IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 July 2024, accepted 26 September 2024, date of publication 4 October 2024, date of current version 25 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3474529

APPLIED RESEARCH

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!)

MARINA BOSI[1], (Senior Member, IEEE), SERGIO CANAZZA[2], NICCOLÒ PRETTO[3], (Member, IEEE), ALESSANDRO RUSSO[2], AND MATTEO SPANIO[2]

[1]Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, Stanford, CA 94305, USA
[2]Centro di Sonologia Computazionale (CSC), Department of Information Engineering (DEI), University of Padua, 35131 Padua, Italy
[3]Media Interaction Laboratory, Faculty of Engineering, Free University of Bozen-Bolzano, 39100 Bozen, Italy

Corresponding author: Sergio Canazza (sergio.canazza@unipd.it)

ABSTRACT This article describes a novel technology for preserving audio documents archived on open-reel magnetic tapes forming the core of the Audio Recording Preservation (ARP) international standard. ARP is part of the Moving Picture, Audio, and Data Coding by Artificial Intelligence (MPAI) Context-based Audio Enhancement (CAE) standard, adopted by the IEEE Standard Association as IEEE 3302-2022 in December 2022. Leveraging automated Artificial Intelligence (AI) tools, ARP analyzes and extracts relevant information from digitized audio and video files of the tape's corresponding digital Preservation Copy. This process includes identifying speed variations and surface irregularities on the tape, automatically rectifying errors to generate a restored Access Copy. By utilizing the ARP standard, archives gain a potent tool for expediting and optimizing the description of the preservation conditions of the tape, as well as automatically correcting any errors that may have occurred during the digitization process. This technology offers an efficient solution for managing both small and large collections of digitized analog items, marking a substantial advancement in the preservation of audio documents.

INDEX TERMS Artificial intelligence, audio documents preservation, audio restoration, IEEE standard, musicological analysis, MPAI standard.

I. INTRODUCTION

> In the summer of 1937, Bird [Charlie Parker's nickname, one of the most important jazz musicians of the twentieth century - Ed. note] underwent a radical change musically. He got a job with a little band led by a singer. . .they played at country resorts in the mountains. Charlie took with him all the Count Basie records with Lester Young solos on them and learned Lester cold, note for note. . .when he came back, only two or three months later, the difference was unbelievable. (Gene Ramey [1])

The legendary Charlie Parker stands as a compelling illustration of how musical documents can shape the course of history. In the early stages of his career, Bird immersed himself in the records of Lester Young—a well-documented instance of a virtuoso jazz musician learning from another. This dynamic exchange gave rise to a new realm of musical improvisation, building upon the foundation laid by previous masters. Such creative evolution would have been inconceivable without easy access to the records of his predecessor. While the preservation of these cultural documents assumes paramount importance and continues to provide incredible opportunities for future generations, it also poses ongoing challenges and rewards, particularly with the advent of new technologies rooted in AI.

When preserving cultural audio heritage, it is fundamental to minimize loss of information. This is particularly significant when dealing with genres like African American jazz music, where a traditional score might not exist. Additionally,

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Alawneh.

in cases such as Tape Music, where the magnetic tape itself is an integral part of the artistic work, careful preservation is essential to safeguard the complete artistic experience.

Unlike recordings of live musicians, Tape Music is not captured on stage or in the studio for later storage and reproduction. Instead, it is composed directly with the assistance of electronic valves, transistors, and similar devices. Tape Music ''exists'' exclusively on magnetic tapes and can be reproduced and experienced through loudspeakers. The viability of these techniques in music composition emerged with the introduction of magnetic tape sound recording technologies. These advancements enabled direct human manipulation and (acoustic-)electromagnetic treatment of the recording medium. As a result, Tape Music captured the interest of prominent experimental and avant-garde creative minds in the mid-twentieth century. Notable figures such as Edgard Varèse (1883-1965), Olivier Messiaen (1908-1992), John Cage (1912-1992), Iannis Xenakis (1922-2001), Luigi Nono (1924-1990), Luciano Berio (1925-2003), Pierre Boulez (1925-2016), and Karlheinz Stockhausen (1928-2007) were drawn to explore its possibilities.

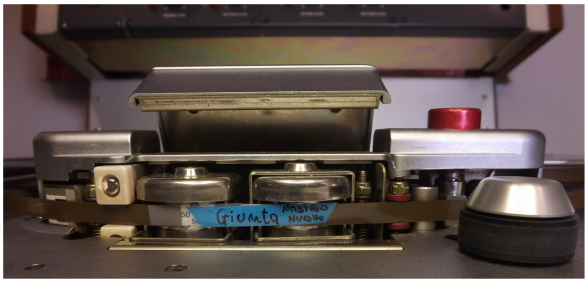


FIGURE 1. Example of markings on a tape splice.

In this context, the primary challenges arise from the relatively short life expectancy of this medium (less than 20 years), in contrast to the longevity of conventional tangible cultural heritage, which can endure for centuries or even millennia. This situation calls for a transition to re-recording these documents in digital form, ensuring their preservation over time. Relying solely on audio copies, however, is insufficient for preservation. Composers actively engaged with the tape, employing techniques such as cutting and pasting, adding annotations directly onto the medium (as illustrated in Fig. 1). Some clues are essential for live performances of the piece, while others, though not directly impacting performance, hold significance from a philological standpoint. Often, composers did not furnish a traditional score; therefore, the tape itself becomes the artwork—the culmination of the creative process. Preserving the tape in its entirety is crucial to safeguarding the essence of the artistic creation.

The integration of electronic and information technology into art has presented fresh challenges for archives and the preservation of cultural heritage. While technology serves as a catalyst for innovative forms of artistic creation [2], it also contributes to the accelerated deterioration and deprecation of formats, thereby reducing the lifespan and accessibility of new artworks.

Critical issues in this context include the compounded sheer volume of material yet to be digitized and the variety of adopted formats. The risks are twofold: firstly, the lack of expertise in digitization may result in the loss of information, and secondly, the limited storage space and bandwidth available pose significant hurdles in the challenge of preserving these archives for posterity [3]. Data analysis, which may occur years after digitization, may highlight error inconsistencies in audio documents that are no longer easily accessible in the original format [4]. Overall, digitization and data analysis represent a significant investment, requiring considerable resources in terms of time, money, and technical expertise [5], [6]. Naive implementations may jeopardize the proper preservation and accessibility of cultural heritage, making it unattainable.

These issues are addressed by the novel technology outlined in the Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) international standard on Audio Recording Preservation (ARP), later adopted as IEEE 3302-2022 [7]. Drawing extensively from contributions of the Centro di Sonologia Computazionale (CSC) at the University of Padua [8], [9], and leveraging considerable experience in music production, the ARP approach revolves around a well-defined scientific methodology anchored on two essential pillars. Firstly, it adopts a multidisciplinary approach that integrates perspectives from engineers, musicians, musicologists, composers, and archivists. Secondly, it upholds a profound commitment to philological accuracy in the development of digital tools. This encompasses the inclusion of metadata and ancillary information deemed crucial for the comprehensive completion of preservation copies [10]. The datasets gathered at CSC were assembled from over 3000 documents digitized through numerous preservation projects [11]. Notably, some of the most representative restored and digitized collections include those from Luciano Berio's archive (Paul Sacher Stiftung, featuring tape music and electronic music), the Luigi Nono Archive of Venice (tape music, electronic music), the Historical Archive of the Teatro Regio of Parma (encompassing opera, Western classical music, and pop/rock), the Tullia Magrini Archive (focused on ethnomusic), the Historical Archive of the Maggio Musicale Fiorentino (covering Opera and Western classical music), and the Fondazione Giorgio Cini of Venice (comprising speech and oral sources).

The structure of this manuscript is as follows: Section II briefly examines existing guidelines, relevant literature, and solutions for audio document preservation, along with the application of AI in music. Section III delves into the preservation methodology forming the core of the ARP standard. Section IV presents an in-depth overview of the foundational infrastructure and novel technology underpinning ARP; while Section V summarizes the performance results of the various technology components adopted, concluding with final remarks.

II. STATE OF THE ART

To ensure the preservation of audio documents, it's imperative to establish internationally shared guidelines that set preservation standards. These guidelines should encompass a regulatory framework that addresses the various stages of the preservation process, such as digitization, archiving, and long-term preservation of audio documents. International organizations, such as the International Association of Sound and Audiovisual Archives (IASA) [12], [13] and the International Federation of Library Associations and Institutions (IFLA) [14] have contributed to the definition of protocols aimed at guaranteeing the quality and sustainability of preservation practices for diverse physical media. Such guidelines, however, often inadequately describe the organization of digitized files, primarily focusing on the correct practices for preserving and managing analog documents. The organization of digitized files, encompassing metadata and storage formats, demands meticulous planning to guarantee the long-term accessibility and integrity of the contents. Hence, shared guidelines must comprehensively embrace the evolving digital landscape, addressing the challenges and best practices pertinent to preserving audiovisual documents in digital formats. To tackle these challenges, the CSC has proposed to MPAI a preservation methodology for audio documents based on [15], elaborated in detail in Section III.

In recent years, numerous archives and private institutions have embarked on extensive digitization projects. These endeavors, however, often encounter the challenge of digitizing a vast quantity of audiovisual documents within a relatively short time frame, which can easily lead to errors during the digitization process. The pressure to meet deadlines while handling such a large volume of materials may result in oversights or mistakes, ultimately compromising the quality and accuracy of the digitized records. Issues such as incomplete signal transfers, mislabeled files, or inadequate preservation of metadata may arise from these digitization efforts. Therefore, archives must allocate sufficient time and resources to minimize the risk of errors when digitizing analog materials. In this regard, AI proves to be an invaluable tool for improving both efficiency and accuracy in the digitization process, addressing challenges related to the quality of audio preservation and restoration [16], [17].

The integration of AI into the realm of music has begun to revolutionize how artists, composers, and producers approach music composition, creation, and production [18], [19]. AI's ability to analyze extensive musical datasets [20], discern patterns, and identify genres [21] empowers it to generate new sounds. Artists can harness AI for inspiration, crafting innovative melodies, and exploring unique sonic landscapes. Moreover, in music production, AI can optimize processes such as mixing and mastering [22], enhancing and streamlining the overall workflow. Some AI-based tools even facilitate automatic composition [23] of personalized musical accompaniments or real-time adaptation of music to listeners' emotions. Alongside these creative opportunities, however, concerns regarding ethics and artistic integrity arise, particularly regarding AI's potential to supplant the human element in music creation and compromise artistic authenticity. As of now, the use of AI in audio preservation has remained relatively limited, primarily focusing on speech restoration [24], [25], quality assessment of digitized audio [17], and, only very recently, enhancing historical recordings by utilizing inpainting [26] and bandwidth extensions [27]. Although AI has found significant applications in music creation, production, and enhancement, its adoption in the preservation of historical audio recordings and management of music archives is still in its very early stages. Nowadays, archives are undergoing a digital transformation and must harness automation, particularly through AI, to effectively manage data [28], [29].

The primary challenge lies in the intricacy and sensitivity of audio preservation, necessitating a meticulous and reverent approach to maintain the quality and authenticity of recordings over time. The research presented in this paper concentrates on investigating novel applications of AI in the conservation and restoration of audio recordings.

III. PRESERVATION METHODOLOGY

The adopted preservation methodology for audio documents is illustrated in Fig. 2. The initial step of the methodology involves photographing each audio document along with its corresponding box to document its preservation status. This information is crucial, as composers frequently made annotations on the boxes, covering not only details about the recording contents but also the adopted channel configuration, equalization curve, and recording speed. While the potential for misalignment between the content and what is reported on the boxes exists, it still serves as a valuable guideline. Visual inspection and pre-reading play an important role in diagnosing evident mechanical issues or identifying chemical/physical syndromes that may impact the tape. These steps provide essential insights before proceeding with the restoration process. Overall, the optimization of the carrier includes fixing old splices through the original tape, applying leader tape at the beginning, cleaning the surface to remove mold and dust, and implementing thermal treatment to address the Sticky-Shed/Soft Binder Syndrome [30].

The second step of the methodology concerns the A/D conversion. To digitize the audio content, it is fundamental to analyze and set the recording formats, digital parameters, and playback configuration correctly. Monitoring the entire A/D process is essential for preventing errors, such as the misinterpretation of channel configuration and recording speed. The digitization process is executed with high-quality converters and fully operational analog devices. Video documentation of the tape is also included to track any irregularities that may be present on its surface.

The final step of the preservation methodology involves data processing and the creation of Preservation and Access Copies. The Preservation Copy comprises a high-quality digital audio file with audio stored at a minimum of 24 bits precision and a sampling rate of 96 kHz, without any restoration

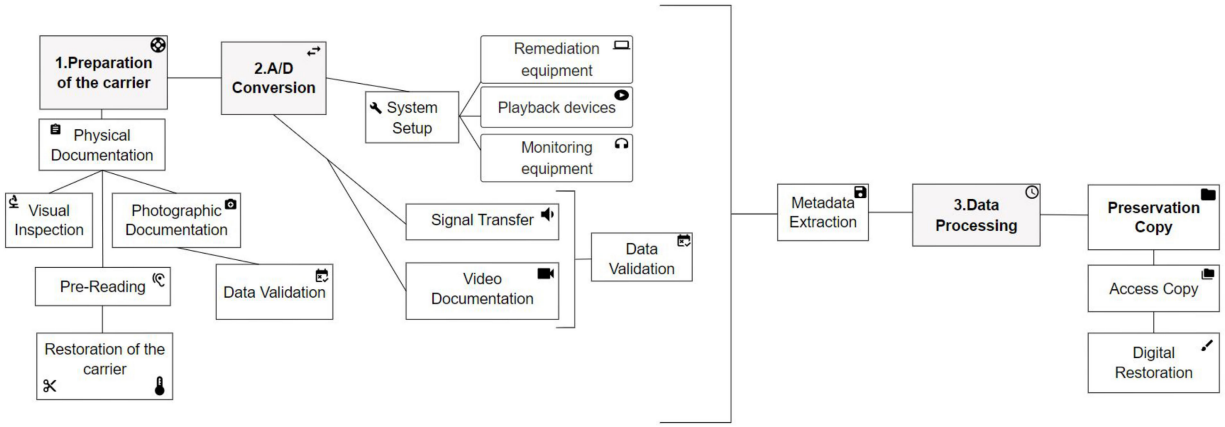


FIGURE 2. Diagram illustrating the preservation methodology.

or filters applied. In the case of multi-channel recordings, a separate audio file is provided for each channel. Multiple acquisitions are conducted when a tape is recorded at different speeds, resulting in separate audio files. In addition to digital audio files, the Preservation Copy incorporates photographic and video documentation, checksums, and scanned images of any accompanying documentation that may have been with the original item. Metadata gathering plays a central role in this process. Data regarding the original document, including brand, reel diameter, channel configuration, recording speed, etc., is stored in a dedicated database and summarized in a .pdf file, which is also included in the Preservation Copy. The Access Copy is typically provided in a compressed format, such as MPEG AAC [31], [32], to enhance portability.

One of the most common challenges in digitizing analog audio tapes is applying the correct equalization (EQ) curve. EQ curves were employed during recording as pre-emphasis to extend the dynamic range and enhance the Signal-to-Noise Ratio (SNR). During playback, inverse post-emphasis curves were applied to restore the original frequency response. Identifying the correct EQ curve is a significant challenge, particularly when dealing with tapes recorded in the early days of sound recording when there were no shared standards. In certain instances, different record labels and/or even individual technicians might have chosen to apply customized EQ curves to improve sound quality or tailor it to the technical characteristics of the equipment used at that time. The introduction of standard EQ curves such as IEC1 [33] (formerly known as CCIR) and IEC2 [34] (formerly known as NAB) has streamlined this process, yet it does not entirely resolve the issue. The digitization process remains complex, as it necessitates identifying and correcting the EQ curves to ensure an accurate and faithful reproduction of the original sound.

IV. CAE-ARP

The ARP technology is part of the MPAI-CAE international standard (aka IEEE 3302-2022[1]). MPAI/IEEE-CAE's pioneering specifications extend across a wide array of applications, including entertainment, communication, teleconferencing, gaming, post-production, preservation and restoration [35]. MPAI/IEEE-CAE encompasses four distinct use cases tailored to enhance the user's audio experience across various contexts, spanning different settings such as the home, car, on-the-go, and studio. The four use cases specified in the CAE standard are: 1) Emotion Enhanced Speech (EES); 2) Audio Recording Preservation (ARP); 3) Speech Restoration System (SRS); 4) Enhanced Audioconference (EAE). These examples highlight the versatility and comprehensive scope of MPAI/IEEE-CAE's innovative specifications, illustrating their adaptability to various contexts and their ability to address a broad spectrum of audio-related needs [35].

The foundational infrastructure enabling the implementation of MPAI-CAE is the MPAI AI Framework (AIF), specified in the MPAI-AIF/IEEE 3301-2022 standard.[2] This provides the operational backbone for executing AI Workflows (AIW), which are constructed from fundamental processing elements known as AI Modules (AIM). MPAI-CAE normatively defines the semantics and syntax of input and output data, the functions of the AIW and AIMs, as well as the connections between AIMs within an AIW. Interoperability is ensured by the ability to substitute an AIW or AIM implementation with a functionally equivalent one while maintaining correct input/output formats. MPAI-CAE's objective is to leverage this embedded structure to enhance user experiences in audio-related applications.

The CAE-ARP technology stands as a groundbreaking advancement in the accurate preservation of information found in open-reel audio tapes. Through this process, not only long-term preservation but also precise playback of the digitized recording is ensured, with the capability for restoration if needed. CAE-ARP leverages automated AI processes to extract crucial information from digitized audio files, facilitating the creation of preservation and access copies. Operating within the framework of the CAE-ARP

[1] standards.ieee.org/ieee/3302/11006/ Last accessed June 27, 2024.

[2] standards.ieee.org/ieee/3301/11096/ Last accessed June 27, 2024.

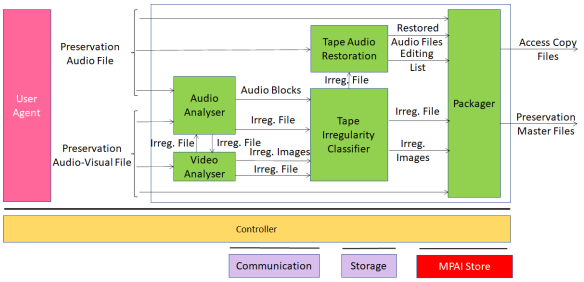


FIGURE 3. MPAI/IEEE-CAE ARP AI Workflow [7].

standard, archives can efficiently manage the wealth of information stored on tapes and their associated metadata. This standardized approach enables the automated preparation of content for immediate storage and/or utilization, streamlining the archival process and enhancing accessibility.

The ARP AIW and its various components are illustrated in Fig. 3. The architecture of the ARP standard comprises five AIMs designed to target and process distinct digital inputs [36]. These include the *Audio Analyzer*, *Video Analyzer*, *Tape Irregularity Classifier*, *Tape Audio Restoration*, and *Packager*. Each AIM plays a specific role in the overall processing and enhancement of audio content, contributing to the comprehensive capabilities of the ARP technology. Preserving audio assets recorded on analog media holds significant importance, considering the valuable information embedded in the magnetic tape of an open reel. In addition to the audio signal, this information may include annotations by the composer or technicians, multiple splices, and various irregularities like carrier corruptions, different-colored tapes, or diverse chemical compositions. The primary ARP objectives are long-term preservation and the creation of an access copy, which is restored if necessary, to facilitate accessibility and correct playback of the digitized recording. The ARP process takes as input the Preservation Audio File, which is generated through the digitization of the analog audio signal recorded on an open-reel tape with 24 bits per audio sample and a sampling rate of 96 kHz. Furthermore, an essential input to the ARP is the Preservation Audio-Visual File, which amalgamates a video file generated by a camera positioned at the playback head of the open-reel tape machine, see Fig. 4, with the audio content digitized at low resolution and synchronized with the video file. This comprehensive input contributes to the preservation process, ensuring that both audio and visual elements are accurately captured and maintained. The first AIMs in the ARP AIW (see Fig. 3), the Audio and Video Analyzers, analyze the audio/video signals in order to detect irregularities (such as Splice, Brands on tape, Start of tape, Ends of tape, Damaged tape, Dirt, Marks, Shadows, Wow and flutter, Play, pause and stop, Speed standard variation, Equalization standard variation, Signal backward) and create an Irregularity File and associated Audio and Image Files. These files feed into the Tape Irregularity Classifier AIM which classifies and selects the ones considered relevant. If the selected Irregularity was detected by the Video Analyzer, in addition to the selected Irregularity File, the corresponding Irregularity Images are also sent to the Packager AIM. The Tape Audio Restoration AIM uses the Irregularity File to identify and restore portions of the Preservation Audio File. It corrects speed, equalization and reading backwards errors in the Preservation Audio File and sends the Restored Audio Files and an Editing List to the Packager AIM. Finally, the Packager AIM collects the Preservation Audio Files, Restored Audio Files, the Editing List, the Irregularity File and corresponding Irregularity Images, and the Preservation Audio-Visual File, producing the Preservation Master Files. These files include the Preservation Audio File, the Preservation Audio-Visual File—where the original audio is replaced with a reduced-resolution version fully synchronized with the video—the set of Irregularity Images, and the Irregularity File. Additionally, Access Copy Files are generated, which contain the Restored Audio Files, Editing List, set of Irregularity Images, and the Irregularity File.

In the following sections, we will provide an in-depth description of the key technologies that are employed in implementing the various MPAI-CAE ARP AIMs. Before diving into the technical details, we will first offer an overview of MPAI, an international, non-profit organization committed to establishing standards for data coding based on AI.

A. MPAI

Established in September 2020 in Geneva, MPAI is an international standards organization committed to advancing the efficient utilization of data. Its mission involves developing technical specifications across diverse fields [37], encompassing Audio, Video, Neural Network Watermarking, Human-Machine Interaction, Avatars, Metaverse, Real and Virtual Environment Performance, Online Gaming, Financial Data, and Health. MPAI operates at the forefront of innovation, incorporating new technologies such as AI to shape standards that address the evolving landscape of data-related applications.

In its first three years of existence, MPAI has successfully developed and released 9 standards, all of which are publicly accessible on their website.[3] Notably, 5 of these standards have been officially adopted by the IEEE Standards Association (IEEE SA), showcasing MPAI's influence and contribution to the broader standards community. Additionally, MPAI continues to work on and has more standards in the pipeline, underscoring its ongoing commitment to advancing technological standards in various domains.

Apart from the technical specifications, MPAI has developed 3 reference software implementations, which are publicly accessible as open source. Furthermore, MPAI has released 2 conformance testing specifications publicly, offering valuable resources for evaluating adherence to standards. Lastly, MPAI has introduced 1 performance assessment specification, publicly available, that assesses factors such as

[3]mpai.community/standards/ Last accessed June 7, 2024.

robustness, replicability, reliability, and fairness, providing insights into the effectiveness and dependability of the implemented standard.

For MPAI-CAE ARP, the technical specifications as well as the conformance testing specifications and reference software are available through the MPAI website.[4] CAE ARP stands out as one of MPAI's most successful technologies, being recognized twice (in 2023 and in 2024) by the prestigious ''Neurons Awards Creativity AI Trophy'' at the World Artificial Intelligence Cannes Festival (WAICF[5]), the world's largest artificial intelligence event.[6]

In the following sections, a comprehensive technical description of the ARP AIMs is presented.

B. AIW ARCHITECTURE IMPLEMENTATION

The current infrastructure of ARP is implemented through a set of docker containers that interact via the Remote Procedure Call (RPC) protocol (using gRPC implementation[7]) and share a volume where to store the data. Each docker container hosts a server with a module implementation and exposes an API. This entire setup is managed by a client that sends organized requests to the services and processes their responses. More specifically, a common interface for all ARP AIMs has been defined via Protocol Buffer (aka Protobuf),[8] which exposes a main method, called ''work'', for starting data processing in each module and receiving responses based on their current state. The Protobuf interface is currently implemented in Python.

The code for this infrastructure implementation, along with its documentation, is also available on Gitlab.[9]

C. VIDEO ANALYZER AND AUDIO ANALYZER

The first two AIMs within the ARP AIW, as illustrated in Fig. 3, namely the Audio and Video Analyzers, are specifically designed to detect tape irregularities and accurately determine the exact moment at which these irregularities occur. The input to the ARP standard comprises two distinct files: a Preservation Audio File (PAF) obtained through the high-quality digitization of the analog audio, encompassing music, soundscape, or speech, recorded on the magnetic tape; and a Preservation Audio-Visual File (PAVF) created by a camera focused on the reading head of the magnetic tape machine (see Fig. 4). Together, these files contribute to the comprehensive preservation of both audio and visual aspects of the magnetic tape content. The PAF plays a crucial role in identifying any errors in the application of EQ curves, tape speed, and reverse audio [38] and it is then processed to be both restored and archived unaltered for philological purposes. In addition, the PAVF proves valuable for managing metadata associated with the carrier and providing additional information related to the context of the recording. This dual-input approach enhances the accuracy and thoroughness of the preservation process within the ARP standard.

1) VIDEO ANALYZER

A fundamental aspect of the Video Analyzer module is the precise identification of Regions of Interest (ROIs). Preliminary studies explored the application of background subtraction algorithms, utilizing prior information to segregate new elements from recurring ones. This approach, however, exhibited limitations, primarily manifesting as false positives due to variations in brightness, reel movement, and undesired artifacts [4]. In response to these challenges, a paradigm shift towards a scene framing-oriented approach was undertaken. It was observed that anomalies consistently exhibited a lack of vertical movement, appearing as small clusters of points within the frame. The strategic decision to focus on stationary elements, notably the capstan area (including pinch roller) and reading head (see Fig. 4), was made to serve as reference points for automatically identifying pixel regions vulnerable to irregularities [39]. The reading head, a pivotal component in tape recorders, proved to be a salient reference point due to its inherent stationary nature during playback. Coupled with the capstan area, these components facilitated the establishment of reliable stationary elements.

A thorough analysis of the central frame of the video associated with the tape, which is assumed to be indicative of a standard scenario, is carried out. After extracting the image (with deinterlacing applied in the case of older PAL videos), the positioning of ROIs is determined by seeking correspondence within grayscale capstan and reading head reference images. The process of element individuation is accomplished using the well-established Generalized Hough Transform [40], [41] and SURF [42] algorithms. The transformation from RGB space to grayscale relies on OpenCV [43], the library employed in the implementation for image processing, which defines the conversion rule as follows:

$$\text{RGB[A] to Gray: } Y \leftarrow 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B$$

Once ROIs are identified, the Video Analyzer proceeds to detect irregularities within the digitized magnetic tape images. This is accomplished by examining the absolute value of the differences between consecutive frames in grayscale. Specifically, the function defined by Equation 1 generates a new grayscale image of the tape based on the difference between consecutive input frames.

$$\mathbf{D}(i,j) = |\mathbf{C}(i,j) - \mathbf{P}(i,j)| \tag{1}$$

[4] mpai.community/standards/mpai-cae/ Last accessed June 27, 2024.

[5] www.worldaicannes.com/en Last accessed June 27, 2024.

[6] Link to the 2023 results: web.archive.org/web/20231210130015/www.worldaicannes.com/en/cannes-neurons; link to the 2024 results: www.worldaicannes.com/en/cannes-neurons Last accessed June 27, 2024.

[7] grpc.io/ Last accessed June 27, 2024.

[8] protobuf.dev/ Last accessed June 27, 2024.

[9] The CAE-ARP reference software can be found at the following link: experts.mpai.community/software/mpai-private/mpai-cae/arp/arp-workflow Last accessed on June 27, 2024. A corresponding repository has also been established on the University of Padua server publicly accessible via the link: gitlab.dei.unipd.it/csc-research/arp-aiw. Last accessed on June 27, 2024.

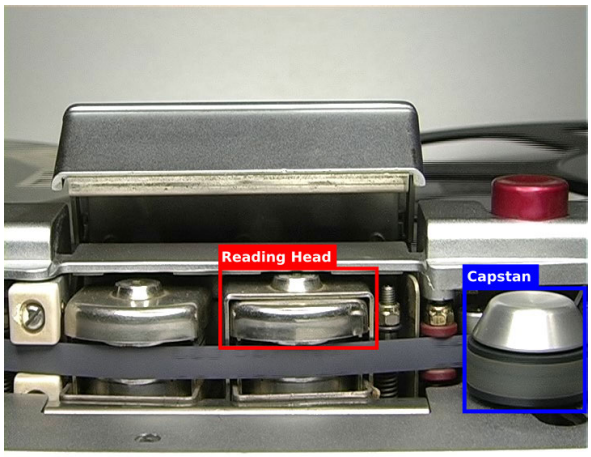


FIGURE 4. Tape machine reading head and capstan area.

where $i = 1, \ldots, n$ and n is the number of rows in the matrix, $j = 1, \ldots, m$ and m is the number of columns in the matrix, matrix $\mathbf{D}$ is the difference frame, $\mathbf{C}$ is the current frame matrix and $\mathbf{P}$ is the previous frame matrix in grayscale.

TABLE 1. Standard deviation S based on tape's speed.

Speed (ips)	Empirical standard deviation
30	2.25
15	2.5
7.5	2.6
3.75	2.75

Given the tape's potential color variations, the standard deviation of the difference image's color is considered to enhance algorithm stability in the presence of color and brightness fluctuations. Instead of merely using the mean, both mean and standard deviation of the pixel color in grayscale are computed using Equations 2 and 3, respectively. These metrics are then compared with the estimated standard deviation S based on the tape's speed, as summarized in Table 1, where the estimated standard deviation has been calculated through empirical tests. The tests involved examining 30 video frames without irregularities for each considered tape speed. For each frame, the mean color value was calculated, and subsequently, the mean and standard deviation were computed.

The following two equations are defined to calculate the mean and standard deviation of the pixel's color (in grayscale, so a single 8-bit channel) of an image $\mathbf{D}$ of dimension $m \times n$:

$$\mu = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \mathbf{D}(i,j) \tag{2}$$

$$\sigma = \sqrt{\frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \big(\mathbf{D}(i,j) - \mu\big)^2} \tag{3}$$

where $\mathbf{D}(i,j), m, n$ have the same meaning as defined in Equation 1.

When $\sigma < S$, it indicates that the colors in the image show minimal visible variation, implying that the difference image does not contain any anomalies. Conversely, for irregular difference images, the Otsu thresholding method [44], [45] is applied to define a threshold T for converting the image to a binary format using Equation 4 and Equation 5 below.

$$T = arg\,max_t\{\sigma_B^2(t) \cdot w_B(t) + \sigma_F^2(t) \cdot w_F(t)\} \tag{4}$$

where $\sigma_B^2(t)$ is the weighted variance of the class above the threshold, $w_B(t)$ is the probability of the class above the threshold, $\sigma_F^2(t)$ is the weighted variance of the class below the threshold and $w_F(t)$ is the probability of the class below the threshold. Thresholding is then applied to the image using Equation 5 below.

$$\mathbf{B}(i,j) = \begin{cases} 1 & \text{if } \mathbf{D}(i,j) \geq T \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $\mathbf{B}$ is the binary image obtained after thresholding.

A denoising function is then applied to highlight irregularity shapes through opening operations (erosion and dilation) with a 3×3 rectangular kernel, as described in Equation 6 below.

$$\mathbf{O}(i,j) = \mathbf{B}(i,j) \circ SE = (\mathbf{B}(i,j) \ominus SE) \oplus SE \tag{6}$$

where SE is the structuring element defined as a 3×3 square matrix and $\ominus/\oplus$ are the morphological image processing erosion/dilation operators.

After this process the matrix $\mathbf{O}(i,j)$ should contain a clearer shape of the irregularity. The count of white pixels in the resulting image is computed, and if it exceeds 5% of the image's area, a significant difference between consecutive frames is inferred. Equation 7 summarizes the decision process: if the area of difference in the image exceeds a fixed threshold, it is considered an irregularity.

$$I = \begin{cases} 1 & \text{if } \sum_{i=0}^{m} \sum_{j=0}^{n} \mathbf{O}(i,j) > \dfrac{m \times n \times 5}{100} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

At the end of the detection process, the frames in which an irregularity was found are stored as Irregularity Images along with their timestamp and a unique id in a JSON file called *IrregularityFile*.

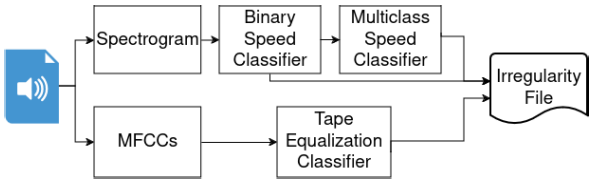


FIGURE 5. Block diagram of the audio analyzer module.

2) AUDIO ANALYZER

The Audio Analyzer AIM is responsible for carrying out a spectral analysis of the Preservation Audio File, identifying playback equalization requirements, detecting tape speed

errors and computing the cross-correlation between the high-quality preservation audio track and the lower-resolution audio track in the associated video file. This process is vital for achieving synchronization between the Preservation Audio File and the Preservation Audio-Visual File, ensuring the alignment and coherence of the audio components during the preservation process. Building on the approach outlined in [36], which utilized a single classifier to identify all audio irregularities, this paper introduces a novel method that divides signal classification into three distinct phases. In the first two phases, machine learning methods are employed to respectively recognize equalization curves and reading/writing tape speeds. The third phase focuses on computing the signal cross-correlation between the audio and video components. This approach enhances efficiency and accuracy in handling the diverse aspects of signal irregularities detected during the preservation process.

Expanding on prior research detailed in [46] and [47], our current spectral analysis utilizes the first 13 Mel-Frequency Cepstral Coefficients (MFCCs) to represent audio for equalization classification (see lower part of Fig. 5). This representation proves to be a suitable approximation of the signal for the task of equalization classification. Given the significant variability of the content recorded on the tape, we decided to adhere to prior research by concentrating the analysis solely on segments of silence on the tape, i.e., portion of the signal with intensity below −50 dBFS. Silence, as defined in [47], where empirical findings indicate that audio signals with intensities ranging from −50 to −63 dBFS represent silence between spoken words, from −63 to −69 dBFS represent noise resulting from the recording head without input, and below −69 dBFS represent noise from sections of pristine tape, tends to produce more consistent results when employed for classification, as opposed to analyzing the entire signal present on the tape. The dataset under analysis comprises 9328 audio segments, each lasting 500 ms and featuring intensities below −50 dBFS. From these segments, the first 13 Mel-Frequency Cepstral Coefficients (MFCCs) are extracted and then normalized. These segments are part of a collection of 25 audio tape recordings, designed to encompass every possible configuration of IEC1 and IEC2 equalization curves at different tape speeds. These tapes were digitized at a sampling rate of 96 kHz and a sample precision of 24 bits. The classification process involved the dataset preparation, the model selection and an assessment over the validation set. The performance of all the models has been validated and tested using a K-Fold cross-validation with $K = 5$ and a 80, 20 train-test dataset split. Numerous experiments were conducted performing grid search cross-validation over K-Nearest Neighbour (KNN), Random Forest Classifiers (RFC), Support Vector Machines (SVM) and Gradient Boosting (XGB) to tune the algorithms hyperparameters and a Deep Neural Networks (DNN) whose structure is described in the following paragraphs. All models were trained and validated using the same dataset to select the most effective model. As illustrated in Table 2, the best results over the test set were achieved using DNNs.

TABLE 2. Models scores over validation set in EQ recognition.

Model	Accuracy	Precision	Recall
KNN	0.80	0.79	0.76
RFC	0.84	0.86	0.82
SVM	0.88	0.88	0.87
XGB	0.87	0.87	0.87
DNN	0.90	0.90	0.90

Fig. 6 presents a comparative analysis of the five classification models, the performance metrics displayed include training and validation accuracy as well as loss, plotted against key hyperparameters for each model. In the KNN model, the training accuracy remains at a perfect score of 1.00 across all neighbor counts, indicating potential overfitting. The validation accuracy hovers around 0.80, showing minor improvements with increasing neighbors, while the validation loss decreases significantly before stabilizing. The SVM model shows that by increasing the regularization parameter C we have an increase in both training and validation accuracy, reaching a peak at around 0.95 and 0.88 respectively. The corresponding loss metrics reflect this trend, stabilizing at lower values as C increases. The Random Forest model maintains a perfect training accuracy of 1.00 regardless of the number of estimators, while the validation accuracy fluctuates slightly around 0.84. This model's loss metrics remain consistently low, suggesting limited generalization improvement with additional estimators. XGB starts with high training accuracy and shows marginal improvements over iterations. However, its validation accuracy improves only slightly, leveling off around 0.87, with the validation loss indicating potential overfitting after initial iterations. Notably, the DNN model exhibits the highest validation accuracy, stabilizing around 0.90, with a corresponding decrease in validation loss over epochs. The training metrics for DNN also show rapid improvement, achieving around 0.95 accuracy, which suggests a strong ability to generalize to unseen data. Overall, the DNN model outperforms the other models, demonstrating its superiority for the given classification task, as evidenced by its balance between high validation accuracy and stable loss metrics.

The DNN architecture used in this study was designed and tuned following an empirical pre-evaluation stage, during which the number of layers, neurons, and activation functions were selected based on initial observations. Leaky ReLU has been chosen over standard ReLU as it showed better performance during this phase. The hidden layers follow a typical structure in which the number of neurons decreases progressively with depth. However, instead of consistently reducing the number of neurons in each layer, it was decided to repeat the number of neurons in consecutive layers, which yielded better results in the initial evaluation. Given the strong performance of the chosen architecture, further detailed

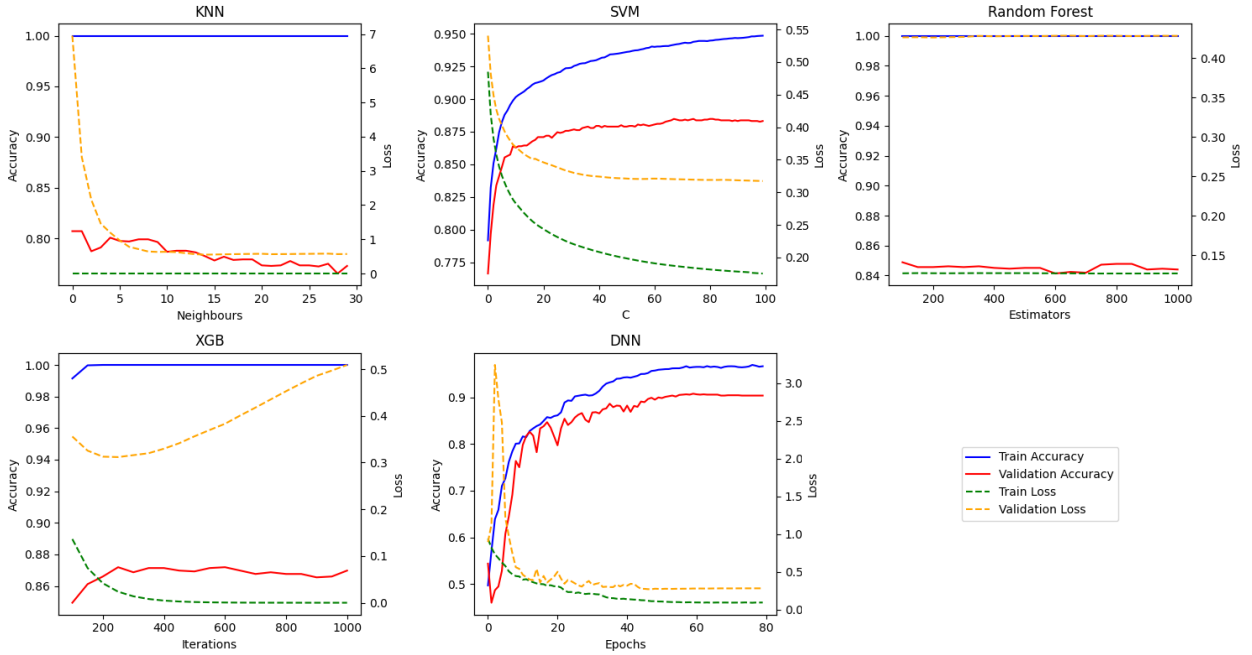


FIGURE 6. Training and validation accuracy and loss for five different EQ classifier models versus key hyperparameters.

analysis and grid search for hyperparameter optimization were deemed unnecessary.

The final neural network architecture consists of eleven fully connected layers implementing a standard feedforward multilayer perceptron neural network as described in [48]. The input layer matches the dimensions of the input data, followed by two layers each containing 128 neurons, two layers with 64 neurons, two layers with 32 neurons, two layers with 16 neurons, and two layers with 8 neurons. A dropout layer is included to mitigate overfitting, followed by an output layer comprising three neurons. The network encompasses approximately 37,000 parameters in total. Activation functions employed within layers containing 128, 64, 32, 16, and 8 neurons consist of Leaky ReLU, complemented by batch normalization techniques to counteract potential issues arising from vanishing gradients. A detailed layer by layer summary of the neural network parameters can be found in Table 3.

The model was trained using the Adam optimizer with an adjustable learning rate starting from 0.01 that was halved with a patience of 5 epochs to prevent local minimum stasis. The batch size was set to 256, and the training was conducted for 100 epochs. Early stopping was implemented with a patience of 10 epochs to prevent overfitting. Dropout with a rate of 0.3 was applied after the penultimate layer to further mitigate overfitting. Categorical cross-entropy was used as the loss function.

Loss and accuracy metrics derived from the validation and training datasets are presented in Fig. 6. Notably, the absence of conspicuous overfitting is evident, as indicated by the consistent behavior of the validation curve across epochs. However, it is discernible from the figure that the model performance tends to plateau after a certain number of epochs. Consequently, the number of training epochs has

TABLE 3. EQ classifier DNN model architecture layer by layer. μ is the momentum parameter of the batch normalization layers, while α is the slope of the leaky ReLU activation function for negative argument.

Name	Type	Output size	Parameters
dense 1	Dense	128	
batch norm 1	Batch Normalization		μ: 0.99
leaky 1	Leaky ReLU		α: 0.2
dense 2	Dense	128	
batch norm 2	Batch Normalization		μ: 0.99
leaky 2	Leaky ReLU		α: 0.2
dense 3	Dense	64	
batch norm 3	Batch Normalization		μ: 0.99
leaky 3	Leaky ReLU		α: 0.2
dense 4	Dense	64	
batch norm 4	Batch Normalization		μ: 0.99
leaky 4	Leaky ReLU		α: 0.2
dense 5	Dense	32	
batch norm 5	Batch Normalization		μ: 0.99
leaky 5	Leaky ReLU		α: 0.2
dense 6	Dense	32	
batch norm 6	Batch Normalization		μ: 0.99
leaky 6	Leaky ReLU		α: 0.2
dense 7	Dense	16	
batch norm 7	Batch Normalization		μ: 0.99
leaky 7	Leaky ReLU		α: 0.2
dense 8	Dense	16	
batch norm 8	Batch Normalization		μ: 0.99
leaky 8	Leaky ReLU		α: 0.2
dense 9	Dense	8	
batch norm 9	Batch Normalization		μ: 0.99
leaky 9	Leaky ReLU		α: 0.2
dense 10	Dense	8	
batch norm 10	Batch Normalization		μ: 0.99
leaky 10	Leaky ReLU		α: 0.2
dropout	Dropout	8	rate: 0.3
dense 11	Dense	3	
softmax	Softmax		

been constrained to 50 epochs to optimize computational efficiency while maintaining satisfactory performance levels.

A final evaluation of the model on the test set can be conducted by examining the confusion matrix in Fig. 7 and Fig. 8. The results indicate that, generally, when the classifier produces an incorrect result, it tends to misidentify digitized tapes with different pre- and post-emphasis equalization applied during recording and playback as correctly processed tapes.

It is important to note that, in this phase, the individual 500 ms audio segments are treated separately. The subsequent modules of the ARP AIW are responsible for aggregating the information, which should help mitigate the issue of incorrect classification results later in the ARP AIW.

In fact, a single classification error in the midst of a sequence of correctly classified segments does not significantly impact the overall outcome of the preservation process.

The audio playback speed detection algorithm is based upon the analysis of spectrogram images of various audio files (see upper part of Fig. 5). These images provide a visual representation of the variations of the audio spectrum over time.

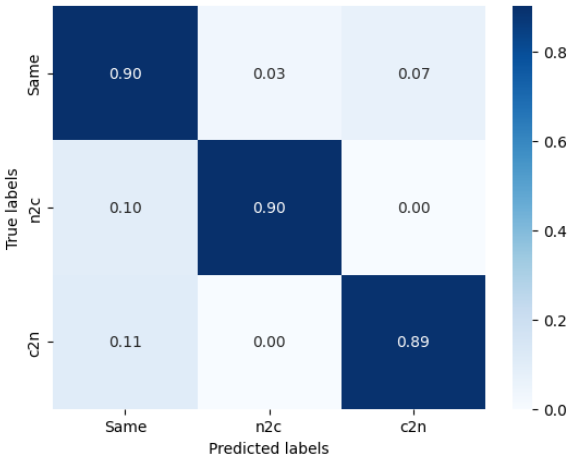


FIGURE 7. EQ classifier normalized confusion matrix evaluating performance over the test set, "same" label means that writing and reading equalization were the same, n2c means that the tape has been read in IEC1 and written in IEC2, while c2n is the opposite.

To assess the speed detection algorithm, a dataset of 300 audio files was selected to encompass sounds with a wide variety of spectral characteristics. The audio files are categorized into different groups: those with correct playback speed and those with speed variations. Since tape speed varies by factors of 2, the speed changes have been labeled with relative changes rather than absolute values. In other words, an audio file with a change in speed from 3.75 ips to 7.5 ips is in the same category (double) as an audio file with a change in speed from 7.5 to 15 ips. The same applies for halving the speed. Since the most common speeds used in professional audio recordings are 3.75, 7.5, 15 ips, the identified categories are: double when the tape is read at double the writing speed, half when the tape is read at half the writing speed writing,

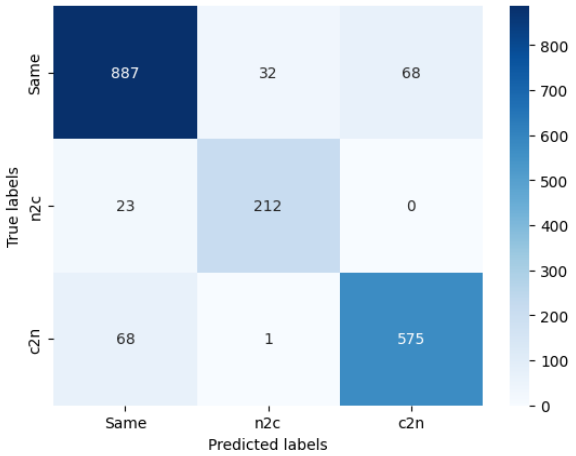


FIGURE 8. EQ classifier unnormalized form confusion matrix evaluating performance over the test set, "same" label means that writing and reading equalization were the same, n2c means that the tape has been read in IEC1 and written in IEC2, while c2n is the opposite.

quarter if the tape is read at a quarter of the writing speed and quadruple as the reciprocal of the previous case.

The audio spectrogram images, similar to those shown in Fig. 9, were extracted separately for each audio channel. Initially, the spectrogram of the entire audio file was generated and then split into chunks of 1 s duration in grayscale at 8 bits, with dimensions of 256 × 128 pixels. Additionally, durations of 250 ms (64 × 128 pixels) and 500 ms (128 × 128 pixels) were tested, but they yielded poorer results as displayed in Table 4.

To minimize errors, the speed classification task has been divided into two stages. The first is a binary classifier that determines if the tape is correctly played or not. The second classifier is activated only in case of anomaly detection and is used to determine the error class (double, half, quarter, quadruple). Both stages utilize Convolutional Neural Network (CNN) models. The two models share the same structure, apart from the last layer responsible for outputting the predicted label. Each model comprises three convolutional layers with a kernel size of 7 × 7 and a ReLU activation function. The layers differ in the number of neurons, progressively increasing from 8 to 16, and finally to 32. Each convolutional layer is followed by a max pooling layer with kernel 5 × 5. After these layers, a global average pooling is performed, and its output connects to a dense layer of size 32, always with a ReLU activation function. The final layer consists of 2 neurons for the binary classifier, activated with the sigmoid function, and 4 neurons for the multi-class classifier, using the softmax activation function. A detailed schema of the convolutional neural network architecture can be found in Fig. 10.

The CNN architecture was chosen following an empirical pre-evaluation phase, during which various configurations of layers, neurons, kernel types, and activation functions were explored. Like the approach used for the DNN, these initial tests identified the current configuration as offering the best

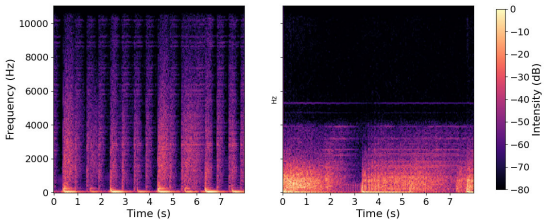


FIGURE 9. Audio spectrogram images showing drum tape recordings played back at the correct speed, 15 ips (left), and at a quarter of the correct speed, 3.75 ips (right).

TABLE 4. Model scores for audio segments with varying durations.

Input Size	F1-score	Precision	Recall
Binary model			
64 × 128	0.77	0.80	0.79
128 × 128	0.94	0.94	0.94
256 × 128	0.96	0.96	0.96
Multiclass model			
64 × 128	0.40	0.71	0.45
128 × 128	0.97	0.97	0.97
256 × 128	0.98	0.98	0.98

balance between performance and accuracy. The architecture closely resembles standard CNN designs, incorporating the typical progression of convolutional layers with increasing neurons, max pooling for dimensionality reduction, and ReLU activations for nonlinearity. This effective combination of simplicity and performance rendered further analysis and tuning unnecessary, as the selected setup already produced excellent results.

The model was trained using the Adam optimizer with a learning rate of 0.01. The batch size was set to 64, and the training was conducted for 50 epochs. Early stopping was implemented with a patience of 10 epochs to prevent overfitting. Dropout with a rate of 0.3 was applied after the penultimate layer to further mitigate overfitting. Categorical cross-entropy was used as the loss function in the case of multiclass model, while binary cross-entropy was used for the binary model. To evaluate the performance of the playback speed detection models, extensive testing and validation were conducted using a diverse set of audio recordings with known speed variations. Both stages achieved impressive accuracy scores, with precision, recall, and F1-score values exceeding 98% on the validation set, while the test set gave a noticeable performance drop of around 85%.

As illustrated in Fig. 5, the output of the Audio Analyzer consists of the Irregularity File, which includes detected irregularities metadata collected in a single JSON file.

D. TAPE IRREGULARITY CLASSIFIER

The Tape Irregularity Classifier is designed to verify and merge, if necessary, the irregularities received as input from the Audio and Video Analyzer AIMs through the irregularity files (see Fig. 14). It utilizes a CNN model tailored specifically for analyzing irregularities extracted from video data and consolidates the classifications related to the individual audio chunks.

In the current implementation the Classifier CNN model is trained to recognize three classes of Irregularities on the tape: Splices, Brands, and Shadows. While splices constitute the primary focus in video analysis, brands and shadows serve to accommodate detections made by the video analyzer that are not strictly content-related. Brands marks on the tape, though prevalent, lack relevance to audio and metadata content. While they recur consistently throughout the tape, the brand information is stored only once, with subsequent brand images segregated into a distinct folder and excluded from the irregularity file. Shadows, conversely, may arise due to specific lighting conditions or irregularities on the tape surface. The latter scenario is of paramount importance in preservation endeavors, necessitating the retention of shadows as irregularities in the Irregularity File to prevent information loss.

Initially, our approach involved leveraging transfer learning by fine-tuning a pre-trained model based on EfficientNet B0 [49], which has shown effectiveness across various computer vision tasks. The classifier architecture consists of an input layer accepting 224 × 224 pixel color images, followed by convolutional layers with frozen weights responsible for extracting relevant features from the input data. A Global Average Pooling layer combined with a dense layer forms the output, with the number of neurons in the dense layer corresponding to the number of irregularity classes to be recognized (in this instance, $n = 3$). Notably, the EfficientNet model accepts images with 8-bit values instead of pixel values scaled between 0 and 1.

TABLE 5. Models scores over test set in Irregularity images recognition.

Model	Accuracy	Precision	Recall
EfficientNet B0	0.91	0.90	0.91
Custom CNN for Speed Classifier	0.96	0.96	0.96

In addition to the initial approach that leveraged transfer learning with the EfficientNet architecture, we conducted another experiment using an adapted version of the CNN architecture designed for the speed detection algorithm as detailed in the preceding section. The former showed excellent results in both training and validation, achieving approximately 99% accuracy. Its performance, however, dropped by approximately 10% on the test set, suggesting potential overfitting. In contrast, the latter model, with slightly lower accuracy during training and validation (around 97%), demonstrated better generalization on the test set with an accuracy of nearly 96%. The summarized results can be found in Table 5.

This indicates that the custom CNN architecture provides more stable performance, underscoring its potential for broader applicability in analyzing tape irregularities. In fact, as shown in Fig. 11, EfficientNet's accuracy on the validation set tends to overfit from the early epochs. In contrast, the accuracy scores of the custom CNN model for the Speed Classifier exhibit slightly more variability across epochs

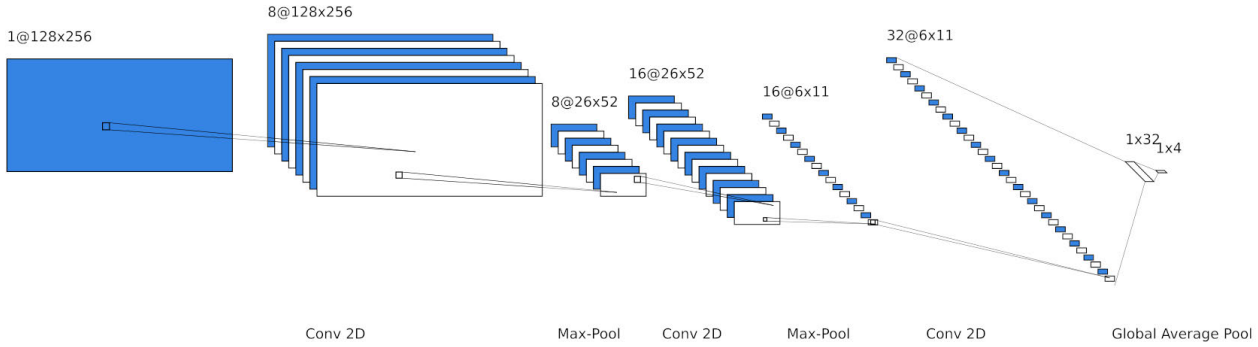


FIGURE 10. CNN model architecture for speed classification.

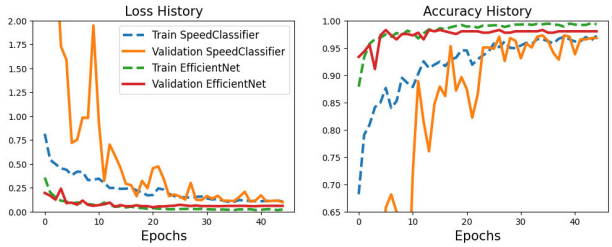


FIGURE 11. Tape irregularity classifier training curve.

but demonstrate consistent improvement as the number of training epochs increases.

The training dataset comprises Irregularity Images corresponding to each class. The training set is partitioned into 80% for training and 20% for validation. Subsequently, the model undergoes testing on Irregularity Images detected from recently digitized magnetic audio tapes, ensuring evaluation on a distinct set of images not encountered during training. This approach facilitates assessing the model's ability to generalize to unseen data and accurately identify irregularities across disparate sources. The dataset, overall, exhibits slight class imbalance, with splices comprising approximately 800 images, while brands and shadows each contain around 600 images. Following 20 epochs of training, the model demonstrates notable efficacy, achieving a 97% accuracy over the validation dataset, indicative of its adeptness in discerning patterns associated with splices, brands, and shadows within the provided dataset.

Looking at the confusion matrix of the model (see Fig. 12 and Fig. 13), no particular class imbalances emerge in the results, Brands and Splices are the classes that reap the greatest successes, while Shadows are occasionally confused with the other classes.

Upon completion of the irregularities selection and aggregation process, the Tape Irregularities Classifier AIM shares the chosen Irregularity Images (along with their metadata) with the Packager. Simultaneously, the aggregated audio irregularities are transmitted to the Tape Audio Restoration AIM to allow the generation of a restored Access Copy. The specific data flow is depicted in Fig. 14.

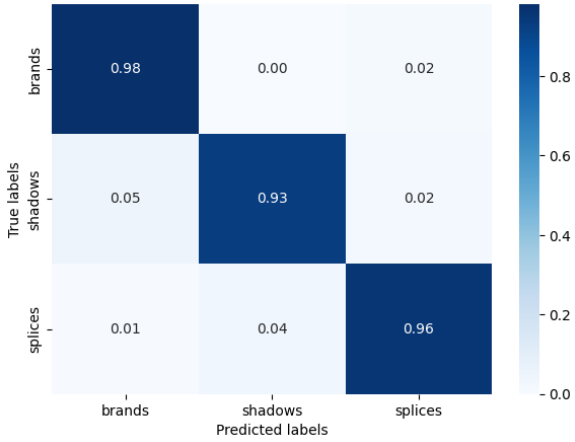


FIGURE 12. Tape irregularity classifier normalized confusion matrix evaluation performance over test set.

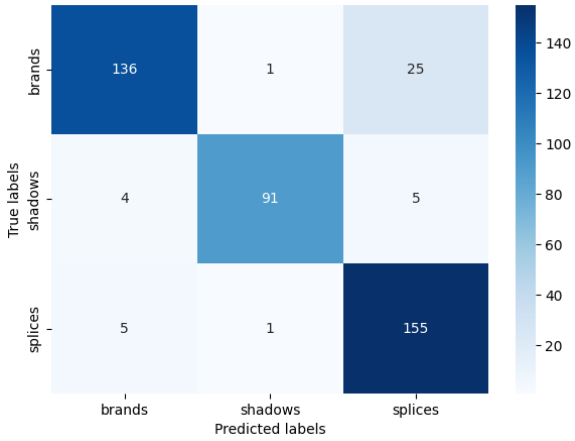


FIGURE 13. Tape irregularity classifier unnormalized form confusion matrix evaluation performance over test set.

E. TAPE AUDIO RESTORATION

The Tape Audio Restoration AIM tackles audio irregularities by rectifying time-reversed segments, conducting sampling rate conversion for accurate playback speed, and applying equalization correction curves in areas identified by the Irregularity File. Inputs to this AIM include the Irregularity

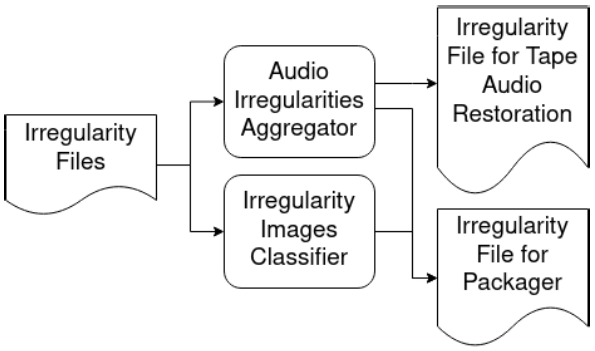


FIGURE 14. Tape irregularity classifier data flow.

File and the Preservation Audio File. The resulting Restored Audio Files guarantee the appropriate playback of the original audio content.

The correction of speed happens through resampling where the new sample rate (sr) is computed as follows: $sr_{new} = sr_{old}$ speed writing speed reading. Then, the equalization correction is performed by applying a filter composed of the inverse of the incorrect curve used during the digitization process and the correct equalization curve (the overall workflow is described in [50]):

$$sr_{new} = sr_{old} \frac{speed_{writing}}{speed_{reading}} \tag{8}$$

Finally, the correction of reversed audio is implemented by simply reversing in time the order of the audio samples based on the starting and ending points of the detected irregularity.

F. PACKAGER

Once all the metadata, magnetic tape images, and restored audio segments are obtained, it is crucial to provide easily accessible and searchable files. The *Packager* AIM is responsible for this task, receiving all the materials generated by the other AIMs and organizing them into folders. One folder is designated for storing a philological copy of the tape (audio and video, synchronized, in high resolution without any corrections) along with the metadata of the identified irregularities (Preservation Master Files). Another folder is created to provide access to a (potentially) restored audio file, sometimes in a compressed format, making it easy to download and play on various devices (Access Copy Files).

V. CONCLUSION

This paper presented the innovative technology integrated into the MPAI/IEEE CAE-ARP standard, showcasing exceptional results in the digital preservation and restoration of magnetic tapes. The test outcomes for the Tape Irregularity Classifier AIM reveal that custom neural network architectures remain relevant and can effectively compete with established CNNs. This indicates that task-specific models can outperform more general ones in certain scenarios. Following rigorous training and testing, the Tape Irregularity Classifier achieves an impressive 96% accuracy on the test dataset, demonstrating its ability to generalize to new data and accurately detect irregularities across various sources. Moreover, the classifier's robustness is evident in its equitable handling of class imbalances within the dataset, ensuring unbiased recognition of different irregularity types.

Similarly, the Audio and Video Analyzer AIMs demonstrate exceptional performance in identifying and characterizing tape irregularities. The Video Analyzer, employing sophisticated techniques such as ROI identification and difference frame analysis, accurately detects anomalies while mitigating false positives arising from environmental variations. By focusing the video on stationary elements in the tape playback system and employing advanced image processing algorithms, the Video Analyzer ensures reliable identification of irregularities, essential for preserving valuable annotations stored on magnetic tapes. The Audio Analyzer AIM, employing meticulous feature extraction and model selection, attains exceptional performance in identifying equalization curves and playback speeds. This highlights the model's ability to classify audio irregularities with high accuracy, thereby aiding in the comprehensive preservation of audio content archived on magnetic tapes.

Standards and their implementations evolve over time following the new possibilities provided by emerging technologies. The MPAI/IEEE community is continually reassessing and updating its standards by incorporating new state-of-the-art approaches. In the specific case of the MPAI/IEEE CAE-ARP standard, new pre-trained models could further enhance the overall performance of audio preservation and restoration tasks, leading to improved accuracy and efficiency in handling diverse audio signals. We are currently furthering the ARP development by designing new algorithms to automatically identify and correct backwards-recorded sections in digitized open-reel tapes, as well as improving the detection of speed variations by including lower and higher playback speeds. Future work will also target the overall performance of the algorithms, towards the development of a real-time system. Last but not least, an expanded dataset featuring additional types of irregularities (for both audio and video content) and various musical genres (such as pop music, opera, and non-Western traditional music) could positively impact the system. This would enable broader use of the tool while enhancing its reliability and overall performance.

In conclusion, the advanced technology integrated into the ARP standard offers effective solutions for detecting and characterizing irregularities, contributing to the preservation of valuable audio records. Adoption of the CAE-ARP standard empowers archives to efficiently identify and rectify errors in various audio files, improving the quality and accuracy of preservation and access copies, streamlining archiving processes, and ensuring interoperability through standardized digital file formats. The ARP standard marks a significant advancement in the preservation of audio cultural heritage, ensuring its enduring accessibility and usability, and paving the way for future developments in this critical field.

ACKNOWLEDGMENT

This work is partially supported by the SYCURI Project, funded by the University of Padova in the Program "World Class Research Infrastructure. The work of Marina Bosi was supported by the Center for Computer Research in Music and Acoustics (CCRMA), Stanford University."

REFERENCES

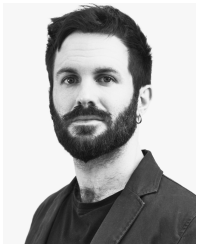
[1] C. Woideck, *Charlie Parker: His Music and Life*. Ann Arbor, MI, USA: Univ. Michigan Press, 2020.
[2] R. Wingström, J. Hautala, and R. Lundman, "Redefining creativity in the era of AI? Perspectives of computer scientists and new media artists," *Creativity Res. J.*, vol. 36, no. 2, pp. 177–193, Apr. 2024.
[3] F. Rumsey, "Will you be mine forever? Audio archiving multitracks, and 90s digital," *J. Audio Eng. Soc.*, vol. 68, no. 4, pp. 304–307, 2020.
[4] C. Fantozzi, F. Bressan, N. Pretto, and S. Canazza, "Tape music archives: From preservation to access," *Int. J. Digit. Libraries*, vol. 18, no. 3, pp. 233–249, Sep. 2017.
[5] M. M. C. Shekar and J. H. L. Hansen, "Historical audio search and preservation: Finding waldo within the fearless steps Apollo 11 naturalistic audio corpus [applications corner]," *IEEE Signal Process. Mag.*, vol. 40, no. 3, pp. 30–38, May 2023.
[6] E. Borin and F. Donato, "Financial sustainability of digitizing cultural heritage: The international platform Europeana," *J. Risk Financial Manage.*, vol. 16, no. 10, p. 421, Sep. 2023.
[7] *IEEE Standard Adoption of Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) Technical Specification Context-Based Audio Enhanced (CAE) Version 1.4*, IEEE Standard 3302-2022, 2023, pp. 1–94.
[8] S. Canazza and G. De Poli, "Four decades of music research, creation, and education at Padua's Centro di Sonologia Computazionale," *Comput. Music J.*, vol. 43, no. 4, pp. 58–80, 2020.
[9] S. Canazza, G. De Poli, and A. Vidolin, "Gesture, music and computer: The Centro di Sonologia Computazionale at Padova University, a 50-year history," *Sensors*, vol. 22, no. 9, p. 3465, 2022.
[10] S. Verde, N. Pretto, S. Milani, and S. Canazza, "Stay true to the sound of history: Philology, phylogenetics and information engineering in musicology," *Appl. Sci.*, vol. 8, no. 2, p. 226, 2018.
[11] N. Pretto, A. Russo, F. Bressan, V. Burini, A. Rodà, and S. Canazza, "Active preservation of analogue audio documents: A summary of the last seven years of digitization at CSC," in *Proc. 17th Sound Music Comput. Conf. (SMC)*, Turin, Italy, 2020, pp. 394–398.
[12] K. Bradley, ed., *Guidelines on the Production and Preservation of Digital Audio Objects (IASA-TC 04)*. Amsterdam, The Netherlands: International Association of Sound and Audiovisual Archives, 2006.
[13] W. Prentice and L. Gaustad, eds., *The Safeguarding of the Audiovisual Heritage: Ethics, Principles and Preservation Strategy (IASA-TC 03)*. Amsterdam, The Netherlands: International Association of Sound and Audiovisual Archives, 2017.
[14] M. Miller and S. Gherdevic, *Guidelines for Audiovisual and Multimedia Collection Management in Libraries* (IFLA). The Hague, The Netherlands: International Federation of Library Associations and Institutions, 2017.
[15] F. Bressan and S. Canazza, "A systemic approach to the preservation of audio documents: Methodology and software tools," *J. Electr. Comput. Eng.*, vol. 2013, no. 1, pp. 1–21, 2013.
[16] N. Wallaszkovits, "Between standards and arts: Digitisation and restoration of audio material—A balancing act between authenticity and manipulation?" *J. New Music Res.*, vol. 47, no. 4, pp. 285–290, Aug. 2018.
[17] A. Ragano, E. Benetos, and A. Hines, "Automatic quality assessment of digitized and restored sound archives," *J. Audio Eng. Soc.*, vol. 70, no. 4, pp. 252–270, May 2022.
[18] M. Mansoori and R. Murali, "A systematic survey on music composition using artificial intelligence," in *Proc. Int. Conf. Advancement Technol. (ICONAT)*, Jan. 2022, pp. 1–8.
[19] K. Tatar, P. Ericson, K. Cotton, P. T. N. Del Prado, R. Batlle-Roca, B. Cabrero-Daniel, S. Ljungblad, G. Diapoulis, and J. Hussain, "A shift in artistic practices through artificial intelligence," *Leonardo*, vol. 57, no. 3, pp. 293–297, Jun. 2024.
[20] L. Moysis, L. A. Iliadis, S. P. Sotiroudis, A. D. Boursianis, M. S. Papadopoulou, K. D. Kokkinidis, C. Volos, P. Sarigiannidis, S. Nikolaidis, and S. K. Goudos, "Music deep learning: Deep learning methods for music signal processing—A review of the state-of-the-art," *IEEE Access*, vol. 11, pp. 17031–17052, 2023.
[21] H. Cai, T. Pu, Y. Luo, and X. Zhou, "Music genre prediction based on machine learning," in *Proc. IEEE Int. Conf. Artif. Intell. Ind. Design (AIID)*, May 2021, pp. 198–201.
[22] H. Zhou, F. Yu, and X. Wu, "Audio mixing inversion via embodied self-supervised learning," *Mach. Intell. Res.*, vol. 21, no. 1, pp. 55–62, Feb. 2024.
[23] M. Catak, S. AlRasheedi, N. AlAli, G. AlQallaf, M. AlMeri, and B. Ali, "Artificial intelligence composer," in *Proc. Int. Conf. Innov. Intell. Informat., Comput., Technol. (3ICT)*, Sep. 2021, pp. 608–613.
[24] J.-W. Hwang, R.-H. Park, and H.-M. Park, "Efficient audio-visual speech enhancement using deep U-Net with early fusion of audio and video information and RNN attention blocks," *IEEE Access*, vol. 9, pp. 137584–137598, 2021.
[25] T. Saeki, S. Takamichi, T. Nakamura, N. Tanji, and H. Saruwatari, "SelfRemaster: Self-supervised speech restoration for historical audio resources," *IEEE Access*, vol. 11, pp. 144831–144843, 2023.
[26] F. Miotello, M. Pezzoli, L. Comanducci, F. Antonacci, and A. Sarti, "Deep prior-based audio inpainting using multi-resolution harmonic convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 113–123, 2024.
[27] E. Moliner and V. Välimäki, "BEHM-GAN: Bandwidth extension of historical music using generative adversarial networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 943–956, 2023.
[28] G. Colavizza, T. Blanke, C. Jeurgens, and J. Noordegraaf, "Archives and AI: An overview of current debates and future perspectives," *J. Comput. Cultural Heritage*, vol. 15, no. 1, pp. 1–15, Dec. 2021.
[29] Z. A. Teel, "Artificial intelligence's role in digitally preserving historic archives," *Preservation, Digit. Technol. Culture*, vol. 53, no. 1, pp. 29–33, Apr. 2024.
[30] F. Bressan, R. L. Hess, P. Sgarbossa, and R. Bertani, "Chemistry for audio heritage preservation: A review of analytical techniques for audio magnetic tapes," *Heritage*, vol. 2, no. 2, pp. 1551–1587, May 2019.
[31] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, and M. Dietz, "ISO/IEC MPEG-2 advanced audio coding," *J. Audio Eng. Soc.*, vol. 45, no. 10, pp. 789–814, 1997.
[32] M. Bosi and R. Goldberg, *Introduction to Digital Audio Coding and Standards*. Cham, Switzerland: Springer, 2002.
[33] *Magnetic Tape Sound Recording and Reproducing Systems—Part 1: General Conditions and Requirements*, IEC Standard 60094-1:1981/AMD1:1994, 1994.
[34] "Magnetic tape recording and reproducing standard, reel-to-reel," Nat. Assoc. Broadcasters, Washington, DC, USA, Tech. Rep. E-416, 1965. [Online]. Available: https://www.richardhess.com/tape/history/NAB/NAB_Reel_Tape_Standard_1965_searchable.pdf
[35] M. Bosi, N. Pretto, M. Guarise, and S. Canazza, "Sound and music computing using AI: Designing a standard," in *Proc. 18th Sound Music Comput. Conf. (SMC)*, 2021, pp. 215–218.
[36] M. Bosi, S. Canazza, A. Russo, N. Pretto, and L. Chiariglione, "An MPAI/IEEE international standard for audio: Overview of CAE audio recording preservation (ARP) technology," in *Proc. Int. Conf. Audio Archiving, Preservation Restoration, Audio Eng. Soc. (AES)*, Jun. 2023, pp. 1–8.
[37] A. Basso, P. Ribeca, M. Bosi, N. Pretto, G. Chollet, M. Guarise, M. Choi, L. Chiariglione, R. Iacoviello, F. Banterle, A. Artusi, F. Gissi, A. Fiandrotti, G. Ballocca, M. Mazzaglia, and S. Moskowitz, "AI-based media coding standards," *SMPTE Motion Imag. J.*, vol. 131, no. 4, pp. 10–20, May 2022.
[38] N. Pretto, E. Micheloni, A. Chmiel, N. D. Pozza, D. Marinello, E. Schubert, and S. Canazza, "Multimedia archives: New digital filters to correct equalization errors on digitized audio tapes," *Adv. Multimedia*, vol. 2021, pp. 1–11, Mar. 2021.
[39] A. Russo, M. Spanio, and S. Canazza, "Enhancing preservation and restoration of open reel audio tapes through computer vision," in *Proc. Int. Conf. Image Anal. Process. (ICIAP) Workshops*, G. L. Foresti, A. Fusiello, and E. Hancock, Eds., Cham, Switzerland: Springer, 2024, pp. 297–308.
[40] D. H. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," *Pattern Recognit.*, vol. 13, no. 2, pp. 111–122, Jan. 1981.
[41] R. Szeliski, *Computer Vision: Algorithms and Applications*. Cham, Switzerland: Springer, 2022.
[42] M. Bansal, M. Kumar, and M. Kumar, "2D object recognition: A comparative analysis of SIFT, SURF and ORB feature descriptors," *Multimedia Tools Appl.*, vol. 80, no. 12, pp. 18839–18857, 2021.
[43] *Studio Encoding Parameters of Digital Television for Standard 4:3 and Wide-Screen 16:9 Aspect Ratios*, ITU-R Standard BT.601-7 (3/2011), 2011, pp. 1–18.

[44] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.
[45] T. Y. Goh, S. N. Basah, H. Yazid, M. J. A. Safar, and F. S. A. Saad, "Performance analysis of image thresholding: Otsu technique," *Measurement*, vol. 114, pp. 298–307, Jan. 2018.
[46] E. Micheloni, N. Pretto, and S. Canazza, "A step toward AI tools for quality control and musicological analysis of digitized analogue recordings: Recognition of audio tape equalizations," in *Proc. 11th Artif. Intell. Cultural Heritage (CEUR) Workshop*, vol. 2034, 2017, pp. 17–24.
[47] N. Pretto, C. Fantozzi, E. Micheloni, V. Burini, and S. Canazza, "Computing methodologies supporting the preservation of electroacoustic music from analog magnetic tape," *Comput. Music J.*, vol. 42, no. 4, pp. 59–74, May 2019.
[48] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE Access*, vol. 7, pp. 53040–53065, 2019.
[49] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, vol. 97, K. Chaudhuri and R. Salakhutdinov, Eds., Jun. 2019, pp. 6105–6114.
[50] N. Pretto, N. D. Pozza, A. Padoan, A. Chmiel, K. J. Werner, A. Micalizzi, E. Schubert, A. Roda, S. Milani, and S. Canazza, "A workflow and digital filters for correcting speed and equalization errors on digitized audio open-reel magnetic tapes," *J. Audio Eng. Soc.*, vol. 70, no. 6, pp. 495–509, Jul. 2022.

MARINA BOSI (Senior Member, IEEE) received her degree in physics from the University of Florence, the degree from the Conservatory of Florence, and the degree from the Executive Program, Stanford Business School. She completed her dissertation at IRCAM, Paris, and was a Faculty Member of the Conservatory of Venice. She has held fiduciary positions on several boards. She was the Chief Technology Officer of MPEG LA, LLC, Denver, CO, USA; the Vice President of technology with DTS Inc., Los Angeles, CA, USA; and a member of the research team that created Dolby Digital with Dolby Laboratories, San Francisco, CA, USA, where she also led the MPEG-2 AAC development for which she received the ISO/IEC 1997 Editor Award. She is currently a pioneer in digital audio coding, she is a Founding Director of MPAI. A sought-after keynote speaker, she holds multiple patents and authored significant contributions to academic literature, including the textbook *Introduction to Digital Audio Coding and Standards* (Kluwer/Springer, 2002). She chairs the IEEE SA CAE WG. She is a fellow and the Past President of the Audio Engineering Society. In recognition of her achievements, she has received numerous awards, including the AES Silver Medal.

SERGIO CANAZZA is currently a Professor of "Fundamentals of Computer Science" and "Computer Engineering for Music and Multimedia" with the Department of Information Engineering, University of Padua. He is also the Scientific Director of the Centro di Sonologia Computazionale. He is also the CEO of Audio Innova srl, a university spin-off enterprise (a Founder Member of MPAI). He has been the Project Manager of European projects. He is the owner of patents on safety and health at work. He is the author or co-author of more than 250 papers in international journals and refereed international conferences. His research interests include expressive information processing, auditory displays, and musical cultural heritage preservation and exploitation. He has been the general chairperson and a member of technical committees at several conferences. He won the Le Palm D'Or at the World Artificial Intelligence Cannes Festival, France, in 2023 and 2024, one of the most important AI events at the world level.

NICCOLÒ PRETTO (Member, IEEE) received the bachelor's and master's degrees in computer science engineering and the Ph.D. degree in information engineering from the Department of Information Engineering, University of Padua, Italy. He is currently an Assistant Professor with the Free University of Bozen-Bolzano, Italy. He is also a part of the Media Interaction Laboratory. He is also a Founder Member of the company MPAI Store Ltd. His research interests include sound and music computing and preservation and access to historical audio documents and cultural heritage in general. More specifically, his work consists of research and development of innovative methodologies, and applications to preserve, analyze, and experience musical cultural heritage, adopting several technologies and methods extending to web and mobile interfaces, embedded systems, and machine learning techniques.



ALESSANDRO RUSSO received the bachelor's degree in technologies for cultural heritage from the University of Turin, in 2012, and the master's degree in materials science for cultural heritage, in 2015. He is currently pursuing the Ph.D. degree in brain, mind, and computer science with the University of Padua, with a project concerning the preservation, re-activation, and documentation of audio-visuals, and multimedia installations. Since 2016, he has carried out research activities collaborating with the Centro di Sonologia Computazionale (CSC), Department of Information Engineering (DEI), University of Padua, mainly focusing on audio restoration; and the La Camera Ottica Laboratory and the Film and Video Restoration Laboratory, University of Udine, working on several projects of digitization and restoration of audio-visual funds belonging to Italian and International foundations and archives. His research interests include the preservation of audio, film, and video archives.

MATTEO SPANIO received the bachelor's degree in computer science (data science) from the Ca' Foscari University of Venice and the bachelor's and master's degrees in performing arts from the Conservatorio di Musica "C. Pollini," Padua. He is currently pursuing the Ph.D. degree with the Brain, Mind, and Computer Science Program, University of Padua. His research interests include the intersection of artificial intelligence and music, focusing on generative AI for music based on cross-modal interaction and the preservation of audio documents. In addition to his academic pursuits, he collaborates as a Software Engineer with the companies, such as Soundfood and Audio Innova. He frequently performs as the first clarinet in many professional orchestras and has graced prestigious stages and theaters across Italy, Hungary, Austria, and Germany.

Open Access funding provided by 'Università degli Studi di Padova' within the CRUI CARE Agreement