Toward an Advanced Web Interface for Analyzing and Annotating Digitized Audio Tape Recordings

Niccolò Pretto[1,*], Andrea Franceschini[2], Alessandro Russo[3], Matteo Spanio[2] and Sergio Canazza[2]

[1]Free University of Bozen Bolzano, Faculty of Engineering, Bolzano, Italy
[2]University of Padua, Department of Information Engineering, Padua, Italy
[3]Národní filmový archiv (NFA), Prague, Czech Republic

Abstract

The access and study of digitized historical audio recordings require more than simple sound playback, as these materials embody complex physical and contextual information. Previous studies have explored the virtualization of tape recorder functionalities; however, such approaches alone are insufficient to fully support scholarly analysis and restoration processes. Recent developments in standards, such as IEEE/MPAI Standard 3302-2024 on Audio Recording Preservation, have emphasized the potential of integrating automated multimodal analysis to support restoration and access, including valuable contributions from artificial intelligence techniques. The main contribution of this work is a standalone web interface for analyzing and annotating digitized historical audio recordings, integrating synchronized audio and video of the tape during the digitization process to provide additional insights, such as annotations and the presence of splices. The proposed interface represents an initial step toward the integration of MPAI-based technologies for restoration. In this context, a computer vision approach is also introduced and evaluated, capable of identifying key components such as the playback head, magnetic tape, and pinch roller, which are essential for supporting future extensions toward (semi-)automated restoration techniques.

Keywords

Historical audio documents, Digitized audio recordings, Web interface for cultural heritage, Computer vision for cultural heritage

1. Introduction

Analog audio recordings are a valuable source of information for several research fields such as linguistics, oral history, and ethnomusicology. However, this important part of the cultural heritage risks being permanently lost due to obsolescence (both playback devices and competencies [1]), degradation of analog carriers with short life expectancy (e.g., magnetic tapes [2]), high digitization and preservation costs, as well as the large amount of audio documents that must be preserved [3].

A significant yet often overlooked cost in the digitization process is the verification of digitized recordings, which typically requires full-length listening. Similarly, the analysis of audio content requires substantial time and effort. In this context, artificial intelligence techniques can contribute to reducing these costs by supporting both technical staff and researchers in their respective tasks, while also reducing potential errors [4].

For magnetic tapes containing culturally significant works, particularly in the domain of electronic music, it is possible to complement the audio signal with video recordings that capture the tape as it passes over the playback head. This visual information can provide relevant insights on the creation process of the work and, in some cases, can be interpreted as a form of score [5]. However, manual

Workshop on Advanced Visual Interfaces for Cultural Heritage (AVICH 2026), June 08, 2026, co-located with the 18th International Conference on Advanced Visual Interfaces, Venice, Italy.

*Corresponding author.
niccolo.pretto@unibz.it (N. Pretto); andrea.franceschini.2@unipd.it (A. Franceschini); alessandro.russo@nfa.cz (A. Russo); spanio@dei.unipd.it (M. Spanio); sergio.canazza@unipd.it (S. Canazza)
0000-0002-3742-7150 (N. Pretto); 0000-0001-8665-6147 (A. Franceschini); 0000-0001-6691-759X (A. Russo); 0000-0002-2436-7208 (M. Spanio); 0000-0001-7083-4615 (S. Canazza)

analysis of these videos is highly demanding, as it requires prolonged observation of continuously moving tape, since recordings may exceed one hour in duration.

In this context, particular attention is devoted to the identification of specific events and conditions occurring during playback, here referred to as irregularities. These include, among others, physical and visual features such as splices, tape brands, start and end of tape conditions, surface damage (e.g., ripples, cupping, and edge deformations), contamination (e.g., mould or dust), annotations or marks on the tape, as well as visual artifacts such as shadows and reflections. This creates a strong case for tools that can support systematic inspection and annotation workflows.

The Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) framework on Audio Recording Preservation (ARP), recently standardized within IEEE as IEEE Std 3302-2024 [6], promotes the use of artificial intelligence techniques applied to both audio and video signals, enabling the integration of multimodal information to facilitate restoration and analysis. Although preliminary implementations of individual functional blocks have been proposed in previous works [5] [7], an integrated prototype interface that supports their practical use in inspection and annotation workflows is still lacking. Moreover, the modular nature of MPAI standards, maintained since their early versions [8, 9], allows for progressive improvement of the underlying algorithms.

This work presents a web interface designed to support the inspection, navigation, and annotation of digitized audio tape recordings, integrating automatically detected regions of interest into an interactive analysis environment. In addition, a computer vision approach is introduced to identify key regions in tape video recordings, namely the playback head, the portion of magnetic tape under the head (hereafter 'tape region'), and the pinch roller, which are essential elements for supporting further analysis and enabling future extensions toward automated processing. While interfaces for visualizing videos of magnetic tape recordings have previously been developed [10], to the best of our knowledge there are no web interfaces that integrate both the annotation of irregularities and the detection of tape recorder components.

The paper is organized as follows. Section 2 presents the proposed web interface. Section 3 describes the computer vision approach for detecting key components in video recordings of tapes. Finally, Section 4 concludes the paper and outlines future work.

2. An integrated interface for digitized audio tape recordings

The proposed web interface (Figure 1) is designed to support both the analysis of digitized audio tape recordings and the manual annotation of irregularities. The system is conceived as a standalone tool for local installation, allowing users to directly upload video files in MP4 or MOV format or simply use a predefined directory. Once a file is loaded, it is automatically processed by the system. The corresponding audio waveform is extracted and displayed, enabling efficient navigation within the recording. In addition, regions of interest are automatically identified through a computer vision module, including the playback head, the tape region, and the pinch roller (see Section 3 for details on the adopted approach). These regions can be optionally visualized as overlays on the video stream and serve as visual support for the annotation process.

Irregularities are presented to the user as markers on the waveform. Each marker is associated with an icon representing its category and can be selected to navigate directly to the corresponding temporal position in the recording. A comprehensive list of irregularities is provided on the right side of the interface, following the taxonomy defined in the MPAI/IEEE Std 3302-2024 [6], thus supporting consistent and standardized annotation.

The interface also enables the manual insertion of new irregularities, which are then stored in JSON format, providing an id, the frame number and the irregularity type. In this context, the tool serves not only as an analysis environment, but also as a means of creating annotated datasets, which are essential to improve the performance of the underlying algorithms and advance the implementation of the modular framework provided by the standard. To facilitate precise annotation, frame-by-frame navigation is supported. Temporal references are expressed using the `hh:mm:ss.ff` format, where `ff`

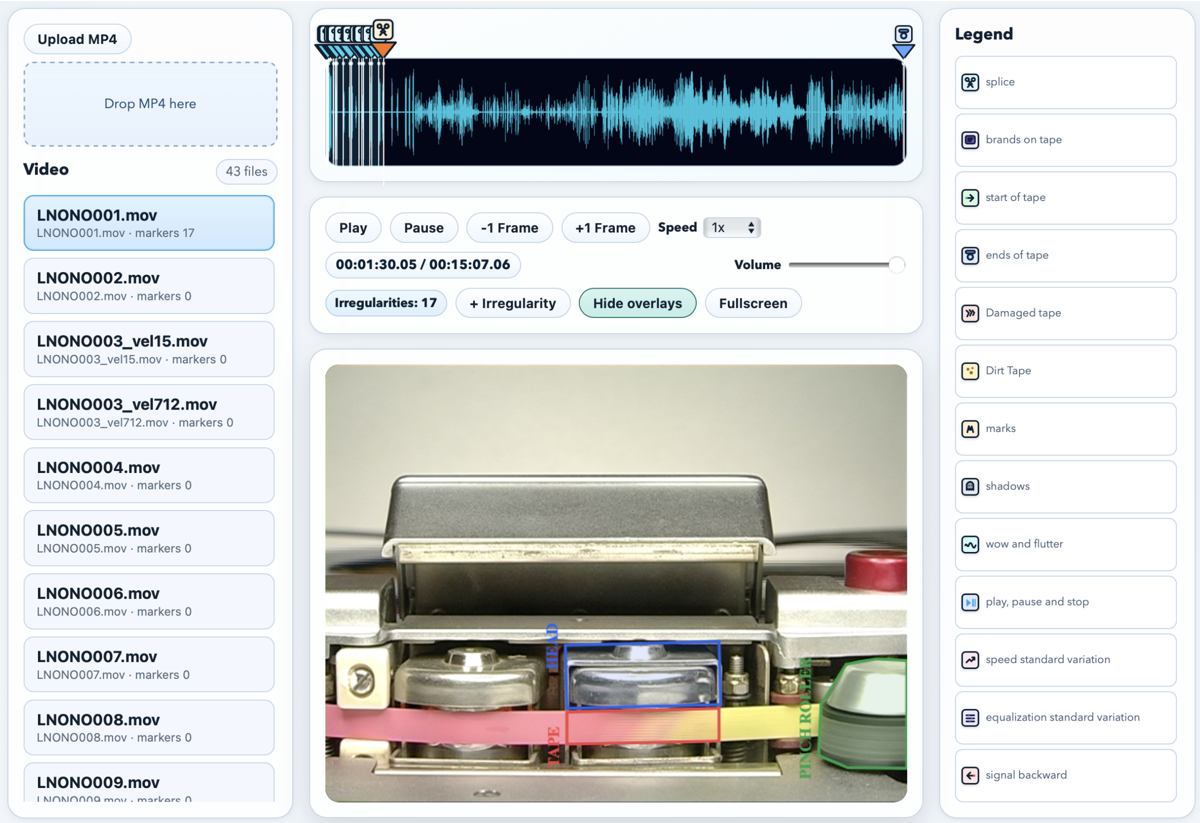

Figure 1: Screenshot of the web interface for digitized audio tape recordings. The player is paused in a frame containing a splice, connecting two distinct tapes. Three polygons indicate the inferred position of the playback head (blue), tape region (red) and the pinch roller (green). On the right, the list of irregularities are reported: splices, brands on tape, start of tape, end of tape, damaged tape, dirt tape, marks, shadow, wow and flutter, play pause and stop, speed variation, equalization variation, signal backwards.

denotes the frame index within the corresponding second.

Finally, the interface provides playback speed control, allowing users to slow down or speed up the reproduction of the recording. This functionality is particularly important when dealing with tapes recorded at different speeds, as it enables proper auditory inspection and analysis, which would otherwise be difficult or impossible [4].

3. Automatic detection of playback head, magnetic tape and pinch roller

The detection of the playback head is required to identify the portion of the tape being reproduced at any given time and is therefore essential for detecting irregularities that may explain the presence of specific sounds, noise artifacts, or silence. The pinch roller is also a critical component: playback begins only when it moves toward the capstan and ends when it returns to its initial position. The identification of these components is fundamental for the development of AI-based tools for irregularity analysis, as well as for the broader study of magnetic tapes. In this section, we present an algorithm capable of detecting these three components, thoroughly evaluated on a subset of a major Italian archival collection.

3.1. Dataset

The experiments are conducted on a manually reviewed ground truth dataset built from digitized magnetic tape recordings from the Luigi Nono archive[1]. The archive contains 209 digitized recordings of the composer's works, as well as preliminary studies and materials related to his electronic compositions. For this reason, the possibility of visualizing irregularities (in particular splices potentially indicating the merging of different tapes or marks containing notes by the composer) is especially important for musicological studies.

The dataset includes 43 of the 209 digitized tape recordings, together with video recordings capturing the tape flowing on the playback head during the digitization. A total of 86 annotated frames were extracted, with polygonal annotations for three components: the playback head, the tape region, and the pinch roller. The stationary components (head and tape) are annotated on one reference frame per video, while the pinch roller is annotated on two frames to capture its motion: the first frame was extracted from the second frame of the video, while the second is extracted from the middle frame of the video.

3.2. Algorithm description and experimental setup

In a previous implementation [7], the detection of regions of interest in tape video frames relied on SURF (Speeded Up Robust Features) [11] for keypoint extraction and description. SURF, like the earlier SIFT algorithm [12], constructs a Gaussian-based scale space in which each pyramid level introduces progressive boundary information loss, which may affect localization accuracy [13]. In contrast, the ORB (Oriented FAST and Rotated BRIEF) algorithm [14] replaces floating-point descriptors with compact binary strings derived from a steered BRIEF scheme, enabling matching via Hamming distance rather than Euclidean distance and resulting in substantially lower computational cost while maintaining robust performance in feature matching tasks [13]. For these reasons, the present work adopts a homography-based geometric transfer approach that relies on ORB feature extraction, brute-force matching, and RANSAC-based [15] homography estimation to project annotated polygons from reference frames to unseen test frames.

Four tests were performed: one for the evaluation of the algorithm performance on playback head and tape, and three for the evaluation of the pinch roller detection. A leave-one-video-out protocol is used to ensure robustness, where each video is tested against all others used as references. As described in Table 1, two distinct tests were used to evaluate a) the initial frame (and therefore the pinch roller position on pause) for all the video and b) the middle frame (pinch roller positioned during the playback) for all the videos. Furthermore, an aggregate test was also conducted using all frames. Considering that two frames were extracted from the a single video, during each iteration the unused frame from the video under evaluation were discarded to avoid data leakage.

Evaluation is performed at the mask level using standard metrics, including Intersection over Union (IoU), Dice coefficient, pixel precision, pixel recall, and success rates at different IoU thresholds.

Table 1
Experimental results on the detection of the playback head, tape region and pinch roller. The Frames column refers to the number of tested frames. The metrics are Intersection over Union (IoU), Dice coefficient (Dice), pixel precision (Prec.), pixel recall (Rec.), and success rates at 0.75 IoU thresholds (S@0.75).

Component	Frames	IoU	Dice	Prec.	Rec.	S@0.75
Playback Head	43	0.9862	0.9930	0.9932	0.9929	1.0000
Tape	43	0.9806	0.9901	0.9902	0.9901	1.0000
Pinch roller (initial)	43	0.9325	0.9610	0.9884	0.9405	0.9302
Pinch roller (middle)	43	0.9762	0.9878	0.9927	0.9833	1.0000
Pinch roller (aggregate)	86	0.9442	0.9679	0.9835	0.9565	0.9419

[1]https://www.luiginono.it/en/ (Last accessed May 25, 2026)

3.3. Results and discussion

The results show that stationary components are highly stable across recordings, allowing nearly perfect geometric transfer without failure cases. In contrast, the pinch roller, being the only moving component, presents slightly greater variability. Despite this, the method achieves strong performance, particularly in the middle-frame protocol, while a slight degradation is observed in the initial-frame condition due to more challenging geometrical configurations. Overall, the results confirm the effectiveness of geometric transfer as a reliable baseline for component localization in tape video recordings, making it suitable for interactive use within the proposed interface.

4. Conclusions

This work presented a preliminary system designed as a human-in-the-loop environment supporting archivists and researchers rather than replacing expert judgment. It allows the analysis and annotation of digitized audio tape recordings, and combines a web-based interface with a computer vision approach for component detection. The experimental evaluation demonstrates that a homography-based geometric transfer method can accurately identify key elements of the tape transport mechanism, achieving near-perfect performance for stationary components and robust results for the moving pinch roller. These findings support the feasibility of integrating automatic analysis tools into interactive systems for the exploration and annotation of historical audio documents, paving the way for future extensions toward more advanced and scalable approaches, and supporting more efficient analysis of large audio heritage collections.

Acknowledgments

The authors would like to thank the Luigi Nono Foundation in Venice (Italy) for providing access to part of its digitized archive of magnetic tapes, which was used as ground truth for the experiments on component detection in tape video recordings and for testing the proposed prototype.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT-5.5 and Grammarly to perform grammar and spelling checks. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. W. Davies, Knowledge: The Missing Element in Archiving and Restoration?, Journal of the Audio Engineering Society (2007).

[2] R. L. Hess, Tape Degradation Factors and Predicting Tape Life, Journal of the Audio Engineering Society (2006).

[3] M. Casey, Article why media preservation can't wait: the gathering storm, IASA Journal 44 (2015).

[4] N. Pretto, E. Micheloni, A. Chmiel, N. D. Pozza, D. Marinello, E. Schubert, S. Canazza, Multimedia archives: New digital filters to correct equalization errors on digitized audio tapes, Advances in Multimedia 2021 (2021) 5410218. doi:10.1155/2021/5410218.

[5] M. Bosi, S. Canazza, N. Pretto, A. Russo, M. Spanio, From tape to code: An international AI-based standard for audio cultural heritage preservation - Don't play that song for me (if it's not preserved with ARP!), IEEE Access 12 (2024) 152544–152558. doi:10.1109/ACCESS.2024.3474529.

[6] IEEE, IEEE/MPAI Standard for Adoption of Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) – Technical Specification: Context-based Audio Enhancement (CAE), Version 2.1, 2024. URL: https://standards.ieee.org/ieee/3302/11691/.

[7] A. Russo, M. Spanio, S. Canazza, Enhancing preservation and restoration of open reel audio tapes through computer vision, in: G. L. Foresti, A. Fusiello, E. Hancock (Eds.), Image Analysis and Processing - ICIAP 2023 Workshops, Springer Nature Switzerland, 2024, pp. 297–308. doi:10.1007/978-3-031-51026-7_26.

[8] M. Bosi, N. Pretto, M. Guarise, S. Canazza, Sound and music computing using AI: Designing a standard, in: Proceedings of the 18th Sound and Music Computing Conference, SMC'21, Virtual Conference, 2021, p. 215 – 218. doi:10.5281/zenodo.5045003.

[9] A. Basso, P. Ribeca, M. Bosi, N. Pretto, G. Chollet, M. Guarise, M. Choi, F. Yassa, R. Iacoviello, AI-based media coding standards, SMPTE Motion Imaging Journal 131 (2022) 10–20. doi:10.5594/JMI.2022.3160793.

[10] C. Fantozzi, F. Bressan, N. Pretto, S. Canazza, Tape music archives: from preservation to access, International Journal on Digital Libraries 18 (2017) 233–249. doi:10.1007/s00799-017-0208-8.

[11] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF), Computer Vision and Image Understanding 110 (2008) 346–359. doi:10.1016/j.cviu.2007.09.014.

[12] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2004) 91–110. doi:10.1023/B:VISI.0000029664.99615.94.

[13] Y. Zhang, Z. Wang, Y. Xu, Y. He, Q. Zhu, H. Sheng, Local adaptive clustering based image matching for automatic visual identification, in: 2023 China Automation Congress (CAC), 2023, pp. 5686–5690. doi:10.1109/CAC59555.2023.10450332.

[14] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: An efficient alternative to SIFT or SURF, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2011, pp. 2564–2571. doi:10.1109/ICCV.2011.6126544.

[15] M. A. Fischler, R. C. Bolles, Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, Communications of the ACM 24 (1981) 381–395. doi:10.1145/358669.358692.