Niccolò Pretto, Carlo Fantozzi, Edoardo Micheloni, Valentina Burini, and Sergio Canazza
Department of Information Engineering
University of Padova
Via Gradenigo, 6b
Padova, 35131, Italy
{niccolo.pretto, carlo.fantozzi,
edoardo.micheloni, valentina.burini,
sergio.canazza}@dei.unipd.it

Computing Methodologies Supporting the Preservation of Electroacoustic Music from Analog Magnetic Tape

Abstract: Electroacoustic music on analog magnetic tape is characterized by several carrier-related specificities that must be considered when creating a copy for digital preservation. The tape recorder needs to be set to the correct speed and equalization, and the magnetic tape could have some intentional or unintentional alterations. During both the creation and the musicological analysis of a digital preservation copy, the quality of the work may be affected by human inattention. This article presents a methodology based on neural networks to recognize and classify the alterations of a magnetic tape from the video of the tape as it passes in front of the tape recorder's playback head. Furthermore, some machine-learning techniques have been tested to recognize a tape's equalization from its background noise. The encouraging results open the way to innovative tools able to unburden audio technicians and musicologists from repetitive tasks and to improve the quality of their work.

In the last 20 years, the preservation of historical audio documents has been one of the main research topics at Centro di Sonologia Computazionale (CSC), the Sound and Music Computing laboratory of the Department of Information Engineering at the University of Padova (cf. Zattra, De Poli, and Vidolin 2001; Canazza and Vidolin 2001a). The problems concerning audio preservation are multifaceted, and a multidisciplinary approach is necessary to exploit the huge potential of this documentary heritage. The methodological framework perfected in these years involves both the active preservation of historical audio documents (Bressan and Canazza 2013) and access to them (Canazza, Fantozzi, and Pretto 2015; Fantozzi et al. 2017), with a focus on analog magnetic tapes. In our laboratory, the consistency of this framework was tested using historical recordings of electroacoustic music during several international projects related to the preservation and restoration of collections of sound recordings carried out with important European audio archives (including the Paul-Sacher-Stiftung, Basel; the Fondazione Arena di Verona; the Historical Archive of the Teatro Regio, Parma; and the Luigi Nono Archive, Venice).

Computer Music Journal, 42:4, pp. 59–74, Winter 2018
doi:10.1162/COMJ_a_00487

Contrary to passive preservation concerning the safeguarding of the material structure of documents, active preservation consists of preserving their content in digital form. This is to ensure, after the initial digitization, a safe transfer of identical copies from one digital carrier to another. There are, nevertheless, many aspects that need to be considered during the digitization of a tape. First, there is the object's material structure–that is, the set of its physical-chemical components, the technology, the production system (acoustic, electroacoustic, magnetic), and the audio and playback format (such as speed and equalization). Next, there is the primary information (i.e., the audio signal recorded). Then there is the secondary (or ancillary) information (Brock-Nannestad 1997; Bressan and Canazza 2013), such as notes on the box, noise signals characterizing the recording system, alterations of the carrier (corruptions, splices, signs, etc.). There may also be other metadata that need to be perserved. Finally, there is the history of the document transmission (storage, duplication, etc.). All of these metadata need to be stored with the preservation copies alongside the digital audio. In this sense, we define a preservation copy of an analog audio document as an organized data set that groups all the information (data and metadata) represented by the source document, stored and maintained in several directories of the archive data center.

The methodology aims to go a step further, making the digital preservation copies more reliable and suitable from a scholarly point of view. The software tools we have developed emphasize the "textual" aspects of a sound document, considering the A/D transfer as a philological operation of *restitutio textus*. The use of philological tools is particularly important in the field of electroacoustic music on analog magnetic tapes (cf. Zattra 2007; Verde et al. 2018). Here, A/D transfer of an audio document might not only be affected by digitization errors (related to speed, equalization, track numbers, etc.), but also cause the loss of useful ancillary information, and therefore generate a proliferation of document "witnesses" with poor value (from a philological point of view). These document witnesses are nonidentical digital audio documents, with "variants"—or differences in comparison to the original analog tape. But they are "poor" readings because they may be imperfect approximations of the original, generating "noise" in the textual critic's task.

There are different definitions of the term electroacoustic music: We use here the term in a technical sense, that is, to refer to any music for which the audio realization inherently requires analog magnetic tape used as a "creative" instrument. In this sense, this music is the most important and innovative music genre of the second half of the 20th century, representing a paradigmatic case of recorded sound art with great implications for preservation as well as for musicological analysis. This "music on tape," or tape music, evolved along with technologies for music postproduction, embracing most genres and aesthetic trends of recorded sound arts. The birth of tape music revolutionized the production system of music itself: The composer became both "luthier" and performer of the music product, recording it on analog magnetic tape, which became a unicum, rather like the products of other art forms (painting, sculpture, etc.). The music changed its artistic category: An *allographic art* (i.e., an art performed with the contribution of multiple actors; in the case of music, composer, and performer) became an *autographic art*, where there is a complete product *in se* at the end of the creation process. As an example, Nelson Goodman (1968), by means of the distinction between autographic and allographic, not only distinguishes replicas from reproductions introducing falsifiability in music, but faces the problem of the identity–difference relation between work and object, and between work and text.

In the first period of tape music several real masterpieces were created without written score; the analog magnetic tapes were the only available documentation. The interpretation of these works was no longer the traditional one, in which one or more musicians performed a score. Instead, it moved towards the projection in a hall by a sound engineer (or by a live-electronics musician), using (or ignoring) the characteristics of the acoustic space; a sort of parallel to the exposure of a painting in an exhibition, where the perception changes according to the conditions of lighting and the layout of the hall.

In this sense, in the field of tape music, a unique feature of the preservation process developed at CSC is the video recording of the tape as it passes the head of the tape recorder, which is important to preserve important ancillary information.

Both the creation and the study of a digital preservation copy share a common problem: human attention. Inattention during several hours of listening to the audio documents and observing kilometers of running tape can lead laboratory technicians and musicologists to overlook some details, such as corruption of the tapes or markings on them. Work becomes messy and could be done without the necessary care. Some of the disorder patterns are similar to the well-known ones affecting transcripts made by the *servus a manu* scribes (i.e., slaves, or Benedictine monks, or professional copyists); others have not yet been studied. Furthermore, in the absence of information related to the audio format of a tape, an audio technician has to make some decisions aurally (Bradley 2009), such as the choice of the correct equalization standard. The probability of an incorrect choice is high and an incorrect decision causes the creation of document witnesses of poor value.

We advocate the use of automatic techniques to extract the aforementioned types of information from audio and video of the tapes to relieve technicians and musicologists of repetitive, tiresome, or otherwise error-prone tasks that are better performed by a machine. This could leave the technicians and musicologists with the last word on issues that require expertise and intelligence. The automatic analysis can also be considered as a preprocessing step that filters relevant data for a high-level interpretation by a professional. The aim of the present article is to: (1) investigate and list possible sources of information, (2) discuss the relevance of each, (3) quantitatively summarize the features of the source that make automatic analysis possible, and (4) propose algorithmic techniques that exploit such information to perform the analysis. Considering the focus of the article, the implementation details of the algorithmic techniques are not described. Rather, a preliminary software implementation and an evaluation of its performance are presented.

The next section of this article introduces the concept of tape discontinuity and the equalization standards related to analog magnetic tape. We then tackle the problem of recognizing tape discontinuities, by using audio and video of the preservation copies, and we propose a methodology based on neural networks to automatically analyze a video recording of an audio tape's playback in the tape recorder. The subsequent section introduces the problem related to equalization and presents a methodology based on machine-learning techniques to detect the correct equalization standard. Finally, we conclude, proposing several awares for future work.

Preserving the Documentary Unit

For historically faithful, active preservation of an original audio document (i.e., to preserve its documentary unit), all information regarding the document must be stored and maintained in the digital domain. Just what information this entails is discussed in the following paragraphs. Usually only the audio content of a recording is analyzed by scholars, however. This section highlights the importance of the analysis of the carrier both for preservation and for musicological studies. Furthermore, the concept of discontinuity will be introduced along with the equalization problem.

Filming Audio

To preserve the documentary unit of a historical audio document, it is necessary to store all information (both intentional and unintentional) stemming from the carrier and from its transmission. The methodology described by Bressan and Canazza (2013) emphasizes the importance of preserving as static images (1) the information reported on edition containers, labels, and attachments, and (2) any clearly visible alterations on the carrier. To preserve all the ancillary information concerning the carrier (physical condition, presence of intentional alterations, corruption, and graphical signs), a video recording of the tape as it passes the playback head during playback should be stored with the preservation copy. This video must also contain the audio, although not necessarily in high quality, to allow synchronization of the video to the high-quality audio signal of the digital preservation copy. The video recording offers:

1. Information on the operations of the magnetic tape assembly, such as the splices used to join different pieces of tape and possible corruptions of the carrier, which are indispensable to distinguish intentional from unintentional alterations during the analysis of the audio document (AES22 1997; Canazza 2007; Canazza and Vidolin 2001b).
2. Instructions for the performance of the piece: From the video analysis, some markings on the tape can be detected; typically, they either represent points to be synchronized with a musical score, or they may indicate particular sound events.
3. A description of the irregularities in the playback speed of analog recordings that cause changes in frequency, such as wow and flutter.

In the CSC methodology (Bressan and Canazza 2013; Fantozzi et al. 2017), only the back of the tape (i.e., the nonmagnetic side) is recorded, to reduce the technical complexity and to facilitate the adaptation to different tape recorders. This means that some details related to the magnetic side will be lost, but usually alterations involve

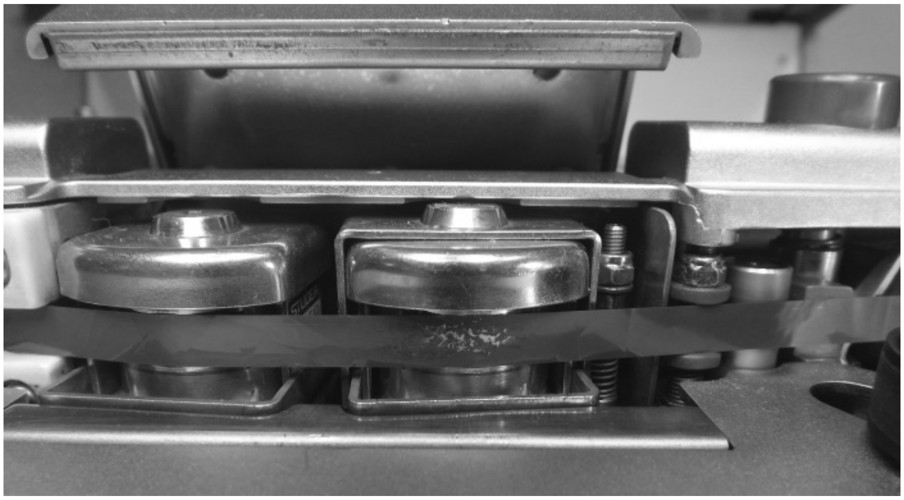


Figure 1. Creases on a tape and localized magnetic coating damage. (Color version available at www.mitpressjournals.org/doi/suppl/COMJ_a_00487-Pretto.)

both the sides. The expression "video of the tape" will be used henceforth to indicate the video of the back side of the tape. The videos are the input data of the original software tool presented in the following section. They are able to automatically locate discontinuities that occurred during the A/D transfer and to classify them, thus increasing information about the signal that may be useful for philological analysis. The videos used in this work derive from several years of preservation projects and were captured by a professional camera with a 720 × 576-pixel resolution.

Tape Discontinuities

According to cataloguing rules set by the International Association of Sound and Audiovisual Archives (IASA, cf. Miliano 1999), an initial visual inspection of the fully wound tape pack by the technician may show the presence of several alterations such as blocking, leafing, windowing, spoking, or embossing. This first stage is essential to determine whether or not the carrier has to be restored prior to digitization. Nonetheless, some degradation and other conditions can only be detected during tape playback, after visual inspection. Consequently, both of these two stages—visual inspection and monitoring playback—become particularly important to evaluate the preservation condition of the tape.

According to the IASA cataloguing rules (Miliano 1999, Appendix C), the main alterations recognizable during tape playback are:

Cupping: an abnormal flexure of the tape surface across or along its width, due to different rates of shrinkage along the substrate and recording layers.

Damage to tape edges: occurring when the edges do not appear flat or straight.

Riffles: formally known in the cataloguing rules as "kink" or "wrinkle," these may be a single crease on a layer of tape or multiple creases in the tape (see Figure 1).

Tape contamination and dirt: presence of mold, powder, crystals, other biological contaminations, or similar sullying.

Interlayer adhesion: stickiness of the surface of one layer to the back of the succeeding layer, which could cause wow and flutter.

Gummy deposit: presence on the tape of gluey substances that can gather on the heads and guides of the playback machine during the tape playback.

Backcoat and magnetic shedding: the first of these involves backcoat particles coming away from the substrate and accumulating on surfaces in contact with the back of the tape. The loss of debris can impair playback quality, leaving deposits on the playing surface of the adjacent layer. The second phenomenon, due to a loss of cohesion, entails magnetic coating particles coming away from the tape substrate, accumulating on the heads and guides of the playback machine.

Brittleness: frequently seen with cupping, results in easy tape breakage.

Marks: signs or words written on the back of the tape (i.e., the nonmagnetic side) or on the adhesive tape of splices. Similarly, the presence of ink or dye on the surface of the tape, or writings on the back seeping through the tape to the front, comprise the phenomenon of "bleeding."

These alterations can be detected only after an attentive monitoring of the entire length of tape during the playback process. The responsibility for both kinds of visual inspections is laid on the tape operator, as the first person involved in the digitalization process, and the director of the collection, who has to manage the preservation intervention. In this regard, appropriate auxiliary evaluation methods have been developed over the years (Casey 2008; Sueiro 2008), but a subjective judgment may undermine the evaluation process because of the great amount of material to check. Therefore an objective and automated intervention is needed. This article moves in this direction, attempting to extract information about tape alterations through automatic computer analysis of the two files, one storing high-quality audio and the other the video.

Another important kind of alteration of the carrier is the splice. In the IASA cataloguing rules, a splice is defined as a small piece of special adhesive tape used to join two pieces of recorded material to form a single piece (Miliano 1999, Appendix C). Splices can join also two pieces of magnetic tape rather than a piece of magnetic tape to a leader tape. According to Delos Eilers (1968), an ideal splice is one that will remain intact for an indefinite period of time and not cause an audible disturbance upon playback. The corruption of the output audio signal mainly depends on the splice's angle, that is, the angle measured between the cut and the tape edge. The most desirable method is to cut the tape at a 45° angle. According to the tape manufacturer Scotch (Eilers 1968), as the angle raises towards a perpendicular cut, the amount of electrical disturbance is increased because the playback head sees the discontinuity at the junction as an abrupt change. On the other hand, an angle smaller than 45° entails a lesser amount of electrical disturbance, but a the tape is then more vulnerable to bending and breakage.

The term *discontinuity* will be used henceforth to indicate all the alterations of the carrier from its original manufacture state that are detectable while the tape is running. Although not strictly an alteration, manufacturers may print their brand name or logo on the back of the tape itself. As necessary and useful information for the identification of tapes, tapes brand markings will also be considered a discontinuity of the tape.

This section has presented the motivations for why, during the digitization process, it is important to be aware of the tape's condition and the possible presence of splices. Discontinuities also have an important role in musicological studies, however. In electroacoustic music on analog magnetic tape, composers manipulated the tapes. Sometimes, even "incorrect" manipulation became part of the artistic act. Therefore, the examination of splices and, in general, discontinuities becomes relevant to studying the genesis of a work of art. Unfortunately, this assessment can be pursued only during playback, while monitoring up to several hours of tape. To avoid human errors, an automatic approach is essential for both the creation of the preservation copies and for musicological analysis of the document. A possible approach is detailed in the section on "Analysis of Tape Discontinuities."

Equalization

As with most analog audio formats, the signal representation on tape is deliberately nonlinear in terms of frequency response. Consequently, correct playback requires appropriate equalization (Bradley 2009). The use of pre- and postemphasis techniques is a method that modifies the spectrum of the audio signal of the source as read by the recording head and then performs the inverse modification during playback (Fielder 1985). Therefore, the postequalizer makes the overall transfer function nearly flat (Mallinson 1976). This technique found wide application in the past because of the limited dynamic range of the audio systems and the fact that the music source generally produces more energy in the low-frequency region, where the ear is less sensitive to noise. The advantages are extension of dynamic-range capabilities (Fielder 1985; MRL 2016) and improvement of the signal-to-noise ratio (Camras 1987).

Different equalization standards exist, and they are commonly identified by the initials of the organization that wrote the tape recording standard (MRL 2016). The most common standard in Europe is the International Radio Consultative Committee (Comité consultatif international pour la radio, CCIR), also referred to as IEC1 (from International Electrotechnical Commission). The standard used most commonly in America is by the National Association of Broadcasters (NAB), alternatively called IEC2. According to standard definitions (NAB 1965; IEC 1994), the equalizations are the results of the combination of two curves:

$$N(DB) = 10log(1 + \frac{1}{4\pi^2 f^2 t_2^2}) - 10log(1 + 4\pi^2 f^2 t_1^2), \tag{1}$$

where f is the frequency in Hz and t_1 and t_2 are time constants in seconds. For CCIR, the curve shape depends only on the constant t_1, and t_2 is set to infinity (IEC 1994). Furthermore, the time constants change in relation to the tape speed. The equalization, as well as the tape speed, needs to be correctly configured to obtain a flat frequency response. According to Kevin Bradley (2009), sometimes any lack of documentation may require the operator to make playback equalization decisions aurally and, as will be presented in the section on "Equalization Errors during Digitization," this lack often leads to errors due to subjectivity of choice.

Analysis of Tape Discontinuities

The audio signal is not sufficient for the detection of discontinuities. The first part of this section will show an unsuccessful attempt concerning splice detection by using audio features. Following that, an innovative approach for discontinuity detection and classification, based on neural networks, will be described and tested. Frames extracted from the video of the back of the tape are the input of this automatic tool.

Use of Audio to Recognize Splices

There is a rich bibliography regarding alterations of audio cues and audio restoration (Godsill, Rayner, and Cappé 2002; Canazza, Poli, and Mian 2010; Canazza 2012). In the specific case of magnetic tapes, some alterations are connected to the discontinuities of the carriers. The first part of the study analyzes relevant features to discover the most significant ones for discerning discontinuities of the carrier. As noted previously, one of the most common discontinuities is the splice. When a splice joins a magnetic tape with a leader tape (which is nonmagnetic), its presence can be inferred by the beginning or end of the empty spectrogram that characterizes nonmagnetic tape. Even if a magnetic portion has no recorded input, it has a distinguishable noise. On the other hand, when a splice joins two pieces of magnetic tape, recognition of the discontinuity is more difficult, as discussed by Verde et al. (2018), whose automatic tools were not able to recognize this kind of splice. For this last reason, the following experiment based on visual inspection was necessary.

As introduced in the section on "Tape Discontinuities," the quality of splices is strictly related to the angle at which the magnetic tape is cut. To identify the features characterizing audio cues of splices that join two pieces of magnetic tape, we created a set of 40 splices as an initial test: half with a cut at 45°, the others with a cut at 90°. In each of these two sets, half of the splices were created on new, pristine tape (i.e., never used for

Table 1. Maximum Frequency Values of 90° Splices in Test 1

	Sample Number	*Channel 1 (kHz)*	*Channel 2 (kHz)*
Unused Tape	1	7	0
	2	0	47
	3	43	44
	4	0	19
	5	10	24
	6	9	10
	7	34	44
	8	6	34
	9	13	23
	10	3	19
Recorded Silence	1	17	19
	2	15	22
	3	22	20
	4	16	22
	5	16	17
	6	14	20
	7	13	12
	8	13	22
	9	13	23
	10	26	19

recording), whereas the other half on a tape recorded with silence. The tapes were recorded and read at 30 in./sec (ips) using a Studer A810 tape recorder. This speed represents the worst case, because the splices run over the heads in the shortest possible time. The samples of the test were digitized at a sample rate of 96 kHz with 24-bit resolution to allow analysis of frequencies over the entire audible range.

The result of the test was unambiguous: All the 90° splices were clearly recognizable in the spectrogram, whereas the 45° splices were not recognizable, except for only one where the splice was weakly distinguishable. Magnetic tape recordings are generally free of clicks (Godsill and Rayner 1998), but, as can be seen in Table 1, for 90° splices there are visible spikes at least in one channel of the digitized samples. Figure 2 shows a clear example of this spike. The peaks are not uniform in all the samples but involve all frequencies from zero to a variable maximum value. Some of them greatly exceed the audible range, nearly reaching the maximum frequency of 48 kHz. Nevertheless, only a few are recognizable solely from listening to the sample.

A second test was based on another 80 samples of splices that were each created by simply cutting a tape recording and then attaching the two resulting loose ends together again in the same place. This aspect may seem trivial but is worth noting, because often in tape music tapes from completely different recordings are joined together to obtain some desired effect. In that case, the recognition is simpler because the two strips of tape are not homogeneous. In this test, we aim to reconstruct a worst case, a splice between two homogeneous strips (or rather, a continuous sound), as could happen in the case of broken tapes. There are four categories of sounds recorded before the cuts were made: instrumental acoustic music, speech, music made with an electric bass, and electroacoustic music.

As can be seen in Table 2, which presents the results of the two tests, the number of detected spectral peaks drastically decreases in the second case. The performance for the 45° splices of the second test is similar to the ones of the first test. Regarding the 90° splices, the number of distinguishable peaks in the spectogram dropped from 100 percent to 42.5 percent. This behavior could be explained by the fact that the peaks have either lower power with respect to the surrounding frequencies, or the discontinuities are described by too few samples for reliable recognition. In the 90° samples of the first test, only 60 percent of the peaks exceed a frequency of 20 kHz, so the result is plausible. In conclusion, the test proves that these kinds of peaks in the spectrogram are not a good feature for recognizing 45° splices. With 62 percent of the 90° splices being distinguishable, the result is better but insufficient to recognize a splice with certainty. This unsuccessful attempt prematurely ended after the analysis of these two tests. The feature identified in some of the samples is not robust enough to be exploited in a reliable automatic tool for detecting splices. An alternative approach was required, which is presented in the following section.

Figure 2. Waveform (a) and spectrogram (b) of the samples: a splice on new, pristine magnetic tape (i.e., never used for recording).

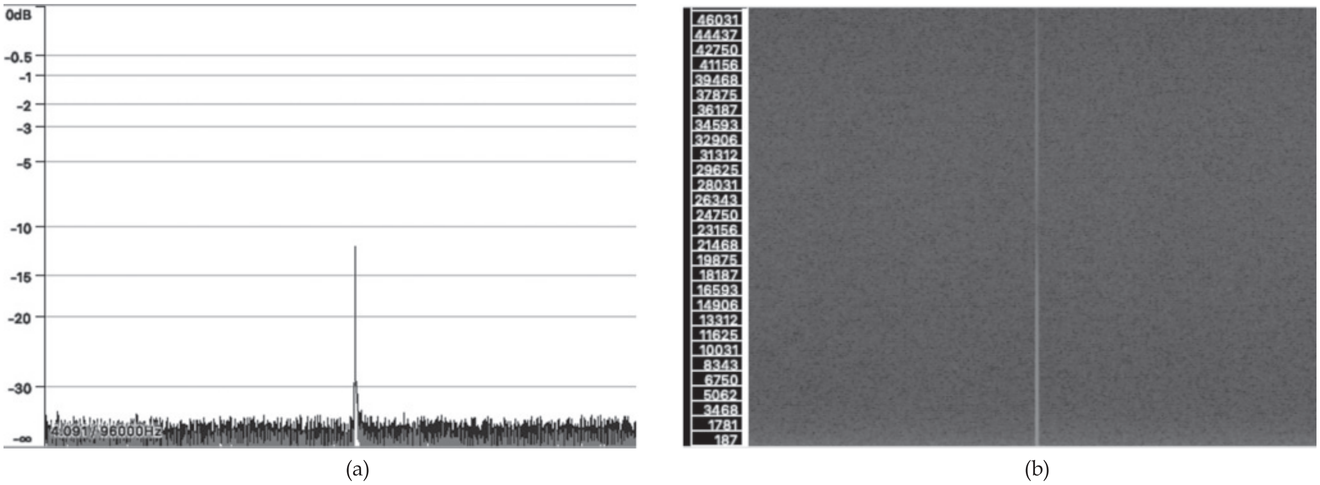


Table 2. Results of the Two Tests

		45° Splices		*90° Splices*	
Test	*Types*	*Recognized*	*Unrecognized*	*Recognized*	*Unrecognized*
1	Pristine tape	0	10	10	0
	Silence	2	8	10	0
2	Acoustic music	7	3	9	1
	Electric bass	0	10	3	7
	Electronic music	0	10	3	7
	Speech	0	10	2	8

Use of Video to Recognize and Classify Discontinuities

Through a video recording of the back of the tape during playback (with framing as in Figure 1), it is possible to extract additional information with respect to the audio signal.

1. It is possible to ascertain whether an alteration is due to a physical problem with the tape or to a specific choice by the composer.
2. Irregularities in playback speed, creating audio artifacts, can be documented.
3. Markings or written notes on the tape, which indicate relevant moments in the composition process, are apparent.

Our automatic analysis of tapes using video recording involves two steps that are applied in sequence. During preprocessing, the first step, the video is examined frame by frame, and each image showing a potentially significant discontinuity is saved. The exact content of the images is not determined. That task is the aim of the second step, classification, in which a classifier is used to determine the content of each image saved during preprocessing. The extraction of frames in the preprocessing step can be performed automatically, as detailed by Fantozzi et al. (2017). For the second step, a possible classifier is described for the first time in the present article. As we will explain in more detail later, it is based on convolutional neural networks.

Before training a classifier for video analysis, the classes of interest must be defined precisely. In our experiments, we considered the following classes:

L–M splices: Splices of leader tape to magnetic tape
M–M splices: Splices of magnetic tape to magnetic tape
Brands on tape
Ends of tape
Ripples
Damaged tape
Markings
Dirt
Shadows

A brand may be the full name of the tape manufacturer, or just a logo. The brand changes in size, shape, and color, depending on the tape used, thus complicating the classification task. We separate brands from other marks (e.g., manual annotations), using two different classes, because they have significantly different meanings for musicologists. The Ends-of-tape class refers to what happens when the tape reaches its end of playback, at which point it is neither under tension nor in contact with the capstan and pinch roller. The distinguishing visual characteristic of this class is the tape coming free—or completely detached—from the capstan. The Ripples class groups all the alterations in the shape of the tape, such as cuppings and damages on tape edges, as described in the "Tape Discontinuities" section. In the Damaged tape class, we group all kinds of damages on the surface of the tape that are not alterations of the shape of the tape: contamination, creases, etc. Markings can be words or symbols written on the magnetic tape, on the adhesive tape, or on both. Again, video frames in this class can be significantly different from each other. If the tape exhibits irregularities that are not physical damages, ripples, or marks, their frames are put into the Dirt class. The Shadows class contains frames in which shadows or reflections are temporarily cast on the tape by external objects in motion. The availability of this class helps the classifier in discriminating frames that actually contain nothing interesting: A shadow, if present, may be continuously changing, depending on the movements of the external object, hence the preprocessing phase repeatedly finds something new and saves several frames that are actually uninteresting.

Data Set

A good data set for supervised training must be large enough to cover the different circumstances that may occur. In addition, class imbalance should be minimal: The number of elements in each class must be similar. Our data set was built from videos of a number of tapes available at CSC and recorded at both 7.5 ips and 15 ips. An artificial tape, containing several different splices, was deliberately prepared and added to the set with the aim of better training for cases that are underrepresented in real tapes. Tens of thousands of images were extracted during the preprocessing step. The first 40,000 potentially significant images were manually analyzed and labeled, or discarded in the event of false positives. After this work, a large number of images still remained to be processed, so we sought a strategy to speed up the work. We noted that, in the videos being preprocessed, the majority of the frames were saved because of the presence of brands on the tape. Because brands appear periodically, they are repeatedly detected as something special that needs to be notified; as a matter of fact, they are not interesting. We searched for a way to discriminate such images automatically; in the end, we created neural networks by fine-tuning modified versions of GoogLeNet, proposed by Szegedy et al. (2015), on a training set of 5,000 brands plus other discontinuities of different types, and a validation set of 3,244 brands. We trained two separate neural networks for videos of tapes running at 7.5 ips and 15 ips. It was decided to manually check the frames classified by the network if the classification confidence was below 90 percent.

We also noticed that some splices between leader tape and magnetic tape were barely visible. It was decided to keep a frame in the data set, and assign it to the L–M splices class, only if the splice was completely visible. The discarded frames were moved to a separate folder and used as a validation set in a dedicated test (more details will be provided at the end of this section). We further had to deal with other extreme situations, such as the simultaneous presence of multiple discontinuities in the same frame. In this case, it was necessary to decide how to classify the frame considering which discontinuity was most important. The criteria we followed are:

1. If a frame contained both marks and other discontinuities (e.g., a splice), then it was assigned to the Marks class.
2. If a frame of damaged tape exhibited a splice, then it was assigned to one of the splice classes (L–M splices or M–M splices, as appropriate).
3. If a frame with shadows exhibited a splice, then it was assigned to the appropriate splice class.
4. If a frame did not exhibit a splice but exhibited adhesive tape applied onto the magnetic tape, then it was assigned to the Marks class because adhesive tape that does not join a cut in the magnetic tape indicates an annotation.
5. Frames were assigned to the Ends-of-tape class only if the capstan was completely detached from the tape.

The final data set, after the application of neural networks and the subsequent manual selection, contains 20,333 images, as summarized in Table 3. It should be noted that, despite our best efforts, the nine classes are not equally represented, because instances of some classes, such as Dirt or Ripples, are rarely found in real tapes.

It is also important to observe that all the discontinuities appear distorted in the videos and, consequently, in any video frame we extracted from them. The discontinuities are elongated and blurred because of the camera's low frame rate (25 frames/sec) and slow shutter speed (order of magnitude: one hundredth of a second). A third observation is that our videos, recorded in the PAL standard, are affected by the interlacing phenomenon. Interlacing is a technique implemented in several video standards that mandates the acquisition of the odd and even lines of a frame at slightly different times to reduce the bandwidth required for video transmission; the technique produces jagged contours in moving areas. We tested various de-interlacing approaches with the aim of improving the quality of the frames in the data set. In the end, we decided to simply discard the even lines for each frame, even if this choice actually eliminates half of the information. As a consequence, the resolution of frames in the final data set is the one of digitized PAL fields, that is, 720 × 288 pixels.

Table 3. Classification from Video: Numbers of Elements in Data Set

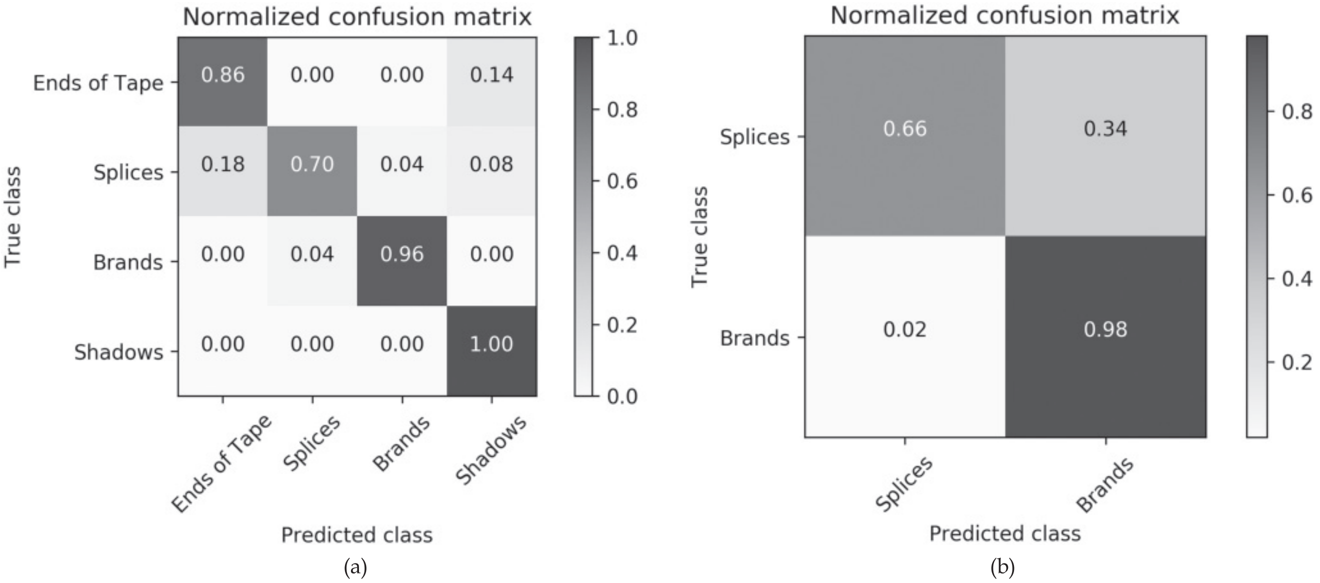
Class	*Elements (7.5 ips)*	*Elements (15 ips)*
L–M splices	1,041	1,459
M–M splices	435	715
Ends of tape	2,484	84
Brands on tape	5,834	5,709
Damaged tape	456	183
Ripples	6	0
Shadows	1,347	55
Marks	145	363
Dirt	8	9
ALL CLASSES	11,756	8,577

Training and Experiments

The convolutional neural networks for the classification of tape discontinuities have the same structure as those previously used to build the data set: They are derived from GoogLeNet, proposed by Szegedy et al. (2015), by preserving the intermediate layers for feature extraction. Preserving layers also has the benefit of making a full, costly retraining unnecessary: A fine-tuning can be adopted instead. The networks were implemented in the Caffe framework (Jia et al. 2014). In all the experiments, distinct classifiers were trained for videos of tapes running at 7.5 ips and 15 ips. Classification time per frame was the same for all classifiers: about 0.4 seconds on a Windows machine with an 8-core Intel Core i7 CPU running at 4 GHz, along with a GeForce GTX960 board for GPU-accelerated computing. This is a consequence of the fact that all the classifiers have the same internal structure.

In a first experiment, classifiers were trained only for the classes where a high number of examples were available in the data set. Moreover, given the

Figure 3. Confusion matrices for classifiers trained in the first experiment: the classifier for 7.5-ips tapes (a) and the classifier for 15-ips tapes (b).



(a)

(b)

imbalance between the M–M and the L–M splices, we decided to merge the two classes. Such choices left us with four classes (Splices, Ends of tape, Brands on tape, Shadows) for the 7.5-ips case, and two classes (Splices, Brands on tape) for the 15-ips case. In each of the two cases, 50 images per class were left out of the training set and used to assess the performance of the classifiers after training. The results of the assessment are summarized by the confusion matrices in Figure 3. The classifiers were able to correctly label from 66 percent to 100 percent of the frames, depending on the class.

In a second experiment, we trained neural networks for all the classes featured in the data set. The results of this experiment must be examined with the knowledge of the class imbalance that affects the data set as a consequence of the class imbalance in real tapes. Nevertheless, we believe that our results may provide useful indications, including the raw learning power in the fine-tuning process. Class imbalance, which is a common phenomenon (Wang and Yao 2012) in machine-learning tasks, can be tackled with a wide array of techniques at both the data set level (e.g., undersampling of the overrepresented classes or augmentation of the underrepresented classes with synthetic data) and the classifier level (e.g., adoption of a classifier ensemble [Nanni, Fantozzi, and Lazzarini 2015] or ad hoc classifiers for some classes). An appraisal of the appropriate techniques for the domain of tape videos mandates a dedicated study, however. In the second experiment, frames from the artificial tape were excluded from the data set and set aside for a test that will be described shortly. This choice left us with a data set containing 11,475 frames from tapes running at 7.5 ips, and 8,387 frames from tapes running at 15 ips. The accuracy of the networks was first assessed by k-fold cross-validation with k equal to 5. In this validation technique, it is established that elements from different sources must be used for different folds (e.g., all frames from video v_1 reserved for validation in fold 1, frames from video v_2 for the second fold, and so on). In our case, given the uneven number of frames contributed to the data set by the different videos, we decided to divide the former into five folds with the same number of images, regardless of the video source. Had we applied the standard strategy, some classes would have contained very few elements, or even none at all. The 7.5-ips network was fine-tuned by splitting the data set into 9,180 items (80 percent) for training and 2,295 (20 percent) for validation. For the creation of the 15-ips network, the fine-tuning was performed using 6,709 elements for training and 1,678 for validation; again, this represents an 80 percent–20 percent split of the data set. In the validation phase, accuracy turned out to be slightly lower for 15-ips tapes than for 7.5-ips tapes, but it was consistently high: The lowest accuracy measured over all the folds was 98.9 percent. In a second, more significant test, the classifiers were evaluated on frames extracted by the preprocessor on videos not previously employed to train the classifiers themselves: two videos of the artificial tape, run at both 15 and 7.5 ips. For this test, considering how the contributions of different videos to the data set and the size of the data set itself influence the possibilities for validation, we put ourselves in a worst-case scenario and used the classifiers with lowest accuracy emerging from the *k*-fold cross-validation. By using the aforementioned 8-core personal computer, the preprocessing step on the 15-ips and 7.5-ips videos was completed in 4:46 and 9:27 min, respectively. (The correlation between the two figures is easy to establish given the fact that, in the latter case, the number of frames to preprocess is doubled.) The step selected 207 frames at 15 ips, and 281 frames at 7.5 ips. At both speeds, the preprocessing step exhibited almost perfect recall, i.e., it was able to correctly detect all frames where a discontinuity was present, with only two exceptions: two annotations with the words "Ancora stacco musica + voce" and "Continuo musica e voce" were not detected.

Classification performance turned out to be heavily dependent on the class. It was low for the Damaged tape, Shadows, and Marks classes, and fair for the L–M and M–M splice classes; the remaining four classes did not appear in the tape. We note that the classifiers may confuse one kind of splice with another, but if both kinds of splices are grouped into a single class then precision and recall are high (0.9 or better). As a final test, frames in the splice class excluded from the data set because the splice is not entirely visible were used to verify how many splices could be correctly classified by the neural networks (i.e., the recall for splices) in such extreme cases. In the same spirit of the previous experiments, the 7.5 and 15-ips classifiers with worst accuracy were used. Results are listed in Table 4: They emphasize the ability of the networks to identify a significant fraction of the splices even in borderline cases.

Table 4. Classification of Frames Discarded from Videos

Frame content	*P*	*TP*	*Recall (TP/P)*
15 ips: L–M Splice	147	97	0.66
15 ips: M–M Splice	135	97	0.72
7.5 ips: L–M Splice	236	177	0.75
7.5 ips: M–M Splice	71	49	0.69

Frames were discarded when the splice was not entirely visible. P *is the number of discontinuities in the frames;* TP *(true positives) is the number of such discontinuities that were correctly classified.*

Analysis of Equalization

As presented in the section "Preserving the Documentary Unit," equalization is an important parameter that requires precise configuration on the tape recorder to obtain a faithful digital preservation copy. In this section, the problem of equalization recognition will be discussed, along with an automatic approach to the detection of this important parameter.

Equalization Errors during Digitization

To obtain a correct digitization of an analog tape, all recording parameters should be known. The equalization curve as well as the playback speed are the most important parameters, but frequently they are not known. As noted in the section on Equalization, in the case of a lack of documentation, IASA guidelines allow the operator to make the decision aurally. We carried out an experiment to study the capabilities of skilled and unskilled testers to discriminate digitized audio samples with different equalization (Burini, Altieri, and Canazza 2017). The experiment consisted of a multistimulus test with hidden reference and anchors (MUSHRA, cf. ITU 2003) that used different audio samples of spoken parts, instrumental music, and vocal music. Reference samples had been previously recorded twice, alternating CCIR and NAB equalization curves. Each recording was also read twice, using both equalization curves. As a result, there were both CCIR and NAB samples, each digitized using both correct and incorrect equalization. They could be compared to each other and to the original reference recording, as well. Each test provided the original reference recording, but nevertheless both skilled and unskilled testers sometimes committed errors in recognizing the uncorrected equalization.

Given that the operators usually do not have a reference during the digitization process, there is evidently a high probability of encountering an incorrect decision. Furthermore, no scientific method is provided in literature to aurally detect the correct equalization. For all these reasons, a new method, going beyond the status quo of digitization processes accomplished by subjective decisions, could be useful to create philologically faithful preservation copies.

Equalization Analysis

As with the video analysis, machine learning techniques were tested to automatically detect the correctness of the equalization applied during the digitization process. In light of the results obtained by the preliminary test described by Micheloni, Pretto, and Canazza (2017), which was based on audio samples generated in the laboratory, a further step in analyzing a real data set has been taken. The preliminary test highlighted the capability to discern between the different chains of pre- and postemphasis filters (both correct and wrong juxtaposition) using cluster analysis and classification algorithms. The features used were the first 13 Mel-frequency cepstral coefficients (MFCCs). The results were so promising—classification with indexes of accuracy, recall, specificity, and precision very close to, or exactly, 1—that these features were used for the new data set. In this experiment, we focused only on recordings at 7.5 ips, since the previous test shows that the results obtained with a specific speed are very close to the ones obtained with a different recording and playback speed. The samples were extracted from six recordings with a CCIR (C) preemphasis curve and four recordings with an NAB (N) preemphasis curve. These files were taken from audio tapes whose pre-emphasis curves were known. From these we created the samples for the data set, with both the correct and the incorrect filter chains. To collect material that could match the behavior of the white noise tracks of the previous test (Micheloni, Pretto, and Canazza 2017), segments of tracks with only background noise were selected.

Table 5. Distribution of Samples in Each Data Set

Pre-emphasis	*CCIR*		*NAB*	
Post-emphasis	CCIR	NAB	CCIR	NAB
Data set A	74	73	67	69
Data set B	221	221	263	263
Data set C	0	0	40	40

From the samples extracted, three different profiles of noise could be recognized: (1) type A is for loudness between −50 and −63 dB (noise in the middle of a recording, i.e., silence between spoken words); (2) type B for loudness between −63 and −69 dB (noise due to the recording head without any specific input signal); and (3) type C for loudness between −69 and −72 dB (noise coming from sections of pristine tape). In a first step, the audio material collected was then manually divided into small, 0.5-sec segments to extract features for the clustering analysis and the classification algorithm. The resulting data set is shown in Table 5. Some further steps were performed with segments in lengths of 0.6, 0.8, and 1 sec, but they did not enrich the results of the algorithms, so the data presented in the table are related to the first step. We tried to recognize the equalization using the same tools as the previous test, i.e., clustering analysis with hierarchical and *k*-means algorithms, changing different distance metrics, and different classification algorithms like decision tree (DT), *k*-nearest neighbors, and support vector machine (SVM).

The results obtained from the cluster analysis on noise types A and B are as follows. For type A, the samples could be divided into two clusters that differ in their preemphasis equalization. The best result is obtained with the hierarchical algorithm using Euclidean distance and the complete linkage method. In a first cluster it allocates 89 percent of the samples having CCIR preequalization, and in a second cluster 76 percent of the samples having NAB preequalization. Similarly, for type B, the samples able to be divided into two clusters that differ in their preemphasis equalization. The best outcome is obtained with *k*-means algorithm using the cosine distance metric. In a first cluster it allocates 84 percent of the samples having CCIR preequalization, and in a second cluster 79 percent of the samples having NAB preequalization. The difference between these two results is that in the first, many more metrics of the algorithms obtain good clustering results, while in the second fewer metrics can be considered. This can be explained by the fact that the type A noise profile is more characterized by the writing head than is noise of type B, where the loudness of the signal is less than with type A. The results obtained by the classification algorithms on noise types A and B confirm the results obtained from the cluster analysis. With type A, samples that differ in their preemphasis equalization can be discriminated. The best results are obtained with DT and SVM. They highlight the capability of guaranteeing that nearly 95 percent of the samples selected are in the correct class but with a precision of 77 percent. This means that there is a 23 percent probability of taking an incorrect sample from a specific class. This work is not meant to implement a tool that allows modification of the original digitization of the audio tape, but only to test whether it was digitized with the correct equalization. So we want to be sure to have the highest probability of selecting the correct class for a sample and that we generate as few false positives as possible. The results obtained from the classification go in this direction. With type B noises, there is some difficulty in discriminating samples that differ in their preemphasis equalization. As for the cluster analysis, the results are less promising than the those obtained by type A samples but still highlight the trend of good performance for pre-equalization classification. Unfortunately, type C noise samples were found only on tapes with NAB pre-equalization. This simplifies the model of detection both for the clustering analysis and the classification algorithm, and it allows a very important result to be highlighted. In this experiment, this type of noise allows detection of the postemphasis curve with a probability near 100 percent. This result was expected, since pristine tape is not recorded, so the recording head does not introduce other noise in addition to that of the playback head. Obviously, this statement needs to be tested by a further experiment with a larger number of samples. If confirmed, it would allow extracting the postequalization on recordings already digitized by third parties who did not provide any information about the equalization used.

Conclusions

This article has tackled several problems affecting the process of preserving and analyzing electroacoustic music on magnetic tape. After the definition of the philological problems, the concept of discontinuities, and the motivations of this work, two main computational methodologies were presented, concerning discontinuity and equalization detection, respectively. A set of methodologies and software for the automatic analysis of the digitized audio documents has been developed to tackle these problems. These tools aim to help the operators involved in the digitization process by providing specific software able to assist in technical decisions related to the playback device configuration and to validate the results of the process. From the scholar’s point of view, automatic tools that detect discontinuities in audio documents will be a powerful aid for a correct and complete analysis.

For the problem of detecting discontinuities, an unsuccessful attempt excluded the possibility of using audio features to recognize splices. Therefore, an alternative approach using the video of the back of the tape was proposed. The developed tool extracts frames of discontinuities from the video and classifies them with a convolutional neural network. It is the first approach of this kind in the open literature, and the positive results prove its potential. To obtain a more reliable tool, however, a larger and more balanced data set of frames is required to train the convolutional networks. Moreover, further research can improve the detection rate and the classification quality by analyzing the impact of more recent video recording technologies and standards. The training and the validation of this tool were performed by using the digital preservation copies of magnetic tapes collected during the digitization projects of the CSC laboratory. Most of the tapes are from the personal collections of two of highly significant composers of electroacoustic music: Luciano Berio and Luigi Nono. Only a small selection of the hundreds of tapes available were used to extract the approximately 20,000 frames of the data set. The use of the entire collections for the creation of the huge ground truth required for the convolutional network approach could improve the classification performance. During the creation of the data set, several annotations by the composers were discovered and collected in the marks class of the data set. But these projects are currently active and the agreements with the foundations prevent us from giving information about the content of these recordings.

Concerning automatic detection of the equalization used, the methodology provides excellent results, laying the foundation for the development of a robust tool. The experiments proposed in this article confirm the preliminary work described by Micheloni, Pretto, and Canazza (2017). To obtain more reliable results it will be necessary to enlarge the data sets used for training the machine-learning algorithms, adding samples from different tape recorders.

Once these algorithms are perfected, natural further developments taking advantage of these algorithms would be software for automatic quality control, validation kits to aid digitization operators, and analysis tools for scholars. All of these would detect and manage discontinuities as well as points of interest synchronized with audio and video. Currently, these kinds of tools are being developed by the authors.

Finally, the philological issue and the use of these kinds of tools can also be extended to other music genres. Since the 1960s, the same working methods, using physical cuts, editing, and synthesized and "concrete" sounds, have been used in rock, in jazz, and for movie sound tracks, as well as in musical works such as *Prometeo* by Luigi Nono and the musical part of multimedia theatrical works such as *Medea* by Adriano Guarnieri.

References

AES22. 1997. "AES Recommended Practice for Audio Preservation and Restoration: Storage and Handling; Storage of Polyester-base Magnetic Tape." *Journal of the Audio Engineering Society* 45(12):1089–1109.

Bradley, K. 2009. *IASA TC-04 Guidelines in the Production and Preservation of Digital Audio Objects: Standards, Recommended Practices, and Strategies.* London: International Association of Sound and Audio Visual Archives. 2nd ed.

Bressan, F., and S. Canazza. 2013. "A Systemic Approach to the Preservation of Audio Documents: Methodology and Software Tools." *Journal of Electrical and Computer Engineering* 2013: Article 5.

Brock-Nannestad, G. 1997. "The Objective Basis for the Production of High Quality Transfers from Pre-1925 Sound Recordings." In *Proceedings of the 103rd Audio Engineering Society Convention*, pp. 26–29.

Burini, V., F. Altieri, and S. Canazza. 2017. "Rilevamenti sperimentali per la conservazione attiva dei documenti sonori su nastro magnetico: Individuazione delle curve di equalizzazione." In *Proceedings of the Colloquium of Musical Informatics*, pp. 114–121.

Camras, M. 1987. *Magnetic Recording Handbook*. New York: Van Nostrand Reinhold.

Canazza, S. 2007. *Noise and Representation Systems: A Comparison among Audio Restoration Algorithms.* Morrisville, NC: Lulu.com.

Canazza, S. 2012. "The Digital Curation of Ethnic Music Audio Archives: From Preservation to Restoration." *International Journal of Digital Libraries* 12(2–3):121–135.

Canazza, S., C. Fantozzi, and N. Pretto. 2015. "Accessing Tape Music Documents on Mobile Devices." *ACM Transactions on Multimedia Computing, Communications, and Applications* 12(1s): Article 20.

Canazza, S., G. D. Poli, and G. A. Mian. 2010. "Restoration of Audio Documents by Means of Extended Kalman Filter." *IEEE Transactions on Audio Speech and Language Processing* 18(6):1107–115.

Canazza, S., and A. Vidolin. 2001a. "Introduction: Preserving Electroacoustic Music." *Journal of New Music Research* 30(4):289–293.

Canazza, S., and A. Vidolin. 2001b. "Preserving Electroacoustic Music." *Journal of New Music Research* 30(4):351–363.

Casey, M. 2008. “FACET (Field Audio Collection Evaluation Tool): Procedures Manual Version.” Technical report. Indiana University, Bloomington. Available online at www.dlib.indiana.edu/projects/sounddirections/facet/facet_procedures.pdf. Accessed January 2019.

Eilers, D. A. 1968. “Splicing Tapes and Their Proper Application.” *Journal of the Audio Engineering Society* 16(4):472–476.

Fantozzi, C., et al. 2017. “Tape Music Archives: From Preservation to Access.” *International Journal on Digital Libraries* 18(3):233–249.

Fielder, L. D. 1985. “Pre- and Postemphasis Techniques As Applied to Audio Recording Systems.” *Journal of the Audio Engineering Society* 33(9):649–658.

Godsill, S., and P. Rayner. 1998. *Digital Audio Restoration*. London: Springer.

Godsill, S., P. Rayner, and O. Cappé. 2002. “Digital Audio Restoration.” In M. Kahrs and K. Brandenburg, eds. *Applications of Digital Signal Processing to Audio and Acoustics*. Berlin: Springer, pp. 133–194.

Goodman, N. 1968. *Languages of Art: An Approach to a Theory of Symbols*. Indianapolis, IN: Bobbs-Merrill.

IEC. 1994. *Magnetic Tape Sound Recording and Reproducing Systems: Part 1, Specification for General Conditions and Requirements*. Geneva: International Electrotechnical Commission.

ITU. 2003. *Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems*. Geneva: International Telecommunication Union.

Jia, Y., et al. 2014. “Caffe: Convolutional Architecture for Fast Feature Embedding.” In *Proceedings of the ACM International Conference on Multimedia*, pp. 675–678.

Mallinson, J. C. 1976. “Tutorial Review of Magnetic Recording.” *Proceedings of the IEEE* 64(2):196–208.

Micheloni, E., N. Pretto, and S. Canazza. 2017. “A Step toward AI Tools for Quality Control and Musicological Analysis of Digitized Analogue Recordings: Recognition of Audio Tape Equalizations.” In *Proceedings of the International Workshop on Artificial Intelligence for Cultural Heritage*, pp. 17–24.

Miliano, M., ed. 1999. *The IASA Cataloguing Rules*. London: International Association of Sound and Audiovisual Archives. Availablle online at www.iasa-web.org/cataloguing-rules. Accessed January 2019.

MRL. 2016. *Choosing and Using MRL Calibration Tapes for Audio Tape Recorder Standardization*. San Jose, California: Magnetic Reference Laboratory. Version 6.10.1.

NAB. 1965. “Magnetic Tape Recording and Reproducing (Reel-to-Reel).”

Nanni, L., C. Fantozzi, and N. Lazzarini. 2015. “Coupling Different Methods for Overcoming the Class Imbalance Problem.” *Neurocomputing* 158:48–61.

Sueiro, M. 2008. *AVDb, Audio and Moving Image Survey Tool: Instruction Manual*. Columbia University Libraries, New York. Available online at library.columbia.edu/content/dam/libraryweb/services/preservation/AV%20Manual-final.pdf. Accessed January 2019.

Szegedy, C., et al. 2015. “Going Deeper with Convolutions.” In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9.

Verde, S., et al. 2018. “Stay True to the Sound of History: Philology, Phylogenetics and Information Engineering in Musicology.” *Applied Sciences* 8(2): Article 226.

Wang, S., and X. Yao. 2012. “Multiclass Imbalance Problems: Analysis and Potential Solutions.” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42(4):1119–1130.

Zattra, L. 2007. “The Assembling of *Stria* by John Chowning: A Philological Investigation.” *Computer Music Journal* 31(3):38–64.

Zattra, L., G. De Poli, and A. Vidolin. 2001. “Yesterday Sounds Tomorrow: Preservation at CSC.” *Journal of New Music Research* 30(4):407–412.