Case Study: Artificial Intelligence in the UNESCO Audio Archives

Richard Arias-Hernandez[1]

Educational applications: This case study can be used in introductions of AI uses for archives, and more specifically for non-textual records. It illustrates the use of AI to enrich the description and metadata of audio-records at the item level in order to improve on their access and discoverability. It motivates, exemplifies, and prompts discussions of AI for automatic speech recognition, speaker recognition, and summarization of audio records. From a diplomatics perspective, this case can be used to discuss modern uses of diplomatics on digitized audio-records to identify structural elements that can be used to train and enhance current AI summarization models.

Educational topics: AI for non-textual records (audio), types of AI/ML for audio-archives, diplomatics for AI, critical AI/ML for archives, evaluations of ML models used for archival tasks.[2]

About: This case study 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 final draft was completed on August 13th, 2024. It has a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International BY-NC-SA 4.0 license, which requires that reusers give credit to the creator. 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]

In 2022, UBC researcher Peter Sullivan and UNESCO archivist Eng Sengsavang embarked on a project to test the effectiveness of applying artificial intelligence (AI) and archival diplomatics to support the processing and description of the UNESCO audio archives. Specifically, by using automatic speech recognition (ASR) to identify language and transcribe recorded speech; by using speaker recognition tools to identify speakers in audio recordings; and by using principles of archival diplomatics to identify and label key structural aspects of the recordings to support accurate AI summarization.

According to Sullivan and Sengsavang: "From 2017 to 2020, the UNESCO Archives digitized 15,316 archival audio recordings dating from the 1950s to the 1980s reel-to-reel magnetic tape.

[1] Associate Professor of Teaching, School of Information, University of British Columbia. richard.arias@ubc.ca

[2] Educational applications map to a Body of Knowledge proposed by InterPARES researchers for AI/ML for the archival professionals. https://docs.google.com/document/d/1UsjkkkGeSJrgCDJGASCAy5q0uo_ZkQpzi_Ch8XUcqYw/edit?usp=sharing

[3] This case study 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] Case Study: Artificial Intelligence in the UNESCO Audio Archives © 2024 by 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/

The recordings document a remarkable range of topics, personalities, geographies, languages, and genres across four decades and in over 70 languages, including radio programmes, interviews, speeches, events, music, and more, reflecting the substance of UNESCO's work and its international, intergovernmental character. An additional 2,314 recordings from the same collection were repaired and digitized between 2021 and 2022. The digitized recordings - approximately 17,630 in total - now form a substantial part of UNESCO's digital archives." (2024:27). By the time their project started, only around 800 of such recordings (4.5%) had been manually described at the item level (Sullivan, 2023), so there was a big motivation to try out AI methods to speed up this process.

Sullivan and Sengsavang used OpenAI's Whisper (Radford et. al., 2023) to automatically identify language and generate transcriptions from the audio recordings. According to the product's website: "Whisper is an automatic speech recognition system trained on 680,000 hours of multilingual and multitask supervised data collected from the web" (OpenAI, 2024). Since the data Whisper was trained on is different from the actual audio records UNESCO was working on, the authors were prompted to conduct thorough evaluation of Whisper's outcomes. They designed multiple experimental evaluations to test the accuracy of the tool on the transcription of UNESCO audio recordings by focusing on variables that represent their unique characteristics, such as: time-range of the recordings, variety of recording techniques, multilingual speech, and different accents (Sullivan and Sengsavang, 2024). In their 2024 report, the authors stated that "on the multilingual speakers set, we found that the newest version of Whisper (Large V3) performs at 92% accuracy" (Sullivan and Sengsavang, 2024).

A second aspect of Sullivan and Sengsavang's project was the use of speaker recognition tools to identify (and name) speakers in the audio recordings based on their unique voice characteristics (i.e., speaker embeddings). These are AI technologies that are currently used in commercial applications such as Amazon's Alexa and Apple's Siri. They rely on deep neural networks or convolutional neural networks to convert the acoustic pattern of human voices into a probability distribution over speech sounds. Identifying speakers automatically in the audio records and naming them in the metadata would help enormously with the description and access to these records. Sullivan and Sengsavang found in their evaluation of these AI tools that they were limited in their accuracy to identify the same speaker in different UNESCO audio recordings when the speaker would change their voice/speech over the years (i.e., cross-age shift/effects) or when the speaker would switch from one language to another (i.e., code-switching) (Sullivan and Sengsavang, 2024).

Lastly, Following Duranti (1998), Sullivan and Sengsavang also proposed an archival diplomatics approach to identify structural components in records grouped by genre (e.g., interviews, speeches, radio programmes, etc.) in order to treat them as key features to be identified and described by AI (2024). Their approach identified consistent structural properties of each recording genre found in transcriptions of UNESCO audio archives at the beginning, body, and end portions of the recordings. For example, at the beginning of recorded interviews a narrator/host introduces the main speaker/guest and states their name. The authors propose that this diplomatic 'element' for recorded interviews can be treated as a label to train AI to learn to accurately identify and describe the name of the interviewee in this group of records (Sullivan and Sengsavang, 2024). Extension of this archival diplomatics analysis for the labeling and training of AI for summarization of the content in the UNESCO audio records was still ongoing in 2024.

Potential Discussion Questions:

1. Explain how AI can be used to enhance access to audio-records. What are specific AI technologies that can be used for this purpose?
2. What are the current limitations of ASR (Automated Speaker Recognition) technologies when applied to historical audio records, as evidenced by this case?
3. Discuss the use of diplomatics for the identification of structural elements in the UNESCO audio records in order to enhance their AI summarization and their theoretical implications for archival diplomatics?
4. What can be learned from this case study regarding evaluation methods of AI technologies for archives?

References

Duranti, L. (1998). *Diplomatics: New Uses for an Old Science*. Society of American Archivists and Association of Canadian Archivists in association with The Scarecrow Press, Inc., Maryland.

OpenAI (2022). *Introducing Whisper*. Retrieved July 5, 2024, from https://openai.com/index/whisper/

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. Proceedings of Machine Learning Research, 2023.

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

Sullivan, P. (2023). Applying AI Tools in Archival Functions. Slides presented at InterPARES Summer School Symposium, July 7 2023, San Benedetto del Tronto, Italy. Accessible at: https://interparestrustai.org/assets/public/dissemination/6-SullivanPlenary8Symposium.pdf