speakers as well as non-native English speakers made it difficult to train ASR systems such as Dragon NaturallySpeaking to improve accuracy. Even today, these systems struggle to reliably transcribe spontaneous spoken speech and capture the emotional context (Gustman et al., n.d.). There was also the added challenge and process of needing to digitize audio recordings before applying ASR as born-digital recordings were not commonly received by archival institutions until the 21st century. Overall, these factors contributed to an environment where implementing and applying ASR technology often required more effort, time, and resources than manually processing and arranging and describing audio records.

Despite these early barriers to adopting ASR systems in archival institutions, as early as the 1990s, archivists, IT specialists, and archival vendors began investigating the application of ASR to assist in the processing of audiovisual materials (Whittaker et al., 1999). The focus on audiovisual materials was due to the extensive hands-on labour required to arrange, describe, and make them accessible, which continues to be a problem today (Byrne et al., 2004). As the following examples will illustrate, the focus was therefore on how ASR technologies could expedite the processing time of audiovisual materials while also improving discoverability and access.

One of the first documented cases of ASR technology for archival purposes was in 2001 when the National Science Foundation (NSF) awarded $7.5 million USD ($13.7 million USD when adjusted for inflation) to the Shoah Visual History Foundation (VHF) to develop multilingual speech recognition software for the processing of interviews of Holocaust survivors (Jackson, 2001). With 52,000 audio and video records in over 32 languages totalling some 180 TB or 116,000 hours of audio and video, the VHF had the "largest and most complex single topic digital video library in the world" (Gustman et al., n.d.). The NSF's funding resulted in the creation of the Multilingual Access to Large Spoken Archives, or MALACH. This was an

international collaboration focused on implementing ASR technology for spoken archives between the VHF at the University of Southern California, IBM, John Hopkins University, the University of Maryland, Charles University, the University of West Bohemia, and AITIA International (MALACH, n.d.-a). The objective of MALACH was to "dramatically improve access to large multilingual spoken archives" by developing and implementing ASR technology to "handle spontaneous and emotional speech with disfluencies, heavy accents, elderly speech, and dynamic switching between multiple languages" (Gustman et al., n.d.). All of which were major barriers that had prevented the application of ASR systems to archives until that point. MALACH sought to utilize advances in ASR technology to assist in metadata creation and segmentation, automate translation of domain-specific multilingual thesauri, and to evaluate the "social and scientific value of [ASR technology] to see how it can be applied to other large archives" (Gustman et al., n.d). Between 2002 and 2006, MALACH had 67 interdisciplinary publications, workshops, and panel discussions on the application of ASR technology in archival institutions, including how to train and implement systems and workflows (MALACH, n.d.-b).

Byrne et al. wrote in greater detail on the scale of the VHF's collection. Before the start of the MALACH project, after 150,000 workhours, only 9% of the oral history collection had been manually processed and annotated with transcriptions of speaker names, locations, dates of creation, and summaries (Byrne et al., 2004). With human transcribers taking 8-12 hours to transcribe a single hour of an English interview, the authors sought to discover whether ASR technologies could be implemented to reduce processing times and streamline archival workflows for recorded interviews of Holocaust survivors in a variety of languages (Byrne et al., 2004). As the interviews were originally recorded on Sony Beta SP tapes, the authors began by digitizing the tapes into a 3 MB/s MPEG-1 stream with a 128 kb/s (44 kHz) stereo audio (Byrne et al., 2004). The corpus or training data for the ASR model was then "generated using 15-min segments of an interview from 800

randomly selected speakers" that was manually transcribed to ensure accurate training data (Byrne et al., p. 422). Initial performance results of transcribed audio of interviewees, measured by word error rates (WER), ranged from 57.3% to 53.1%, although additional training via manual transcription and consensus decoding reduced the WER to as low as 39.6% (Byrne et al., 2004). Using ASR, the VHF was able to successfully process 500 hours of English records at this early stage of the MALACH project and make them accessible via metadata enrichment based on the generated transcripts. Given the time and resources required to train and implement the ASR system, the authors noted that while implementing a similar system in other archives would not be feasible, it was an exciting proof of concept that revealed a number of future applications for the MALACH project.

By 2006, the MALACH project was a success in using AI to support the processing and description of audio and video recordings, a task once considered impossible prior to the implementation of ASR technology. The project also resulted in the MALACH Interviews and Transcripts English corpus that can be used to train other ASR systems (Ramabhadran et al., 2012). The training dataset is made up of 375 hours of testimonies featuring some of the most complex aspects of ASR, such as accented and emotional speech, non-English named entities, uncued language switching, and multiple speakers (Picheny et al., 2019). In 2019, the system had a WER of roughly 21.7%, although further development in 2023 using contemporary AI and ML systems, such as OpenAI's Whisper, reduced the WER further to 13.5% (Picheny at al., 2023). As the MALACH corpus is a dataset that can be used in archives for the training of ASR models to support the processing of audio records, there is still ongoing development to improve the dataset today, proving the longevity and impact of the MALACH project for the use of ASR systems in archives (Picheny at al., 2023).

Regarding other developments, in 2006, Munteanu et al. began what would become a years-long investigation into the application of ASR technology to transcribe webcast lectures at the University of Toronto to

improve access and discoverability. They started by measuring how the quality of transcripts affects user experience (Munteanu et al., 2006). The authors found that speech recognition accuracy linearly influences both user performance and experience, that 45% WER is unsatisfactory, and that transcripts with a WER of 25% or less is an acceptable error threshold from the perspective of the user (Munteanu et al., 2006). A year later, Munteanu et al. (2007) found that ASR accuracy could be improved by combining them with large vocabulary Language Models (LMs) based on corpora from the internet as well as telephone conversations. This resulted in a reduction in the WER by relatively 11%, proving the viability of combining LMs and ASR technology as early as 2007 (Munteanu et al., 2007). By the end of their work, Munteanu et al. (2008) concluded that ASR systems, when combined with LMs and enhanced with different corpora, were an effective way to transcribe webcasts to enhance access and discoverability.

Writing from the perspective of archival technology vendors, Kummer and Backfried (2007) demonstrated the viability of modifying and integrating off-the-shelf ASR products with existing archival infrastructure and systems to assist in the processing of audiovisual materials (Figure 1). Using the Media Mining System (MMI) from SAIL LABS Technology, they found it was possible to utilize ASR technology to automate content transcription and speaker and keyword identification with roughly 80% accuracy to reduce archival processing times (Kummer and Backfried, 2007). However, Kummer and Backfried also found that their ASR tools struggled to accurately recognize spoken language when the audio had a poor quality or when it contained dialects that were different from their training datasets (Kummer and Backfried, 2007). As a result, the authors emphasized that ASR technologies were best utilized as tools to streamline the earlier, more tedious stages of processing audiovisual materials to empower archivists to more strategically utilize their time (2007).

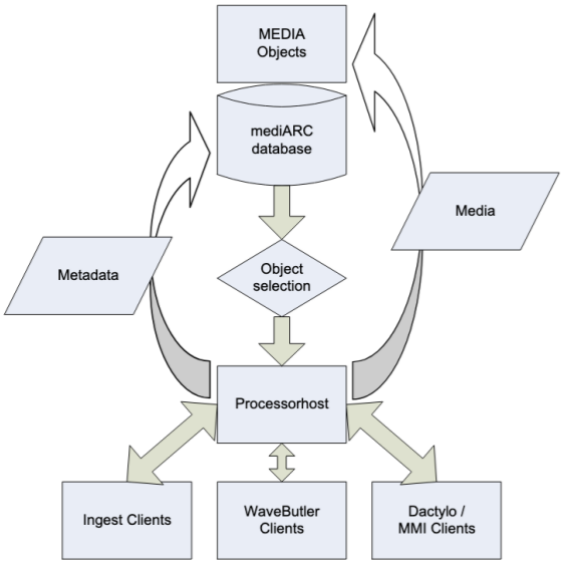


Figure 1. A framework for how Kummer and Backfried implemented ASR into an existing system (Kummer and Backfried, 2007)

Information technology specialists at Ryukoku University in Japan considered the role of ASR in the creation and maintenance of records such as meeting minutes from the National Congress of Japan (Nanjo et al., 2006). They designed a computer assisted speech transcription (CAST) system with an interface that enabled users to easily correct ASR errors in real-time, reducing the transcription time to about half of the time (Nanjo et al., 2006). Their language model was trained on meeting minutes from the National Congress that totalled over 86.7 million words. However, because these minutes did not contain spoken expressions such as filler words, the authors also prepared a true transcription of meetings with an additional 353,000 words to ensure a more accurate starting point. When combined with a spontaneous speech corpus of 228 hours, the CAST system had a vocabulary size of 52,093 words (Nanjo et al., 2006). The work of Nanjo et al. (2006) demonstrates the innovation taking place at the time to adapt ASR

technologies to archival and records management contexts where off-the-shelf ASR products were not yet available.



ACTIVITY #1

ASR technologies have existed for decades with a number of proven use cases and archival-based applications, although their accuracy and usefulness may vary greatly.

In small groups (2-4), have students test out how well speech-to-text tools work in different environments (e.g., multiple speakers, high background noise, different languages, different styles of speech such as fast speech, emotional speech, genre-specific or domain-specific speech, etc.). Free speech-to-text tools include Apple iPhone Dictation and Google's Voice Search or apps such as MacWhisper or Aiko. Afterwards, have students discuss how the tools worked, whether some were better than others, any issues they encountered, and what situations they think ASR could be useful in.

How Does ASR Work?

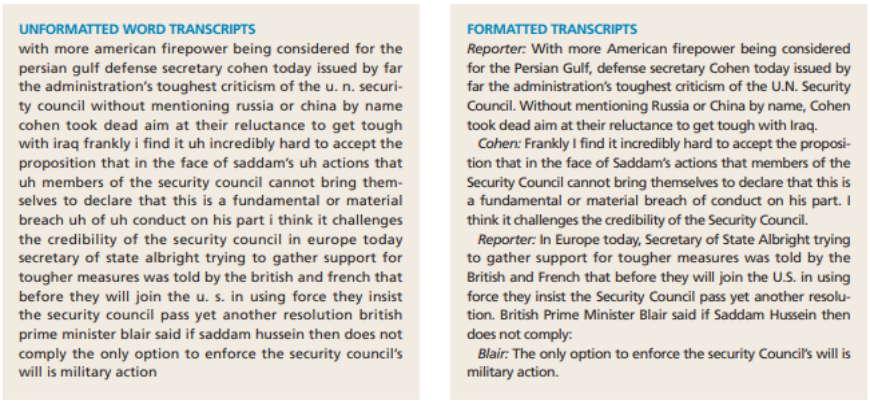
Before ASR technologies can be applied to the processing of audiovisual records, the materials in question must first be adequately prepared. In the case of an analogue record such as an audio cassette tape, it would first need to be digitized into an audio file with key properties that an ASR tool can recognize. This means that the quality of audio samples should be a priority to ensure accurate ASR outputs. Key audio priorities include a high sample rate of at least 16kHz, bit depths of 16 or 24, mono audio channels, and non-lossy file formats such as WAV or FLAC (Loeber, 2025). For purposes of digital preservation, IASA recommends a frequency sampling of at least 48kHz and up to 96 kHz and at least 24-bit depth for the digitization of analogue audio (IASA, 2009). Pre-processing techniques such

as noise reduction, normalization, and echo cancellation also occur at this stage to improve the clarity and consistency of an audio recording to ensure more accurate outputs (Harris, 2024). However, quality often comes at the price of increased processing, storage demands. A delicate balance must be found between best practices and an institution's resources. The size of the files matter as well. Several ASR tools impose restrictions on the size of the processed files (e.g., 25MB, 50MB, 100MB, 1GB, etc. depending on the tool and whether files are uploaded to the cloud or stay local and use an API service). This means that archival institutions using these types of tool will have to either compress the audio files or split them to file sizes under the imposed limit. An alternative is to use open-source versions of ASR tools, such as Whisper, locally or on the archives' own infrastructure to remove these file size limits. At this point the digital audio file is ready to enter an AI-powered ASR processing pipeline.

This pipeline has been streamlined during the last ten years. Between 2011 and 2019, ASR significantly improved with AI relying on deep learning techniques by using hidden Markov models in combination with Deep Neural Networks (HMM-DNN). In this AI end-to-end approach, an expert in linguistics was required to annotate and create a lexicon and phonetic equivalents for the model and the deep neural network would take care of the language models over multiple steps to reliably transcribe speech to a written transcript (Foster, 2025). This was the original approach behind Siri and Alexa. An issue with the HMM-DDN approach described here is that archives and other organizations have differential access to the human expertise required to create these models. This constituted a hurdle for most archives attempting to create in-house ASR models. However, this changed by the end of 2019: state-of-the-art ASRs completely switched to pure Deep Learning/Deep Neural Networks surpassing ASRs using HMM-DNN and reaching accuracy rates of above 95%, making the need to develop acoustics, lexicon, and linguistic models unnecessary and streamlining the

ASR pipeline. Issues with hallucinations still remain, though, especially when processing audio in low-resource languages (Timmel et al., 2024).

The next step in the processing pipeline is speech segmentation. A speech segmentation task divides the continuous audio stream into smaller units such as individual phonemes, syllables, words, or sentences to improve ASR accuracy. Segmentation is important in the context of audio records because, unlike text, it lacks the natural clues such as punctuation, capitalization, and formatting that traditionally convey key contextual information (Ostendorf et al., 2008). For example, segmentation can identify points in audio where a speaker or audio type changes, such as when someone stops speaking and only silence is heard. Human listeners often do this automatically by drawing on "sophisticated syntactic, semantic, acoustic, prosodic, pragmatic, and discourse knowledge" to arrive at something like the formatted transcript in Figure 4 (Ostendorf et al., 2008). A machine, however, does not. It instead produces an output of a stream of words without any contextual boundaries as shown in the unformatted transcript in Figure 4 such as a comma to indicate a natural pause in a spoken sentence. Therefore, segmentation is a vital component in increasing readability of outputs.



UNFORMATTED WORD TRANSCRIPTS

with more american firepower being considered for the persian gulf defense secretary cohen today issued by far the administration's toughest criticism of the u. n. security council without mentioning russia or china by name cohen took dead aim at their reluctance to get tough with iraq frankly i find it uh incredibly hard to accept the proposition that in the face of saddam's uh actions that uh members of the security council cannot bring themselves to declare that this is a fundamental or material breach uh of uh conduct on his part i think it challenges the credibility of the security council in europe today secretary of state albright trying to gather support for tougher measures was told by the british and french that before they will join the u. s. in using force they insist the security council pass yet another resolution british prime minister blair said if saddam hussein then does not comply the only option to enforce the security council's will is military action

FORMATTED TRANSCRIPTS

Reporter: With more American firepower being considered for the Persian Gulf, defense secretary Cohen today issued by far the administration's toughest criticism of the U.N. Security Council. Without mentioning Russia or China by name, Cohen took dead aim at their reluctance to get tough with Iraq.

Cohen: Frankly I find it incredibly hard to accept the proposition that in the face of Saddam's actions that members of the Security Council cannot bring themselves to declare that this is a fundamental or material breach of conduct on his part. I think it challenges the credibility of the Security Council.

Reporter: In Europe today, Secretary of State Albright trying to gather support for tougher measures was told by the British and French that before they will join the U.S. in using force they insist the Security Council pass yet another resolution. British Prime Minister Blair said if Saddam Hussein then does not comply:

Blair: The only option to enforce the security Council's will is military action.

Figure 2. An example of how segmentation can help preserve key contextual and structural clues in speech (Ostendorf et al., 2008). Note inclusion of speakers, punctuation, and capitalization on the right.

Speaker diarization (SD) can further segment audio files by further partitioning audio files by speaker. In other words, SD answers the question of *who* spoke and *when* using segmentation and labelling (Figure 5). SD helps ASR programs distinguish individuals from one another during conversations, such as during recorded interviews with multiple speakers (IBM, n.d.). There are two main approaches to SD: supervised and unsupervised. The former is trained to recognize a set number of speakers and thus only works with speakers that appeared in the training data, increasing its overall accuracy but limiting its versatility without specific training and additional resources (La Javaness, 2023). In comparison, the latter uses a model to cluster "audio segments according to the speaker based on extracted audio features", making it capable of identifying a wider range of speakers it has no prior knowledge or training of (La Javaness, 2023). With recent advancements in AI, most SD models are unsupervised.

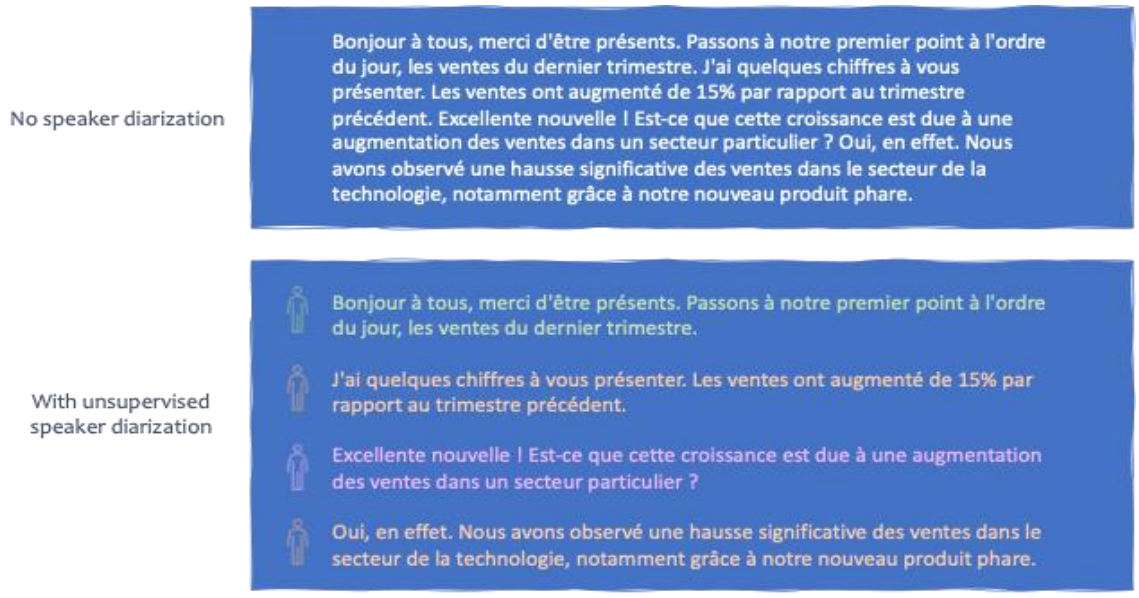


Figure 3. An illustration highlighting the importance of speaker diarization for ASR systems to identify one or more speakers in an audio recording (La Javaness, 2023).

Another key component of the AI-powered ASR pipeline is speaker recognition (SR), which is "the process of identifying or verifying a speaker using their voice" (Hansen, 2024). While ASR focuses on transcribing spoken words to text and determining *what* is being said, SR looks at *how* it is being said to determine *who* the speaker is by analyzing the "unique characteristics of a person's speech, such as pitch, tone, and speaking style" (Hansen, 2024). This is sometimes referred to as a voice or audio fingerprint, which can be compared to a dataset or training model to predict the most likely speaker based on how well they match existing voice patterns in the system (Hansen, 2024). Speaker verification, as the name implies, can then be applied to ensure that speakers have been correctly identified by reviewing extracted audio features (i.e., embeddings) and comparing them to the original references to return the most likely speaker. SR and speaker verification offer a unique opportunity to expedite the processing and description of audiovisual materials. Additionally, when combined with recent advancements in large language model (LLM) technologies, it is now possible for SR and speaker verification tools to produce structured multilingual outputs previously thought impossible. Meng et al. (2024) found that their model, Multi-Talker LLM or MT-LLM, can simultaneously transcribe multiple speakers into text, transcribe a specific target speaker when given a reference audio clip, and transcribe speech based on a speaker's sex, occurrence order, or when they say a keyword or speak in a specific language. LLMs can also make probabilistic inferences about the identity of the speakers identified in the audio records based on the content of the audio record itself. For example, if the content of the transcribed audio record includes personal identifiable information (i.e., the speaker introduces themselves at the beginning of the audio providing their name and affiliation), this information is then used by the LLM to identify the speaker in the associated audio segments with a level of certainty (Abdul-Mageed, n.d.) LLMs also allow users to interact with transcripts and return contextual

information such as the main topics that were discussed, speakers, or summaries as demonstrated in Figure 3. (Abdul-Mageed, n.d.).

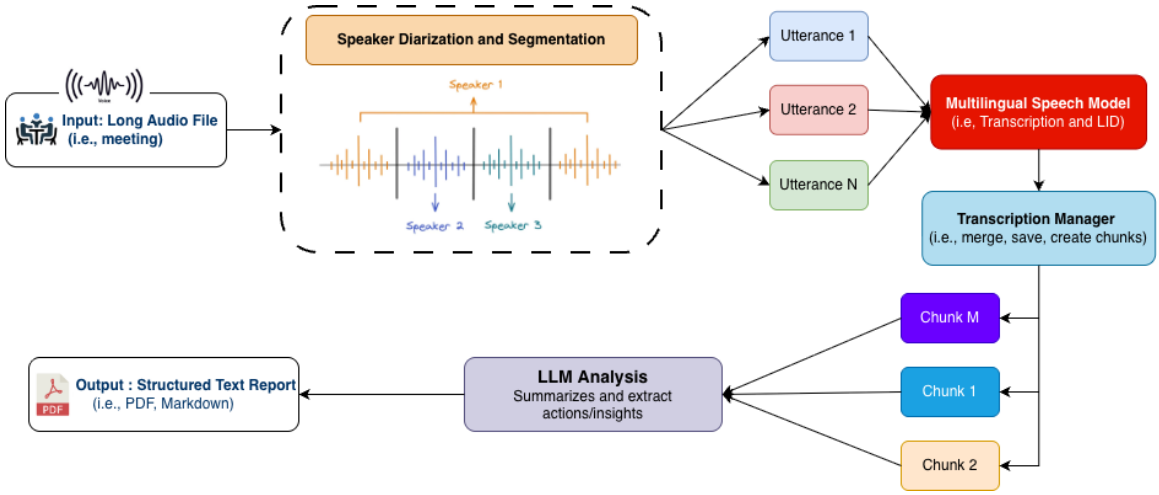


Figure 4. An InterPARES AI pipeline for audio records, which demonstrates how LLMs can be combined with ASR technologies to produce complex and versatile outputs based on a user's goals and needs (Abdul-Mageed, n.d.).

Finally, Human-in-the-Loop (HITL) post-processing techniques are then used to correct mis-recognitions and other errors in decoded transcriptions to improve accuracy. HITL refers to a workflow or process where a human is actively involved in the operation, supervision, or decision-making of an automated system or AI technology to "ensure accuracy, safety, accountability or ethical decision-making" (Stryker, n.d.). In the context of automated post-processing tools for audio records, a HITL approach is crucial to ensure that any changes to a file do not impact its authenticity, trustworthiness, or quality. Moreover, it provides the opportunity to explicitly include paradata about transformations on the digitized audio records during any step of the AI pipeline in the processing of such records (Cameron et. al., 2023). Once an output is generated, such as a transcript from an interview in the form of an XML, SRT, VTT, or JSON file, that data must also be synchronized with the original audio file to ensure

alignment. In other words, synchronization connects what was said and when it was said in the audio file, often in the form of time stamps. It is a crucial component in being able to verify the accuracy of ASR tools, training new models, and performing additional tasks such as keyword or speaker identification (Navon et al., 2023).



ACTIVITY #2a: ASR in a digital audio record using AssemblyAI

In this activity, students will transcribe an audio file with AssemblyAI's free speech-to-text tool to see how features like speaker diarization and speech recognition work in real-time.

1. Have students make a free account with AssemblyAI.
2. Listen to a 5-mins, digital-audio file extracted from a 1-hour-, digitized-, public-, audio record from the Archives of the University of British Columbia.
3. Download this file to your computer and then upload it to AssemblyAI. Process the file and review the transcript. Optionally, you could use another digital audio file for this activity. However, note that this is a cloud service, so you should not be using audio that infringes copyrights or privacy laws.
4. As a group, have students discuss their experience. Consider how having a synchronized transcript with speaker diarization could expedite the processing of audio records, including metadata capture. Additionally, have students compare the transcript with the raw output found in the 'API Response' and consider their advantages and disadvantages.

ACTIVITY #2b: ASR in a digital audio record using OpenAI's Whisper

In this activity, students will transcribe an audio file with OpenAI's Whisper to experience how speech recognition works in real-time (note that Whisper alone does not include diarization). Students will need a Google Account for this activity.

5. The instructor walks through Python code using Whisper on a Google Colab notebook and explains each section of the code and how it works. Students save a copy of this notebook in their Google Drive and follow the instructor step-by-step executing code in their own copy of the Google Colab notebook.
6. Use the same digital audio file used in Activity 2a to be processed by Whisper.
7. As a group, discuss the differences in the experience and output provided by AssemblyAI in Activity 2a and by Whisper in this activity. Consider as well the possibilities and conveniences of using Python code and Whisper as a Python Library to run speech recognition locally (without accessing cloud services) and maintaining the digital audio files being processed locally as well (without being sent to the cloud).

Current and Emerging Processing of Audio Records with AI/ML Tools

The application of AI/ML tools poses unique opportunities for audio materials in archives, in large part because these records often entail time-consuming workflows to arrange, describe, and process. For example, in 2012, the Canadian National Archival Development Program (NADP) stated that the appraisal and selection, arrangement, description, and physical

processing of sound recordings should — on average — take four times as long as the running time of the recordings (NADP, 2012). As archival institutions continue to struggle with increasing backlogs and decreasing resources, devoting that kind of time to a single object is often impractical, if not outright unfeasible. This was the case at the UNESCO Radio Archives. When Sullivan and Sengsavang (2024) started a project to describe UNESCO's collection of over 17,000 reel-to-reel tapes, only 800 had been described at the item level with key metadata such as speaker and language identification. The disparity in described items stemmed from how time-consuming and laborious describing audio records at the item level is. Additionally, staff would need extensive knowledge of foreign languages UNESCO's history and key figures to reliably identify speakers and languages (Sullivan and Sengsavang, 2024). Still, given the fact that these tapes covered a "remarkable range of topics, personalities, geographies, languages, and genres across four decades and in over 70 languages…reflecting the substance of UNESCO's work and its international, intergovernmental character," expediting the description and metadata enrichment process to make them more accessible to users was of an utmost priority (Sullivan and Sengsavang, 2024, p. 27).

This presented a unique opportunity to incorporate AI/ML ASR technology with the traditional archival method of diplomatic analysis as a means to anchor the authors' approach and provide more control over the application of AI tools (Sullivan and Sengsavang, 2024). As such, the authors used an archival diplomatics-informed methodology to "identify patterns in the underlying structures of similar types of recordings…explicitly labelling the important structural parts, [and] thus simplifying the AI problem to be solved" (Sullivan and Sengsavang, 2024, p. 28). With a strong foundation to base their work upon and ensure record authenticity, the authors used OpenAI's Whisper to automatically identify languages, speakers, and generate transcripts from the audio files (Sullivan and Sengsavang, 2024). Even with accented speech, the authors found that Whisper had a 94%

accuracy rate and was a success in streamlining the capture of that key metadata element (Sullivan and Sengsavang, 2024). They had less success with consistent and accurate speaker recognition, in part due to linguistic, age, and gender biases that require in-house model training. The project continues today with plans to improve speaker recognition tools and transcript accuracy that can eventually be used for summarization experiments.

The Computational Linguistics Application for Multimedia Services (CLAMS) project at Brandeis University in Massachusetts is also working to develop and implement ML-based tools to optimize the processing of audiovisual archive material and metadata generation to improve access, search, and exploration of archival audiovisual material (Rim et al., 2025). With a focus on interoperability and open source tools that are free to use and customize, tools such as the inaSpeechSegmenter are now available for integration into processing environments with future developments in progress (CLAMS, n.d.).

Meanwhile, Anderson and Rowe (2025) at the Wildenstein Plattner Institute used Amazon Transcribe, a speech recognition service, to transcribe more than 400 hours of audio and video from the Romare Bearden collection. While being resource-conscious was the primary driving factor in Anderson and Rowe's decision to use an AI tool like Amazon Transcribe, the authors also prioritized access. They argue that "transcripts are an essential component of web accessibility" that offer "equitable access to individuals [with] difficulty processing audio or visual materials" (Anderson and Rowe, 2025, p. 6). However, while Amazon Transcribe quickly generated transcripts of material, Anderson and Rowe (2025) noted that the outputs were not accurate enough to guarantee true accessibility, in part due to varied audio quality, incorrect labelling of speakers, inaccuracy with underrepresented accents, and an inability to identify foreign languages. This resulted in an additional several months of work to correct the transcripts and complete the processing according to their project goals and standards. Despite these

challenges, Anderson and Rowe (2025) considered the project to be a success and a proof of concept that AI tools like Amazon Transcribe can play a valuable role in archival processing with correct planning and oversight.

In Canada, AI tools are being used to assist in the processing of audio records to preserve and promote Indigenous languages. The Computer Research Institute of Montreal, the Canadian Broadcasting Corporation (CBC), and the Piruvik Centre are developing "language labelling and speech segmentation tools for recordings of Indigenous languages to support more efficient annotation and ultimately, enable automatic speech recognition" (National Research Council Canada, 2025). The goals of this project are to expedite metadata enrichment to make it easier to access recordings of Indigenous languages being spoken, perform speech segmentation for easier data annotation, and to determine the viability of ASR for languages such as Inuktitut, East Cree, Innu, and Dénésuline (National Research Council Canada, 2025). Because Indigenous languages are considered low-resource, meaning there are either no or very few existing models and data sets to start from, CRIM and its collaborators sourced thousands of hours of spoken Indigenous audio files to develop AI models in-house. The CBC, for instance, provided access to over 1,343 hours of radio programming in East James Bay Cree (National Research Council Canada, 2025) while the Inuktitut models were trained on parliament proceedings and recordings of oral stories (Gupta and Boulianne, 2020). CRIM continues to develop tools for voice activity detection, speaker retrieval, speaker diarization, and language labelling to enable increased searchability and discovery (National Research Council Canada, 2025). To meet the needs of Indigenous communities, linguists, and researchers, CRIM notes that the tools will be made publicly accessible with VESTA and ELAN, a collaborative work platform for research software and an open source software for annotating oral recordings, respectively (National Research Council Canada, 2025).

AI tools and ASR technologies are similarly being used to preserve Indigenous languages in Mexico. Deance, Hernández, and Varela (BUAP)

(2024) are currently working on a project to preserve the oral and sound heritage of the Nahuatl and Totonac speaking peoples in the Sierra del Norte de Puebla, Mexico. Hernández is associated with the Meritorious Autonomous University of Puebla while Deance and Varela are from the Benemérita Universidad Autónoma de Puebla and are native speakers of Totonac and Nahuatl, respectively. Using Amazon Transcribe, the authors are currently digitizing and transcribing 46 reel-magnetic tapes with ethnographic recordings of oral testimonies, prayers, and music (Deance et al., 2024). The tapes were originally recorded by French researchers in the 1960s and feature a mix of Spanish, French, Nahuatl, and Totonac, making them an excellent candidate for ASR technologies such as speaker and language identification to expedite the otherwise time-consuming process of manual translation and transcription (Deance et al., 2024). Unfortunately, the accuracy and quality of the resulting transcriptions has been inconsistent. It has become evident to the researchers that while Amazon Transcribe currently supports the automated identification of over 80 languages, such as Spanish and French in these recordings, the tool has proven inefficient when encountering low-resource languages such as Totonac and Nahuatl (Deance et al., 2024). While this has naturally limited the application of ASR technologies to this project, the authors nonetheless emphasize how much time Amazon Transcribe has saved them by automatically identifying and translating the French and Spanish segments or timestamping segments with no detected speech, such as music or silence (Deance et al., 2024). Therefore, instead of spending hours manually translating, transcribing, and timestamping audio recordings, the team is able to focus their time and energy only on the spoken Nahuatl and Totonac recordings or segments (Deance et al, 2024). The lack of ASR tools for low-resource languages has also inspired the authors to start development on an open-source model capable of recognizing, identifying, and verifying spoken Nahuatl and Totonac based on these preliminary findings (Deance et al., 2024). Overall, the work of Deance et al. (2024) highlights the multifaceted role that AI and ASR

technologies can have in processing audiovisual records, even if Amazon Transcribe did not fully meet their project needs.

Finally, AI tools for audio files have also been instrumental in records management. At the Dutch Institute of Image and Sound, content is not manually analyzed by humans at the intake and evaluation stages of the records lifecycle unless it is "really strictly necessary to do so" (Sanabria Medina and Rodríguez Reséndiz, 2023, p. 80). This streamlines the appraisal and review process of active records which, in the context of today's ever-increasing record creation, offers invaluable tools for records managers. Moreover, in terms of storing records, AI tools for and algorithms for cataloguing and extraction of keywords from an audio file's transcript "are essential to guarantee its subsequent recovery" and access, tasks that become "slower and more complex [with digital files]" (Sanabria Medina and Rodríguez Reséndiz, 2023, p. 81). Indeed, the efficient cataloguing of records according to series and retention schedules are an essential yet time consuming component of active records management, and AI tools that expedite the process are becoming more necessary with increasing record creation.

Challenges and Constraints Using AI/ML Tools for Audio Records

While the opportunities of ASR and AI/ML tools for processing audio records have been outlined, it is equally as important to consider the challenges they present as well. As discussed already, utilizing ASR technology requires having a digitized audio file, thereby creating additional processes when working with analogue audio records. Most of the historical audio records in archives have not been digitized and they use a wide variety of media in different formats (e.g., reels of different sizes, analogue cassettes, etc.) which demands different digitization approaches, play-back and digitization equipment, and may require specialized vendors. This process of digitization itself is time consuming and requires financial resources whether it is conducted in-house or outsourced to a vendor. Thus,

digitization of analogue audio records is a considerable challenge for archives in itself. The condition of the analogue audio records themselves may present yet even more challenges as analogue audio records are at risk of physical degradation. For example, magnetic media such as tapes or reels may be damaged, torn, or re-recorded over, resulting in distorted or overlapping speech that will negatively affect ASR accuracy on digitized files.

Additionally, because ASR tools often rely on pre-existing models or data sets, one must consider the potential biases in that data and how they may affect the output. This is, essentially, the concept of 'garbage in, garbage out', where the quality of the output is determined by the quality of the input. This is especially true in previously mentioned cases of metadata enrichment, where for "AI-generated metadata to accurately represent its inputs, data quality is a critical concern, since it predetermines the results of metadata generation" (Thomas and Dineen, 2026, p. 11). Corrupted or low-quality audio files could, for example, incorrectly tag segments as being speech when they contain music, "rendering [the] data invisible to users searching for the correct […] data type" (Thomas and Dineen, 2026, p. 11). This highlights the importance of a Human-in-the-Loop approach and pre-processing techniques to improve audio quality.

In the context of ASR and AI/ML tools, bias can manifest in lower accuracy rates depending on gender and race. Koenecke et al. (2020) tested the ability of five ASR systems by Amazon, Google, Apple, IBM, and Microsoft to transcribe 19.8 hours of structured interviews with 42 white speakers and 73 black speakers. The authors found that "all five ASR systems exhibited substantial racial disparities, with an average word error rate (WER) of 35% for black speakers compared with 19% for white speakers" (Koenecke et al., 2020, p. 7684). The implications of this disparity are significant, as it suggests that "it is considerably harder for African Americans to benefit from the increasingly widespread use of speech recognition technology" and may "actively harm [them] when for example, speech recognition software is used by employers to automatically evaluate candidate interviews or by criminal

justice agencies" (Koenecke et al., 2020, p. 7688). Harris et al. (2024) had similar findings when comparing the performance of ASR systems such as Whisper and Wav2vec2 with different dialects of American English. The authors found that Standard American English (SAE) had significantly lower WERs on every system when compared to minority dialect speakers of African American Vernacular English (AAVE), Chicano English, and Spanglish (Harris et al., 2024). Within this documented racial and language bias, Harris et al. (2024) also found that there were disparities depending on gender. Within minority dialect groups, women outperformed men in terms of accuracy, although within SAE, men outperformed women across every system (Harris et al., 2024).

These biases are not limited to gender or dialect, either. Because ASR systems often rely on existing models or datasets for innovation and improvement, efficacy and accuracy are thus dependent on the availability of resources and research. Languages such as English, German, or Mandarin Chinese are considered high-resource because there is an abundance of well-documented research, literature, and available datasets to fine-tune systems for improved accuracy (Amodei et al., 2016). In turn, for languages that are 'low-resource' and do not have robust and existing datasets and research available, they do not have the same access to ASR systems and tools and the opportunities and privileges they may afford. For example, this includes languages such as Finnish, West-Frisian, Malayo-Polynesian (Bartelds et al., 2023) and Indigenous languages such as Inuktitut and Cree (National Research Council Canada, 2025) or Nahuatl and Totonac (Deance et al., 2024). In an assessment of ChatGPT's performance with language identification, Abdul-Mageed (2024) found it "[fell] short of serving the wide and diverse linguistic needs of global communities in their languages", with African languages receiving the least support (p. 10).

In terms of rectifying the issue of language, gender, and racial biases, Koenecke et al. (2020) and Harris et al. (2019) argued that the disparities of ASR accuracy based on gender, race, or dialect stemmed from biases

inherent in the ASR models themselves. While the specifics of how mainstream ASR models from the likes of Apple, Google, or Amazon are trained remain unknown, the disparities in accuracy may reveal gaps in their modelling and datasets that appear to leverage one type of speaker over another (Harris et al., 2024). While this makes solving the issue challenging, Bartelds et al. (2023) found success in training an existing ASR system with in-house data to improve accuracy for low-resource languages, although this may not be feasible for all archival institutions due to resource constraints. In the context of using ASR to process or manage audio records, this means it may be necessary to investigate whether an ASR system is capable of reliably transcribing a speaker and, if not, whether an institution has the resources and ability to do additional in-house training. Archivists must also consider whether an institution can support the integration of advanced technologies like ASR and AI in the first place, including having in place or acquiring the expensive computing resources to support the technology permanently or on a larger scale (e.g., acquiring multiple GPUs to run ASR) (Brodsky, 2024). Even if the archival organization decides to go with a commercial ASR tool on cloud services rather than using in-house computing resources, relying on a third party raises concerns related to cost, privacy, and ownership of data and records in different jurisdictions. For example, in the European Union (EU), the General Data Protection Regulation (GDPR) classifies voice recordings as personal data. This means that an archival institution may have to get consent prior to using ASR on recordings of individuals and ensure that any tools used do not store data outside of the EU's jurisdictions (European Data Protection Board, 2023).

Integrating ASR into archival workflows also requires adjustment of labour practices. Anderson and Rowe (2025) adapted existing transcript guides and workflows at their institution to streamline the integration of ASR tools. They also chose a commercial product rather than developing an ASR model due to resource constraints and project scope (Anderson and Rowe, 2025). The Archives & Records Association (ARA) in the United Kingdom and

Ireland have also recently released their AI preparedness guidelines for archivists preparing to launch an AI project. Written by Colavizza of the University of Copenhagen and the University of Bologna and Jaillant of the Loughborough University (2026), the authors suggest identifying a clearly defined problem and use case, understanding how complete or partial one's digital corpus is and potential gaps and exclusions, and developing clear human-in-the-loop workflows for reviewing and approving AI outputs before starting a project (Colavizza and Jaillant, 2026).

In summary, the potential applications of AI tools such as ASR for the processing of audio records are evident, although so too are their challenges. Ensuring adequate research, preparation, and documentation of processing workflows alongside continued collaboration is key to successfully adopting these tools to process audio records.

Current and Emerging Processing of Video Records with AI/ML Tools

So far in this module, we have explained how AI/ML tools like ASR can be used to assist the processing of audio records and what challenges need to be addressed before archives adopt these tools in their workflows. In conjunction with AI tools like emotion recognition and facial recognition, these tools can also play a role in processing digitized or born-digital video records as well. As previous modules have already explained how facial and emotion recognition work and their history (Hernandez and Fewster 2024a, 2024b, 2025a, 2025b), the focus of this section is on their application to video-based records.

While Microsoft's integration of AI-based tools such as Copilot has recently received mixed reviews from the public (Corden, 2025), their suite of cloud-based Azure AI services such as the Video Indexer and OpenAI Services has had greater success in records management. Azure AI Video Indexer is a comprehensive AI tool that "extracts deep insights from video (live and uploaded) and audio content" to support "transcription, translation,

object detection, and video summarization" (Microsoft, 2025b). In this way, the Video Indexer effectively combines many of the technologies discussed in this and previous modules, including ASR, language translation, transcription as well as facial and object recognition to enable real-time analysis of digital audiovisual records (Microsoft, 2025b). However, one technology that the Video Indexer utilizes that has not been discussed is emotion recognition. Building on the capabilities of facial recognition software and utilizing AI algorithms, Microsoft's Video Indexer can analyze, identify, and timestamp emotions that appear in parts of a video using Azure AI's Face Service (Microsoft, 2025c). It is important to note, however, that at the time of writing, Microsoft has limited or retired certain facial recognition capabilities due to the potential of discrimination on the basis of race and gender, although it may grant permission to use these functionalities for "responsible use case[s]" (Microsoft, 2025c). The limitations and challenges of AI tools for video-based records will be discussed in greater detail later in this module. Still, while facial recognition is only available on a case-by-case basis, the Video Indexer is also capable of text-based sentiment analysis. Using a collection of ML and AI algorithms, the Video Indexer reviews transcripts for clues to identify words or phrases most commonly associated with key emotions like joy, sadness, anger, and fear (Microsoft, 2025b). Using labelling and segmentation, those emotions are then indexed and timestamped for greater segment discoverability and video navigability. For researchers, this offers a more efficient way to interact with video records if they are looking for specific information (Jansen and Cruz, 2024). The Video Indexer is currently available as a cloud service, but it can also be deployed to edge location (i.e., local data centres) via Azure Arc to meet operational and compliance needs (Microsoft, 2025b).

The viability of Microsoft's Video Indexer application to archival and records management contexts was tested in 2023 by Jansen and Cruz (2024) at the Hawai'i State Archives, where Jansen works as the State Archivist. In this role, Jansen was responsible for the records management of video

recordings of legislative hearings, including arrangement and description, storage, preservation, and ensuring access (Jansen and Cruz, 2024). As there are thousands of bills — and, in turn, records and videos — produced every year, the Hawai'i State Archives naturally struggled to keep up with the influx of records while also guaranteeing access (Jansen and Cruz, 20204). For example, before integrating AI technology, records and videos were placed into different containers, documents, and repositories, including YouTube for oral testimonies (Jansen and Cruz, 2024). This was a fragmented approach that made it difficult for users such as members of the public, researchers, and government representatives to find the records they needed while also limiting the Hawai'i State Archives' ability to fulfill its mandate (Jansen and Cruz, 2024). It is in this context that the State Archives partnered with Microsoft to utilize the Video Indexer to expedite the processing of audiovisual records. Jansen and Cruz (2024) used their Video Indexer to automatically produce timestamped transcripts with speaker diarization and keyword identification to inform description, enrich metadata, and enhance user access and discoverability by being able to search by keyword, speaker, and even topic in any given video and video segments in a matter of minutes. As processing audiovisual materials typically takes four times the length of a video (NADP, 2012), and state legislature recordings often have a running time of many hours, Jansen found the Video Indexer's efficiency and outputs to be unparalleled by existing technologies and approaches (Jasen and Cruz, 2024). Additionally, Jansen and Cruz (2024) continued to work on integrating a large language model (LLM) to further improve the access and discoverability of the Hawai'i State Archives' audiovisual records. They demonstrated the preliminary viability of training an LLM on their holdings to enable interactive searching where a user could simply ask the LLM to retrieve videos that feature specific keywords rather than conducting an advanced search using boolean operators, for example (Jansen and Cruz, 2024). The authors felt this enabled a more organic searching method for archival holdings that may feel more familiar to users

and remove barriers to access records due to lack of experience or willingness to request research and reference help (Jansen and Cruz, 2024).

Another example of applying AI technologies to managing audiovisual records comes from Tsipas et al. (2020). The authors utilized a multi-modal approach using audio and video channels for speaker diarization to "[automate] semantic analysis of multimedia content" in the context of the recent growth of multimedia content on websites such as YouTube (Tsipas et al., 2020, p. 3751). The authors found that facial recognition technology can be used to validate and improve the accuracy of speaker diarization by mapping a speaker to the appearance of their face on a video frame (Tsipas et al., 2020). Their findings demonstrate the viability and potential of a multimodal, audiovisual approach to processing digital audiovisual records, although barriers of cost and resources should still be considered (Tsipas et al., 2020).

Challenges and Constraints Using AI/ML Tools for Video Records

Many — if not all — of the challenges present in the application of AI tools for audio records apply to video records. In fact, challenges relating to computing and storage considerations are often magnified with video records as they have larger file sizes and processing requirements that are otherwise not present with photographs or text-based records. This, in turn, equates to higher costs, which are always of concern in the context of resource-driven archival institutions. For example, as AI infrastructure like data centres around the world increases, critical components for video AI tools such as graphics processing units (GPUs) have skyrocketed in price (Leswing, 2026). This comes as the result of GPU manufacturers such as Nvidia prioritizing multi-billion GPU contracts with companies like OpenAI over individual retail customers (Leswing, 2026). This may make implementing in-house training and models inaccessible to some archival institutions, especially in the context of resource constraints. Similar to ASR technology, one solution may be adapting off-the-shelf products such as Microsoft Azure's Video Indexer

for select archival purposes as they offer variable pricing depending on the duration of the input file (Microsoft, n.d.). Although Microsoft requires interested customers to contact a representative for a full quote, the cost per input minute ranges from $0.1 to $0.2 CAD for cloud options (Microsoft, n.d.). There are also cloud options for processing power, although this may be expensive for video records with longer runtimes and, in turn, processing times and raise concerns surrounding the storage of potentially copyrighted or private data (Ohiri, 2025). These third party options may very well be a cost-effective option for many archival institutions looking to utilize AU tools for the processing of video records depending on the project. Still, as with the use of ASR tools, the total cost of a project and record runtimes should always be estimated first to avoid surprise research expenditures.

The matter of bias is also amplified with video records and AI tools. Not only will video tools have the same language and gender biases as ASR technologies, but with the addition of computer vision and video analysis, appearance-based biases such as race must also be considered. As discussed in Module 4, facial recognition software had an average error rate of just 0.8% for light-skinned men compared to 34.7% for darker-skinned women (Fergus, 2024). These discrepancies stem from unequal and homogenous training data, resulting in biased and potentially inaccurate outputs. Additionally, Stančić (2024) found that some computer vision models had lower accuracy rates when used on older, historical images as they are often underrepresented in training data. This issue may be amplified in the context of moving images or film as the technology has not been in use as long as photography and thus represents a smaller and less accessible dataset.

Finally, an institution must consider the privacy and data implications of implementing video AI tools and workflows given the potential misuse or improper storage of a subject's likeness, which often has greater legal protections than their voice. For example, a bill in Denmark is currently tabled to amend their copyright law to protect citizens' rights in determining how their body, facial features, and voices are used (Bryant, 2025). While

the bill continues to be debated and is primarily intended to combat the production and misuse of deepfakes online (Bryant, 2025), it nonetheless speaks to growing legal and moral concerns regarding the use of AI technologies for video records. It is therefore critical that an archival institution considers the legal ramifications of misusing AI video tools for audiovisual records prior to implementation and establishes a risk profile to help guide development.



ACTIVITY #3

In this activity, students will have the opportunity to analyze a video record using a free Azure AI Video Indexer account to see how ASR and video and image AI tools can be used to process audiovisual records and assist in metadata generation.

Please note that while it is free to create an Azure account with $200 USD worth of free credits, a credit card is required to complete the registration. The credits must be used before they expire within 30 days. Additionally, should a user exceed their initial $200 of credits, the linked credit card will be automatically charged. A live demo using your own account may thus be more appropriate and accessible for students depending on the context. **You may consider using an alternative video analysis tool which does not require a credit card to sign up, such as TwelveLabs, which comes with 10 hours of free runtime.** Please note that TwelveLabs does not have the same features and capabilities as the Azure AI Video Indexer, such as sentiment analysis. However, it does have an interactable LLM that students can use to request a transcript with timestamps or generate key metadata elements such as a title, director, runtime, and key

scenes and events with descriptions. This LLM feature is similar to the one demonstrated by Jansen and Cruz (2024).

1. Have students make a free Azure or TwelveLabs account.
2. In groups, have students record a video or download a public access video file from, for example, the Internet Archive's collection of moving images. Ensure that students select a file that includes audio, preferably speech with multiple speakers. Shorter video clips are also preferable as the process to download and analyze them are faster (and cheaper).
3. Upload the file to Azure or TwelveLabs.
4. Ask students to review either platform's features, including transcript and metadata generation, video summaries, and speaker diarization.
5. As a group, have students share their experiences. For example, have them consider how a tool like may — or may not! — be helpful for processing video records. What workflows could these tools help expedite? At what points do they think a human should be involved to review the outputs? Additionally, if you have groups that used Azure *and* TwelveLabs, have them share their thoughts and whether one tool is better than the other or better suited for specific tasks.



MODULE COMPREHENSION ACTIVITY

Consider the following scenario/project (or make your own!) and in teams develop a plan for how — or if — you would use AI/ML tools to assist with the processing.

"This is a collection of audiovisual records from a private donor that documented local history and politics in their town. No other copies of the materials exist and you are expected to accession everything due to high user demand. You are a full-time processing archivist and have access to a part-time staff member that speaks Spanish fluently. The collection includes:

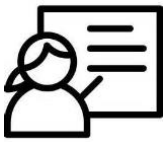
- 15 digital video recordings of high-profile town council meetings. Runtime: ~2 hours each
- 5 analogue audio cassettes featuring interviews with local celebrities, two of which feature a mix of English and Spanish. Runtime: ~1 hour each
- 5 digital audio files of group interviews with prominent and local business people, three of which are in Spanish. Runtime: ~1.5 hours each"

Students may refer to this module's required and recommended readings, Anderson and Rowe (2025) and Colavizza and Jaillant (2026), as starting points. They might also consider the following questions:

- If the NADP (2012) states that processing audiovisual records generally takes four times as long as an item's runtime, what is the estimated time required to process this donation?
    - As there is only one full-time staff member, can this project be completed in a reasonable amount

of time? How will you utilize staff effectively? For what tasks?

- Is there a defined use case and potential benefit for AI tools such as ASR to be used? Are there any concerns or limitations?
- If AI/ML tools will be used, which ones? Will a cloud or in-house option be used and why? At what steps is there human intervention (e.g., to validate outputs)?



SUMMARY

AI demonstrates considerable potential in streamlining the otherwise time-consuming and labour-intensive processing of audiovisual records. This is especially true in the context of metadata enrichment projects, active records management, and financial and staffing constraints that archival institutions continue to face. The development of in-house ASR products for records featuring low-resource languages also presents a unique opportunity for improved awareness, education, and cultural preservation for languages such as East Cree, Inuktitut, Nahuatl, and Totonac.

However, the limitations and challenges of these AI tools regarding privacy, copyright, cost, and processing power must be acknowledged. Implementing AI tools, whether through an on-site product or via cloud infrastructure, requires existing staff knowledge, support from management or an institution, and ongoing maintenance. Proper project planning, guaranteeing a human-in-the-loop approach, and relying on

existing frameworks and methodologies to inform implementation (e.g., Colavizza and Jaillant [2026]) are therefore crucial in successful implementation.

REFERENCES

Abdul-Mageed, M. (2024). AI in the I Trust AI Partnership. In: Duranti, L. and Rogers, C. (Eds.). (2024). Artificial intelligence and documentary heritage. SCEaR newsletter, Special issue 2024, 27-33. https://unesdoc.unesco.org/ark:/48223/pf0000389844

Abdul-Mageed, M. (n.d.). *Audio Guide: Multilingual Audio Analysis*. InterPARES. https://demos.dlnlp.ai/InterPARES/

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., Engel, J., Fang, E., Fan, F., Fougner, C., Gao, L., Gong, C., Hannun, A., Han, T., Johannes, L., Jiang, B., Ju, C., Jun, B., LeGresley, P., Lin, L., Liu, J., Liu, Y., Li, W., Li, X., Ma, D., Narang, S., Ng, A., Ozair, S., Peng, Y., Prenger, R., Qian, S., Zongfeng, Q., Raiman, J., Rao, V., Satheesh, S., Seetaun, D., Sengupta, S., Srinet, K., Sriram, A., Tang, H., Tang, L., Wang, C., Wang, J., Wang, K., Wang, Y., Wang, Z., Wang, Z., Wu, S., Wei, L., Xiao, B., Xie, W., Xie, Y., Yogatama, D., Yuan, B., Zhan, J., & Zhu, Z. (2016). Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. *Proceedings of The 33rd International Conference on Machine Learning*, 48, 173-182. https://proceedings.mlr.press/v48/amodei16.pdf

Anderson, D. & Rowe, S. (2025). What 400 Hours of AI Transcription Taught the Wildenstein Plattner Institute. Archival *Outlook, 6-7. https://mydigitalpublication.com/publication/?i=858398*

Arias-Hernández, R. & Fewster, K. (2024a). Teachable AI for the Archival Professions - Module 1: Introduction to Artificial Intelligence for the Archival Professions. InterPARES Trust AI. https://interparestrustai.org/assets/public/dissemination/AD_01_Module1_v_1_2_202410291.pdf

Arias-Hernández, R. & Fewster, K. (2024b). Teachable AI for the Archival Professions - Module 1: Introduction to Artificial Intelligence for the Archival Professions. InterPARES Trust AI. https://interparestrustai.org/assets/public/dissemination/AD_01_Module2_v_1_0_20241211.pdf

Arias-Hernández, R. & Fewster, K. (2025a). AI/ML for processing textual
records in Archives. InterPARES Trust AI. https://interparestrustai.org/assets/public/dissemination/AD_01_Module3_v_
1_0_202504141.pdf

Arias-Hernández, R. & Fewster, K. (2025b). AL/ML for Processing Image-based Records in Archives. InterPARES Trust AI. https://interparestrustai.org/assets/public/dissemination/AD_01_Module4_v_1_0_20240918.pdf

Bartelds, M., San, N., McDonnell, B., Jurafsky, D., & Wieling, M. (2023). Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation. *Proceedings*

of the 61st Annual Meeting of the Association for Computational Linguistics, *1*, 715-729. https://doi.org/10.18653/v1/2023.acl-long.42

Brodsky, S. (2024, October 14). *The hidden costs of AI: How generative models are reshaping corporate budgets*. IBM. https://www.ibm.com/think/insights/ai-economics-compute-cost

Bryant, M. (2025, June 27). *Denmark to tackle deepfakes by giving people copyright to their own features*. The Guardian. https://www.theguardian.com/technology/2025/jun/27/deepfakes-denmark-copyright-law-artificial-intelligence

Byrne, W., Doermann, D., Franz, M., Gustman, S., Hajič, J., Oard, D., Picheny, M., Psutka, J., Ramabhadran, B., Soergel, D., Ward, T., & Zhu, W. J. (2004). Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Transactions on Speech and Audio Processing*, 12(4). https://doi.org/10.1109/TSA.2004.828702

Cameron, S., Franks, P., & Hamidzadeh, B. (2023). Positioning Paradata: A Conceptual Frame for AI Processual Documentation in Archives and Recordkeeping Contexts. *Journal on Computing and Cultural Heritage, 16*(4), 1-19. https://doi.org/10.1145/3594728

CLAMS. (n.d.) *CLAMS App Directory - inaspeechsegmenter-wrapper*. https://apps.clams.ai/#inaspeechsegmenter-wrapper

Colavizza, G. & Jaillant, L. (2026). *AI Preparedness Guidelines for Archivists*. Archives & Records Association (UK & Ireland). https://static1.squarespace.com/static/60773266d31a1f2f300e02ef/t/

69789fd03543a1698d645315/1769512912257/AI-Preparedness-Guidelines_February_2026.pdf

Coniam, D. (1999). Voice recognition software accuracy with second language speakers of English. *System*, *27*(1), 49-64. https://doi.org/10.1016/S0346-251X(98)00049-9.

Corden, J. (2025, December 8). *Microsoft has a problem: nobody wants to buy or use its shoddy AI products — as Google's AI growth begins to outpace Copilot products*. Windows Central. https://www.windowscentral.com/artificial-intelligence/microsoft-has-a-problem-nobody-wants-to-buy-or-use-its-shoddy-ai

Council of Archives. (2012). *National Archival Development Program: Time Guidelines for Arrangement and Description Projects.* https://www.councilofnsarchives.ca/sites/default/files/ProjectTimelineGuideArrangementandDescription.pdf

Deance, I., Varela, E., & Hernández, H. (2024, October 29). *Net-titlan: uso de IA para la recuperación de información en materiales orales y sonoros de pueblos indígenas de México* [Conference presentation]. Encuentro Iberoamericano de Archivos e Inteligencia Artificial. Ibermemoria Sonora, Fotográfica y Audiovisual. https://www.youtube.com/watch?v=IgEpPAWi8qY

Dragon Medical Transcription. (n.d.). *History of Speech & Voice Recognition and Transcription Software*. https://www.dragon-medical-transcription.com/history_speech_recognition.html

European Data Protection Board. 2023. Guides 01/2022 on data subject rights - Rights of access. https://www.edpb.europa.eu/system/files/2023-04/edpb_guidelines_202201_data_subject_rights_access_v2_en.pdf

Fergus, R. (2024, February 29). Biased Technology: The Automated Discrimination of Facial Recognition. ACLU of Minnesota. https://www.aclu-mn.org/en/news/biased-technology-automateddiscrimination-facial-recognition

Foster, K. (2025, October 15). *What is Automatic Speech Recognition? A Comprehensive Overview of ASR Technology*. AssemblyAI. https://www.assemblyai.com/blog/what-is-asr

Gupta, V. & Boulianne, G. (2020). Automatic Transcription Challenges for Inuktitut, a Low-Resource Polysynthetic Language. *Proceedings of the 12th Conference on Language Resources and Evaluation*, 2521-2527. https://aclanthology.org/2020.lrec-1.307.pdf

Gustman, S., Ramabhadran, B., Picheny, M., Franz, M., Kambhatla, N., Byrne, W., Psutka, J., Hajic, J., Soegel, D., & Oard, D. W. (n.d.). *MALACH: Multilingual Access to Large Spoken ArCHives*. [PowerPoint Slides]. https://www.ee.columbia.edu/~stanchen/e6884/slides/lecture12.malach.pd

Hansen, U. S. (2024, December 12). *A Guide to Speaker Recognition: How to Annotate Speech*. Encord. https://encord.com/blog/guide-to-speaker-recognition/

Harris, C., Mgbahurike, C., Kumar, N., & Yang, D. (2024). Modeling Gender and Dialect Bias in Automatic Speech Recognition. *Findings of the Association for Computational Linguistics: EMNLP 2024, 15166-15184. https://doi.org/10.18653/v1/2024.findings-emnlp.890*

Hux, K., Rankin-Erickson, J., Manasse, N., & Lauritzen, E. (2000). Accuracy of Three Speech Recognition Systems: Case Study of Dysarthric Speech. *Augmentative and Alternative Communication*, 16. https://www.proquest.com/scholarly-journals/accuracy-three-speech-recognition-systems-case/docview/220492192/se-2?accountid=14656

IASA, International Association of Sound and Audiovisual Archives Technical Committee (2009). Guidelines on the Production and Preservation of Digital Audio Objects. Ed. by Kevin Bradley. Second edition 2009. Standards, Recommended Practices and Strategies, IASA-TC 04. www.iasa-web.org/tc04/audio-preservation

IBM. (n.d.). *What is speech recognition? https://www.ibm.com/think/topics/speech-recognition*

Jackson, J. (2001, September 26). *NSF Hands out $156 Million to Spur IT Innovation*. Washington Technology. https://www.washingtontechnology.com/2001/09/nsf-hands-out-156-million-to-spur-it-innovation/345197/?oref=wt-next-story

Jansen, A. and Cruz, M. (2024). iTrust AI Public Symposium. Using AI To Interrogate Archives - Enhancing Access to Video Using Lox Code/No Code AI Services (Conference Presentation). iTrust AI Public Symposium, Honolulu, Hawai'i. February 23, 2024. https://www.youtube.com/watch?v=hJjifghl6AQ&t=7074s&ab_channel=Hawai%CA%BBiStateArchives

Juang, B.H. & Rabiner, L. (2004). Automatic Speech Recognition - A Brief History of Technology Development. https://web.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf

Karatas, G. (2025, July 2). *Speech Recognition: Everything You Need to Know*. AI Multiple. https://research.aimultiple.com/speech-recognition/

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quarry, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 117(14), 7684-7689. doi: 10.1073/pnas.1915768117.

Kummer, J. C., & Backfried, G. (2007). Archiving meets automatic speech recognition - curse of blessing? https://www.thejts.org/wp-content/uploads/2015/03/JTS_NOA_SLT_Paper.pdf

La Javaness. (2023, July 17). *Speaker Diarization: An Introductory Overview*. Medium. https://lajavaness.medium.com/speaker-diarization-an-introductory-overview-c070a3bfea70

Leib, A. (n.d.). *An Introduction to Nuance Dragon NaturallySpeaking and Dragon Dictate*. Western University. https://www.westernu.edu/mediafiles/cdihp/an_introduction_to_nuance_dragon_naturallyspeaking.pdf

Levin, S. (206, September 8). *A beauty contest was judged by AI and the robots didn't like dark skin*. The Guardian.

https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people

Leswing, K. (2026, January 10). *AI memory is sold out, causing an unprecedented surge in prices*. CNBC. https://www.cnbc.com/2026/01/10/micron-ai-memory-shortage-hbm-nvidia-samsung.html

Loeber, P. (2025, November 18). *The best audio file formats for speech-to-text: A guide.* AssemblyAI. https://www.assemblyai.com/blog/best-audio-file-formats-for-speech-to-text

MALACH. (n.d.-a). *MALACH*. https://www.scribbr.com/apa-examples/citing-online-sources-no-author-date-title/

MALACH. (n.d.-b). *Publications*. https://malach.umiacs.umd.edu/malach_pubs.html

Meng, L., Hu, S., Kang, J., Li, Z., Wang, Y., Wu, W., Wu, X., Liu, X., & Meng, H. (2024). Large Language Model Can Transcribe Speech in Multi-Talker Scenarios with Versatile Instructions. International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.48550/arXiv.2409.08596

Microsoft. (2025a, November 23). *Audio concepts in Azure Speech in Foundry Tools. https://learn.microsoft.com/en-us/azure/ai-services/speech-service/concepts/audio-concepts*

Microsoft. (2025b, December 11). *Azure AI Video Indexer Overview*. https://learn.microsoft.com/en-us/azure/azure-video-indexer/

Microsoft. (2025c, November 18). *Azure Vision in Foundry Tools Documentation*. https://learn.microsoft.com/pdf?url=https%3A%2F%2Flearn.microsoft.com%2Fen-us%2Fazure%2Fai-services%2Fcomputer-vision%2Ftoc.json

Microsoft. (n.d.). *Azure AI Video Indexer pricing*. https://azure.microsoft.com/en-us/pricing/details/video-indexer/

Moskvitch, K. (2017, February 17). *The machines that learned to listen*. BBC. https://www.bbc.com/future/article/20170214-the-machines-that-learned-to-listen

Munteanu, C., Baecker, R., Penn, G., Toms, E., & James, D. (2006). The Effect of Speech Recognition Accuracy Rates on the Usefulness and Usability of Webcast Archives.

Munteanu, C., Penn, G., & Baecker, R. (2007). Web-Based Language Modelling for Automatic Lecture Transcription. https://www.cs.utoronto.ca/~mcosmin/pubs/interspeech2007.pdf

Munteanu, C., Penn, G., & Baecker, R. (2008). Usable speech recognition: toward improved access to webcast lectures. https://www.cs.toronto.edu/~gpenn/papers/chiuai08.pdf

Nanjo, H., Akita, Y., & Kawahara, T. (2006). *Computer assisted speech transcription system for efficient speech archive* [Paper]. WESPAC IX, South Korea. http://sap.ist.i.kyoto-u.ac.jp/EN/bib/intl/NAN-WESPAC06.pdf

Navon, A., Shamsian, A., Glazer, N., Hetz, G., & Keshet, J. (2023). Open-vocabulary keyword-spotting with adaptive instance normalization. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 11656-11660. https://doi.org/10.48550/arXiv.2309.08561

Nuance. (2023, September 12). *The Evolution of Dragon Medical*. https://www.nuance.com/asset/en_us/collateral/healthcare/infographic/ig-dmo-evolution-en-us.pdf

Noordegraaf, J., & Schjøtt, A. (2025). From preservation to access and beyond: the role of AI in audiovisual archives. In L. Jaillant, C. Warwick, P. Gooding, K. Aske, G. Layne-Worthey, & J. S. Downie (Eds.), *Navigating Artificial Intelligence for Cultural Heritage Organisations* (pp. 93–112). UCL Press. https://doi.org/10.2307/jj.24215718.11

Oard, D. (2012). *Can Automatic Speech Recognition Replace Manual Transcription*? Oral History in the Digital Age. https://ohda.matrix.msu.edu/2012/06/automatic-speech-recognition/

Ohiri, E. (2025, August 29). *AItraining cost: Hyperscalers vs specialized platforms*. CUDO Compute. https://www.cudocompute.com/blog/ai-training-cost-hyperscaler-vs-specialized-cloud

Ostendorf, M., Favre, B., Grishman, R., Hakkani-Tür, Haper, M., Hillard, D., Hirschberg, J., Ji, H., Khan, J. G., Liu, Y., Maskey, S., Matusov, E., Ney, H., Rosenberg, A., Shriberg, E., Wang W., & Wooters, C. (2008). Speech Segmentation and Spoken Document

Processing. *IEEE Signal Processing Magazine*, 25(3), 59-69. https://doi.org/10.1109/MSP.2008.918023

Parente, R., Kock, N., & Sonsini, J. (2004). An analysis of the implementation and impact of speech-recognition technology in the healthcare sector. *Perspectives in Health Information Management*, *1*. https://pmc.ncbi.nlm.nih.gov/articles/PMC2047322/

Picheny, M., Tuske, Z., Kingsbury, B., Audhkhasi, K., Cui, X., & Saon, G. (2019). *Challenging the Boundaries of Speech Recognition: The Malach Corpus* [Paper]. INTEERSPEECH, Austria. https://doi.org/10.48550/arXiv.1908.03455

Picheny, M., Yang, Q., Zhang, D., & Zhang, L. (2023). *The MALACH Corpus: Results with End-to-End Architectures and Pretraining* [Paper]. *INTERSPEECH*, Ireland. https://www.isca-archive.org/interspeech_2023/picheny23_interspeech.pdf

Rim, K., King, O. C., Lynch, K., Verhagen, M., & Pustejovsky, J. (2025). A platform for AI-Assisted Archival Metadata Generation. *International Conference on Human-Computer Vision*, 183-203. https://doi.org/10.1007/978-3-031-93160-4_12

Ramabhadran, B., Gustman, S., Byrne, W., Hajic, J., Oard, D., Olsson, S. J., Picheny, M., & Psutka, J. (2012). *USC-SFI MALACH Interviews and Transcripts English*. Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC2012S05

Sanabria Medina, G. & Rodríguez Reséndiz, P. O. (2023). Inteligencia artificial en los procesos documentales de los archivos digitales sonoros. *Investigación Bibliotecológica Archivonomía Bibliotecología e*

Información 36(93), 73-88. https://doi.org/10.22201/iibi.24488321xe.2022.93.58618

Stryker, C. (n.d.). *What is human-in-the-loop?* IBM. https://www.ibm.com/think/topics/human-in-the-loop

Sullivan, P. and Sengsavang, E. (2024). UNESCO Audio Archives: AI for Metadata Enrichment. In: Duranti, L. and Rogers, C. (Eds.). (2024). Artificial intelligence and documentary heritage. SCEaR newsletter, Special issue 2024. pp. 27-33. PURL: https://unesdoc.unesco.org/ark:/48223/pf0000389844

Timmel, V., Paonessa, C., Kakooee, R., Vogel, M., & Perruchoud, D. (2024). Fine-tuning Whisper on Low-Resource Languages for Real-World Applications.In Proceedings of the 10th edition of the Swiss Text Analytics Conference, pages 57–65, Winterthur, Switzerland. https://doi.org/10.48550/arxiv.2412.15726

Tsipas, N., Vrysis, L., Konstantoudakis, K., & Dimoulas, C. (2020). Semi-supervised audio-driven TV-news speaker diarization using deep neural embeddings. The Journal of the Acoustical Society of America, 148(6), 3751-3761. https://doi.org/10.1121/10.0002924

Whittaker, S., Hirschberg, J., Choi, J., Hindle, D., Pereira, F., & Singhal, A. (1999). *SCAN: Designing and evaluating user interfaces to support retrieval from speech archives* [Paper]. ACM SIGIR, USA. https://doi.org/10.1145/312624.312639

Visual History Archive. (n.d.). *Home*. University of Southern California Shoah Foundation. https://vha.usc.edu/home