Dr. Eng. Rafał Korycki
Expert in the field of audio and video engineering
Forensic Bureau ISA

Authenticity investigation of digital audio recorded as MP3 files

Summary

In the work, the problem of detecting discontinuities in lossily compressed audio recordings was outlined and new methods that can be used to examine the authenticity of digital audio records were presented. The described solutions are based on statistical analysis of the data, calculated on the basis of the value of MDCT coefficients. Designated vectors, consisting of 228 features, were used as the training sequences of two machine learning algorithms under the supervision of the linear discriminant analysis (LDA) and the support vector machine (SVM). Detection of multiple compression was both used to detect modification of the recording as well as to reveal traces of montage in digital audio recordings. The effectiveness of the algorithms for the detection of discontinuities was tested on the database of recorded music consisting of nearly one million MP3 files, specially prepared for this purpose. The results of the research were discussed in the context of the influence of parameters of the compression on the ability to detect interference in digital audio recordings.

Keywords authenticity examination of digital recordings, detection of montage, testing digital evidence, double and multiple MP3 compression, MDCT, supervised machine learning methods, support vector machine (SVM), linear discriminant analysis (LDA)

Introduction and purpose of the article

Audio authenticity examinations, which rely on the investigation of the originality of recordings and on the detection of traces of interference in their continuity, are among the essential tasks of modern forensics. The basic scope of investigation related to the analysis of the authenticity of audio recordings was defined by the Supreme Court in the case of Ref. Act III K 49/61 of 10 March 1961. In the ruling, it was written [1]:

"...evidence from magnetic tape acting as physical evidence requires proof of the identity of both the recorded voices and the tape, as well as the lack of any changes to it".

According to the message of the cited judgements, the task of experts, performing caseworks at the request of the judicial authorities and law enforcement agencies, is to determine whether the investigated recording is authentic (specific, identical), as well as revealing any traces of interference in its continuity. The definition of an authentic recording, according to AES27-1996 standard developed by the international organization AES (Audio Engineering Society) can also be cited, which reads as follows:

"As applied to audio recordings, a recording made simultaneously with the acoustic events it purports to have recorded, and in a manner fully and completely consistent with the method of recording claimed by the party who produced the recording; a recording free from unexplained artifacts, alterations, additions, deletions, or edits" [2].

This definition also points to the necessity of examination into the authenticity of recordings and detecting traces of montage, specifying the need for the analysis of the compliance of the adopted recording techniques with those applied during the production of the recording. In addition, attention should be paid to the integrity of the recording with the entirety of its associated acoustic events.

Methods of examining the authenticity of audio recordings have changed over the years. Most of them were created for the analysis of recordings on analogue magnetic tape. After the promulgation of digital recording, some of the previously used methods, such as time course observation and spectrogram recordings, were also used to assess the integrity of digitized recordings [3, 4]. Interference in the continuity of a digital recording can be detected using methods based on the analysis of current

frequency fluctuation of the power grid (ENF Criterion) [3, 5, 6, 7]. The test method has been approved and is widely used by experts belonging to the European Network of Forensic Science Institutes (ENFSI) [8]. It may, however, be applied only when a signal with the frequency of the power grid induces the recording device and is revealed within the recording. The instantaneous frequency of the signal must be determined, and then, the changes in the values of current frequency of the power grid can be compared to the database (e.g. proprietary or obtained from the power grid operator).

The signal of the power grid frequency, from the point of view of manufacturers of recording equipment, is an interference signal, therefore, there are various technical operations aimed at impeding its fixation in the recording, such as shielding elements of the analogue path or the use of high-pass digital filters. Moreover, it happens that the induced signal, associated with the current flow in the power network, is characterized by a short distance from the noise or other disturbances, which sometimes makes it difficult or impossible to use the *ENF Criterion* method effectively. In view of the above and bearing in mind the increasing number of requests for testing the authenticity of digital recordings from the criminal justice and law enforcement authorities, it was necessary to develop effective methods for detecting montage in digital audio recordings.

As for the analysis of video recordings, many more solutions are available for testing their authenticity than in the case of audio recordings. Many methods have been proposed involving, among others, the detection of the double quantization of MPEG-2 compressed videos [9], block analysis of artefacts on the basis of differences in quantization error between the neighbouring blocks [10] and investigation of resolution changes resulting from the integration of another image into the tested one [11, 12]. Methods have also been developed for both the analysis of lossily compressed and uncompressed video recordings [13]. As a result, it is now possible to detect the violation of integrity of individual photos as well as entire video sequences.

However, the above-mentioned methods cannot also be used to test audio recordings because of the fundamental differences in the construction of image and sound encoders. Despite this, one can find some similarities between them. The use of lossy compression of digital audio data causes the adjacent samples of the audio recording to become more dependent on each other (correlated). Also, the frames and sampling the domain of frequency leave some "traces" in the recording, which can be identified. In many publications, authors have pointed out the possibility of obtaining information about the applied encoder based on the compressed recording [14, 15]. Ways of determining the bit rate at which the recording is encoded have been described [16, 17], as well as methods of the detection of double compression [18, 19]. Solutions are also presented to enable the detection of discontinuities in compressed and decoded recordings using modified cosine transform [20].

The objective of this paper is to present research results involving the development of new methods for testing the authenticity of digital audio recordings. The author focuses on the analysis of MPEG-1 Layer 3 lossily compressed recordings [21]. This selection choice was dictated by a considerable amount of literature on the subject, the availability of documentation and the widespread use of the MP3 algorithm. It is worth mentioning that methods presented in this paper can also be used for other compression algorithms that use filter assemblies DCT (*Discrete Cosine Transform*) with perfect or quasi-perfect reconstruction, such as AAC (*Advanced Audio Coding*) or Ogg Vorbis [20, 22]. Based on the analyses described in the above-mentioned publications, the author proposed his own algorithms for initial bit rate prediction, detection of double and triple compression, as well as for detecting montage in repeatedly encoded and decoded recordings. To check the effectiveness of the developed methods, a proprietary database of audio recordings was created, consisting of nearly one million MP3 files. In the opinion of the author, this approach is conducive to the objectification of research and allows easy comparison of research methods. Moreover, the continually popular algorithms for machine learning are used, which can be useful in the recognition and classification of types of recording modification (e.g. re-compression using other parameters, removal of a section of the recording, etc.). Studies have shown that by properly prepared learning sequences, it is possible to create an algorithm which allows examination of the authenticity of MP3 compressed recordings.

For the purpose of describing the research, the author decided to introduce the following terms: "source recording" refers to fragments of music from original CDs, and "original recording" to those subjected to single compression.

Studies have been carried out within the framework of the project: "Design of empirical research and analysis of the materials concerning the specificity of forensic science methods in the work of the special services of public order services" funded by the National Centre for Research and Development in the form of a grant no. 0023/R/ID3/2012/02.

Selected issues related to MP3 compression

The MPEG-1 standard was created as a result of the work of the research group MPEG (Moving Pictures Expert Group), which was formed as the merger of two technical committees of international organizations: ISO (International Organization for Standardization) and IEC (International Electrotechnical Commission). Its task was to develop guidelines for moving picture compression algorithms, along with audio, requiring transmission speeds of less than 1.5 Mbps (megabits per second) [22]. In the MPEG-1 standard, the audio signal can be encoded using three different algorithms, i.e. three different layers, referred to as MP1, MP2, and MP3. The compression standard permits mono or stereo signal, sampled at a frequency of 32 kHz, 44.1 kHz and 48 kHz, providing a bit rate of 32 to 448 kbps (to 320 kbps in the case of MP3) [21]. Given the third layer of the MPEG-1 standard, the compression rate for stereo recordings sampled at a frequency of 44.1 kHz is approximately 1 : 44 do 1 : 4.

The acoustic signal on the input of the MP3 encoder is grouped into frame lengths of 1152 samples and passed through a polyphasic filter (the first set of filters), which divides the frequency domain signal into 32 sub-bands of identical width (in Hertz). Each of 32 sub-band signals is then broken down into 6 or 18 sub-channels (the second set of filters), i.e. the signal samples obtained for each of the sub-bands are multiplied by one of four pre-defined time windows and then the modified discrete cosine transform (MDCT–*Modified Discrete Cosine Transform*) [23] is allotted for them. As a result, for each frame of the input signal there are 576 (long window and two types of transition windows: *start* and *stop)* or three realizations of the 192 (short window) spectrum bands. The decision, which of the divisions is to be used, shall be made based upon psychoacoustic entropy of the signal fragment spectrum, set in a psychoacoustic model block [22].

In parallel, the input signal is grouped into frames of 1024 samples (the remaining samples which are complement to the value of 1152 are rejected), based upon which, the Fast Fourier Transform FFT (*Fast Fourier Transform*) spectral bands are calculated to define a psychoacoustic model. The purpose of the model is to provide information, helpful in switching the length of the time window in the second set of filters and the coefficients used to quantize 576 sub-band samples (following the second set of filters), grouped in predefined frequency sub-ranges reflecting the occurrence of the critical bands (*scale factor bands*) [22]. The psychoacoustic model uses the sound masking effect based on the properties of the human auditory system consisting, to simplify, of the different perception of particular audio signals in the presence of other sounds. For each of the critical band of a given width, in which there is a signal at a set frequency, a masking curve can be marked, below which the other sounds (occurring in this band and beyond it) will not be audible [22]. The distance between the level of the masking tone and the masking curve is referred to as SMR *(Signal to Mask Ratio).* If the masked signal is quantization noise, it will not be heard in this critical band until the SNR *(Signal to Noise Ratio)* is greater than the SMR [22]. Then the values of the 576 samples of the spectrum are quantized with an accuracy dependent on the estimated parameters related to the masking curve, in order to provide the best possible way to "hide" quantization noise. The last stage of the lossy coding process is the creation of the data stream and writing the file.

Extremely important, from the point of view of discontinuity detection in lossy MP3 compression recordings (as well as for other compression algorithms using DCT filters of perfect or quasi-perfect reconstruction), is the process of resetting the part of the 576 samples of the spectrum (obtained after the second set of filters), carried out in the encoder while performing the quantizing operation. Moreover, thanks to one of the properties of MDCT, it is possible to observe the minima of the spectral characteristics of the decoded recording, occurring in the zeroed bands of the spectrum. The following shows the relationship allowing cosine transform designation based on N signal samples $x[n]$ grouped into frames with a 50% overlap [24]:

$$X_{(p)}[k]=\frac{2}{N}\sum_{n=0}^{2N-1}x_{(p)}[n]\cdot h[n]\cdot\cos\left(\frac{\pi}{N}\cdot\left(n+\frac{N+1}{2}\right)\cdot(k+0{,}5)\right) \quad (1),$$

where $0 \le k \le N$ - 1, p is the number of frames, and $h[n]$ is a function of the time window. After applying the inverse MDCT transformation, $2N$ samples are obtained, which also overlap each other [24]:

$$\hat{x}_{(p)}[n]=\frac{2}{N}\sum_{k=0}^{N-1}X_{(p)}[k]\cdot\cos\left(\frac{\pi}{N}\cdot\left(n+\frac{N+1}{2}\right)\cdot(k+0{,}5)\right) \quad (2),$$

where $0 \le n \le 2N-1$. In order to eliminate distortions (*aliasing*), it is necessary to conduct the OLA procedure (*overlap-and-add*), in which the inverse MDCT transformation is marked for the previous and succeeding frames [25]. Then, each of the overlapping segments is multiplied by the corresponding window, and the time domain segments are summed [26]:

$$x_{(p)}[n]=\begin{cases}\hat{x}_{(p-1)}[n+N]\cdot h[N-n-1]+\hat{x}_{(p)}[n]\cdot h[n], & 0\le n\le N-1\\ \hat{x}_{(p)}[n]\cdot h[2N-n-1]+\hat{x}_{(p+1)}[n-N]\cdot h[n-N], & N\le n\le 2N-1\end{cases} \quad (3).$$

If the signal has a local symmetry [26]:

$$\begin{cases} \tilde{x}_p[n] = \tilde{x}_p[N-n-1], \ 0 \le n \le N-1 \\ \tilde{x}_p[n] = -\tilde{x}_p[3N-n-1], \ N \le n \le 2N-1 \end{cases} \quad (4),$$

where:

$$\tilde{x}_p[n] = x_p[n] \cdot h[n], \ \ 0 \le n \le 2N-1 \quad (5),$$

after performing the MDCT transformation the spectrum bands are reseted [26]. Thanks to this feature, it is possible to observe the minima of spectral characteristics of the recording subjected to the compression and decoded only in the same position of the analysis window which was used during the encoding process. Figure 1 shows the spectral characteristics marked for a selected single frame of the recording making use of the time window located in the same position as during the encoding process and shifted by one sample point, respectively "left" (offset = -1) and "right" (offset = + 1). As one can notice, the lack of synchronization resulting from the shift of the window, even by one sample, does not allow to observe the phenomenon of quantization of the levels of spectrum samples.

Extraction of the feature vector

Based on the analysis of ISO standard [21] and preliminary experimental studies, a feature vector was constructed, which allows both examining authenticity as well as detecting montage in audio recordings, subjected to MP3 lossy compression. For its construction, modified the values of discrete cosine transform were used, determined on the basis of binary data encoded in the MP3 file, using the Huffman compression algorithm.

During quantization, the compression algorithm converts the values of MDCT spectrum bands to discrete values. This means that to any variable from the required interval, the values are assigned derived from a finite (countable) set of levels. The process is non-linear, which means that more quantization levels are attributed to low values than the high one. Therefore, it can be assumed that low MDCT coefficients will show a greater sensitivity to double compression [19]. In the paper [17], the authors investigated the possibility of distinguishing between original (single compression) and re-compressed MP3 files, on the basis of the number of MDCT coefficients of the smallest values. It was observed that the original recordings were characterized by a greater number of coefficients adopting levels ± 1 than the recordings undergoing double compression. Another observation is the possibility of analysing these coefficients in conjunction with the frequency ranges in which they occur. It is associated with the

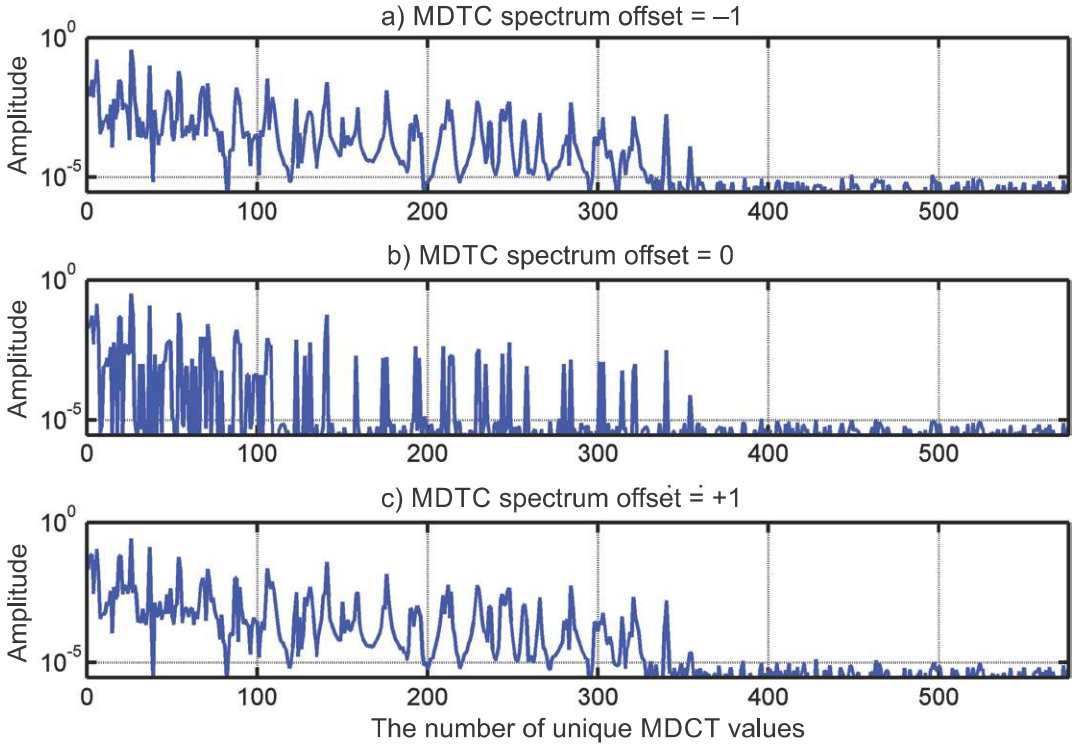

Fig. 1. MDCT spectrum of the decoded fragment of the audio recording computed using a time window offset: a) one sample to the left (offset = -1), b) without movement, c) one sample to the right (offset = +1) relative to the frames set used in the encoder. The amplitude is shown in a logarithmic scale.

filtration process, carried out prior to marking the MDCT transformation (division into 32 sub-bands in the first set of filters), and re-division into the frequency sub-ranges which reflect the occurrence of the critical bands (*scale factor bands*) in the process of quantizing and coding. It can be presumed that the values of spectrum samples, contained in the same sub-range, are mutually correlated to a greater extent than the ones occurring outside of a given band.

During the analysis of each of the MP3 files, the matrix **M** of quantized MDCT coefficients $m_{p,i}$, is obtained, where $0 \leqslant p \leqslant N$ is the number of frames, and $0 \leqslant i \leqslant 576$ denotes the MDCT spectral band index [19]:

$$\mathbf{M} = \begin{bmatrix} m_{0,0} & \cdots & m_{0,575} \\ \vdots & \ddots & \vdots \\ m_{N-1,0} & \cdots & m_{N-1,575} \end{bmatrix} \quad (6).$$

Taking into account the above observations, the author proposed to determine the vector of 228 characteristics used to describe properties of the compressed recordings. On the basis of the magnitudes of the elements of the matrix **M**, a vector including a number of unique non-zero magnitudes (U_i), a vector containing the number of occurrences of these magnitudes(H_i) and a vector which is the measure of the distance between adjacent non-zero unique magnitudes (D_i) were calculated. The first element of feature vector ($F1$) represented the mean number of zero MDCT coefficients per frame [19]:

$$F1 = \frac{1}{N}\sum_{i=0}^{575}\sum_{p=0}^{N-1}\delta(m_{p,i} = 0) \quad (7),$$

where N is the number of frames in the tested MP3 file, so that the function $\delta(m_{p,i} = 0)$ takes the value equal to the unit $m_{p,i} = 0$, and otherwise, zero. The next elements of the feature vector reflected the number of non-zero spectral bands ($F2$), the number of unique MDCT values ($F3$), the average distance between adjacent non-zero unique magnitudes ($F4$) and the variability and continuity of MDCT coefficients ($F5$–$F7$) [19].

$$F2 = \frac{1}{N}\sum_{i=0}^{575}\sum_{p=0}^{N-1}\left|\delta\left(m_{p,i} \neq 0\right) \times m_{p,i}\right| \quad (8),$$

$$F3 = \dim\left(U_i\right) \quad (9),$$

$$F4 = \frac{1}{N-1}\sum_{i=0}^{N-2} D_i \quad (10),$$

$$F5 = \sqrt{\frac{1}{N-1}\sum_{i=0}^{N-2}\left(D_i - F4\right)^2} \quad (11),$$

$$F6 = \frac{1}{N-1}\sum_{i-0}^{N-2}\left[D_i \times \left(h_i + h_{i+1}\right)\right] \quad (12),$$

$$F7 = \sqrt{\frac{1}{N-1}\sum_{i-0}^{N-2}\left\{\left[D_i \times \left(h_i + h_{i+1}\right)\right] - F6\right\}^2} \quad (13).$$

The author also proposed the appointment of three characteristics that describe the average number of instances of quantized MDCT spectrum bands with the magnitudes equal ±1, ±2 and ±3:

$$F8 = \frac{1}{N}\sum_{i=0}^{575}\sum_{p=0}^{N-1}\delta\left(\left|m_{p,i}\right| = 1\right) \quad (14),$$

$$F9 = \frac{1}{N}\sum_{i=0}^{575}\sum_{p=0}^{N-1}\delta\left(\left|m_{p,i}\right| = 2\right) \quad (15),$$

$$F10 = \frac{1}{N}\sum_{i=0}^{575}\sum_{p=0}^{N-1}\delta\left(\left|m_{p,i}\right| = 3\right) \quad (16).$$

Furthermore, due to the division of the MDCT spectrum into 22 sub-ranges for long windows (*scalefactor bands*) reflecting the existence of critical bands, for each of them the additional $F3$–$F7$ features were set [19]. During the preliminary experimental studies, an increase in efficiency of classification algorithms was also observed, where in each of the mentioned sub-ranges, the characteristics of $F1$ and $F8$ were also calculated. The remaining 64 variables of the feature vector were marked on the basis of the method proposed in the article [18]. The presented algorithm has been slightly modified by the author in connection with the need to replace the detection threshold of 0.00001 with the value of 1, since the decoded MDCT coefficients, based on the Huffman data, have integer values. Thus the prepared vector of 228 characteristics, determined for each investigated recording (or the portion of the recording), can then be used as the input data for the classification algorithms.

Machine learning algorithms

Machine learning under supervision involves inference of an unknown function, based on the training data sequence. The training data consist of a set of examples that contain a data vector and the desired output state (called the control signal). The algorithm analyses the training sequences and deduces an unknown function (of the model), which is called a classifier (in the case of discrete data) or the regression function (for continuous data). A properly selected classifier should correctly predict the output state for each correct set of data [27].

The multilayer perceptron (MLP) network is currently the most popular type of artificial neural network. The perceptron model is a single neuron with the linear function network of weight w, and a threshold activation function [29]. The input data comprise the feature vector of $x = (x_1, x_2, \ldots, x_n)$ in n-dimensional feature space. The function of the network is a weighted sum of inputs [29]:

$$u(x) = w_0 + \sum_{i=1}^{n} w_i x_i \quad (17),$$

however, the output function is obtained on the basis of the activation function [29]:

$$y(x) = \begin{cases} 1 & u(x) \geq 0, \\ 0 & u(x) < 0 \end{cases} \quad (18).$$

The learning process takes place on the basis of the training sequence, as follows [29]:

$$w(k+1) = w(k) + \eta\big(d(k) - y(k)\big)x(k) \quad (19),$$

where η determines the adaptation speed, $d(i) \in \{0,1\}$ is the desired state at the output of the classifier for a particular training sequence $\{(x(i), d(i))\}$. Index k is used herein to indicate that the learning vectors are transmitted sequentially to the perceptron, but in random order. If the state at the output of the perceptron $y(k)$ is consistent with the desired state $d(k)$, then the weight vector is not changed. Otherwise, vector w is updated.

The classic artificial neural network together with the learning algorithms, making use of non-linear minimization of the error function, shows some disadvantages [30]. The error function can adopt a lot of local minima, in which the learning algorithm may stop without reaching the optimal solution. A support vector machine (SVM), which appeared in recent years, presents a new approach in constructing and training neural networks, free from the defects of previously used algorithms [31]. SVM uses a learning algorithm that maximizes the margin of separation between the two classes defined by the set p of the pairs containing the input data vector x and the class d to which it belongs [32]:

$$D = \{(x_i, d_i)\}_{i=1,2,\ldots,p}, \quad x \in X, \quad d_i \in \{-1,1\} \quad (20).$$

The optimal hyperplane is defined as a linear decision function with the maximum margin of separation between the two classes of vectors. For constructing such a hyperplane, only some of the selected training vectors should be used, the so called carrier or support vectors, which define the margin of separation. In the case of training vectors that are linearly separable, the hyperplane takes the form [32]:

$$g(x) = w^T x + b = 0 \quad (21),$$

where w are the normal vectors to the hyperplane, b is the shift constant, and T the transposition operator. Determining the widest margin comes down to the solution of the main problem of optimization [3]:

$$\min_{w} \frac{1}{2} \| w \|^2 \quad (22)$$

with the following limitation [3]:

$$d_i\left(w^T x_i + b\right) \geq 1, \quad i = 1,\ldots,K \quad (23).$$

The above-mentioned optimization problem can be represented by a single expression that is minimized due to w i b and maximized due to Lagrange multipliers $(a_1, a_2, \ldots, a_k)$ [3]:

$$\min_{w,b} \max_{\alpha} \left\{ \frac{1}{2} \| w \|^T - \sum_{i=1}^{K} \alpha_i \left(d_i \left(w^T x_i + b \right) - 1 \right) \right\} \quad (24).$$

The relationship between α_i and the position of the vector x_i relative to the margin, is described by the Karusha–Kuhn–Tucker conditions (KKT). They show that the coefficients α_i take values other than zero only in those equations that contain support vectors [3, 4]. Therefore, the classification function is dependent only on a small number of support vectors S in comparison with the total number of training vectors.

Another popular method used for finding a linear combination of features that best characterize or isolate two or more classes of objects or events, is discriminant analysis [33]. It consists in the projection of multidimensional input data vectors on the space of a smaller number of dimensions by simultaneous maximizing the separation of data from different classes and minimizing the dispersion of data belonging to the same class. In this way, the maximum ability to classify the data vectors and reducing their dimensionality is obtained at the same time.

The linear discriminant analysis (LDA) is implemented by gathering information about the classes, which can be attributed to new observations with the aid of position indicators and the dispersion of sub-samples of the learning sample. The mean values for the classes can be defined as [34, 35]:

$$\overline{x}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ki} \quad (25),$$

where k is the number of populations, whereas n_k is a sub-sample of observations from a given class. Assuming that x_k is the mean of a group and the variance of vector data is the same for all g populations, a common covariance matrix can be set on the basis of the intra-group covariance matrix [34, 35]:

$$W = \frac{1}{n-k} \sum_{k=1}^{g} (n_k - 1) S_k = \frac{1}{n-k} \sum_{k=1}^{g} \sum_{l=1}^{n_k} (x_{kl} - \overline{x}_k)(x_{kl} - \overline{x}_k)^T \quad (26),$$

where S_k is the covariance matrix of the k samples of the population $n = n_1 + n_2$. The purpose of the LDA method is to find a direction a that maximizes the distance between projected mean trials, taking into account the variation of the projection [34, 35]:

$$\arg\max_{a}\left(\frac{\left(a^T\bar{x}_2 - a^T\bar{x}_1\right)^2}{a^TWa}\right) \Rightarrow a = W^{-1}\left(\bar{x}_2 - \bar{x}_1\right) \quad (27)$$

assuming the presence of the two classes. The classification is performed by projecting the observation *x* in the direction of *a* and assigning it to a definite class, depending on whether the projection is closer to the centre of projection of the first or second trial. In the case of a greater number of classes, such *a* direction is sought that would maximize the following expression [34, 35]:

$$\arg\max_{a}\left(\frac{a^TBa}{a^T\mathrm{W}a}\right) \quad (28),$$

where in *B* is the inter-group covariance matrix, and *W* is the intra-group covariance matrix [34, 35]:

$$B = \frac{1}{g-1}\sum_{k=1}^{g} n_k\left(\bar{x}_k - \bar{x}\right)\left(\bar{x}_k - \bar{x}\right)^T \quad (29),$$

$$W = \frac{1}{n-g}\sum_{k=1}^{g}\left(n_k - 1\right)S_k \quad (30).$$

Database of compressed recordings

To evaluate the effectiveness of the detection of interference in the content of the recording by the suggested detection algorithms, a database, designed and created especially for this purpose, was used. The subject of the study was thought to be materials recorded in conditions similar to real ones, using recorders of unknown parameters. Such recordings are characterized by the presence of noise and interference and it is difficult to predict the shape of the spectral characteristics. Therefore, the attempt at copying accurately these "ideal" real conditions was abandoned and, as the reference material, all kinds of music recordings were used. Such solution makes it possible to ensure reproducible research and similar solutions are practised in the scientific community, in particular, with regard to the problem of detecting modifications to digital audio recordings [4, 17, 18, 19]. Also, the attention should be paid to the lack of available sources of recordings that could be useful to examine the authenticity of audio recordings. The existing database of distorted speech recordings contains files recorded with sampling frequencies less than or equal to 16 kHz, which does not correspond to the typical values described in the ISO standard for MP3 algorithm [21]. It should also be noted that some recordings of music reflect perfectly the wide variety of artefacts that appear in the actual recordings of the evidence, such as crackle, hitting (drums), band noise (keyboards, effects, electric guitars), and sinusoidal interference (string instruments).

Accordingly, 2940 pieces of music (of the length of 10 seconds each) from 36 albums from the private collection of the author were used in the research. Among the files added to the database, there were, among others, records of symphonies by Ludwig van Beethoven performed by the London Symphony Orchestra, three symphonies by Wolfgang Amadeus Mozart, conducted by Rafael Kubelik, a compilation of *Adagios* signed by Herbert von Karajan, the conductor, digitally restored versions of the works of Ray Charles, two Norah Jones albums and another two of Iron Maiden, the live album "Pulse" of Pink Floyd and the Beyonce album "4". All source files (2940) captured in the linear and lossless LPCM format (*Linear Pulse Code Modulation*) and stored in "wav" files were compressed by means of 21 different implementations of MP3 encoders (see Table 1), in each case using six available rates: 64, 96, 112, 128, 192 and 256 kbps (kilobits per second). Four of these coders were an equipment implementation in the form of portable recorders (Zoom, Tascam, Olympus, Sony), so in their case, the music was played using the active speakers Adam P22 and recorded taking advantage of built-in microphones.

The recordings were compressed using the three selected implementations of the MP3 algorithm: Lame 399, iTunes and Adobe CS6 were then decoded into lossless format (LPCM with the extension of "wav") using decoders corresponding to the coders, which were contained in the same applications (for the Lame 399 encoder, also appearing alone, the application Lame Front-End 1.8 was used, containing both Lame 399 encoder and the decoder). Then the copies of the decoded fragments of the recordings were transferred using a cyclic buffer (*circshift* function in the MATLAB computing environment) of 10 ms [36]. The implemented shift was aimed at representing montage involving cutting or inserting a fragment of speech that can be executed in the time domain after prior decoding of a recording to lossless format. In the process of compressing the input data, are organized into frames and then processed in such form. If montage has occurred, the existing order of frames will be affected. To restore the original layout, the appropriate transfer, identical for each next frame, from the place of performing the edition, should be made. Thus, by introducing the transfer, e.g. by 10 ms., it is possible to model the occurrence of montage with respect to those fragments of a recording that follow the location of the montage.

Thanks to the above operations, the material was obtained for examining the files in which the montage

Table 1

List of MP3 encoders used in studies

Designation	Name of encoder/publisher or creator	Version (year)
Lame 396	LAME/Mike Cheng, Mark Taylor et al.	3.96.1 (2004)
Lame 398	LAME/Mike Cheng, Mark Taylor et al.	3.98 (2008)
Lame 399	LAME/Mike Cheng, Mark Taylor et al.	3.99.5 (2012)
Adobe 2	Adobe Audition/Adobe Systems Inc.	2.0 (2005)
Adobe CS6	Adobe Audition CS6/Adobe Systems Inc.	5.0 (2012)
iTunes	iTunes/Apple Inc.	10.3.1.55 (2011)
8 Hz	8Hz MP3 Encoder/8Hz Production	02b (1998)
BladeEnc	BladeEnc/Tord Jansson	0.94.2 (2001)
GoGo	Gogo-no-coda/Herumi and Pen	3.13 (2004)
Helix	Helix/RealNetworks Inc.	5.1 (2005)
Mp3enc	Mp3Enc (Demo)/Fraunhofer IIS-A	3.1 (1998)
Plugger	Plugger+ Pro/Alberto Demichelis	0.4 (1998)
SCMPX	SCMPX/Shinji Chiba	1.51 (1999)
Shine	Shine/Gabriel Bouvigne	0.1.4 (2001)
SoloH	Mpeg Encoder/SoloH	0.07a (1998)
Real	Real Player/RealNetworks Inc.	15.0.6.14 (2012)
Xing	Xing MP3 Encoder/Xing Tech. Corp.	1.5 (1999)
Zoom	Zoom Handy Recorder/Zoom	H4n (b.d.)
Tascam	Linear PCM Recorder/Tascam	DR-40 (b.d.)
Olympus	Digital Voice Recorder/Olympus	WS-812 (b.d.)
Sony	ICD Recorder/Sony	UX513F (b.d.)

in the form of removal or insertion of a fragment of the recording was made. The prepared recordings, in lossless format, were then re-compressed, using the same three encoders (Lame 399, iTunes and Adobe CS6) with the following bit rates: 112, 128 and 192 kbps. The created database of recordings consisted in total of nearly one million MP3 files. The study, conducted within the framework of the project entitled "Planning the empirical research and analysis of the material on the specifics of the methods of forensic science in the work of the special services of public order services" included various topics, such as identification of the type and version of the encoder and the recording equipment based on the analysis of MP3 files, and the impact of training models on classification accuracy. This study, due to the need to reduce its volume, contains results of research carried out mainly on the files compressed using only Lame 399 encoder. It is worth noting that the encoders of the Lame family are the most common implementations of the MP3 algorithm, and they are in use in the majority of free programs which are used to edit digital audio recordings.

The research took advantage of the MATLAB computing environment, in particular, the following tool packages: *Signal Processing Toolbox*, *Parallel Computing Toolbox*, *Statistics Toolbox* and the *Bioinformatics Toolbox* [36]. For the classification, a linear discriminating analysis was applied, with the function of inverting the matrix by the Moore-Penrose method [33, 36], and a support vector machine with a core linear function. For each of the performed classifications, feature vectors representing one group (e.g. recordings subjected to single and double compression or recordings unmontaged and edited) were divided into two equal sequences: training and testing. The training sequence was used to create models by LDA and SVM algorithms, whereas the classification efficiency, and therefore also the accuracy of interference detection in the recording, was examined in the test sequence.

Determination of bit rate

The quality of the MP3 recording is measured, in addition to, among others, sampling frequency, in the value of its bit rate expressed in kilobits per second (*kbps*). Based on this, one may, for example, decide to purchase the digital version of an album in an online shop. Since the bit rate value is stored in the file and only indicates the parameters of the last made compression, it is expected that some of the recordings can be intentionally transcoded (i.e. decoded into a lossless format and re-encoded with other parameters, e.g. changing bit rate) from a lower to higher bit rate in order to demonstrate their superior quality. Such actions constitute fraud and significantly reduce the confidence of internet trade. Recordings can also be transcoded to hide

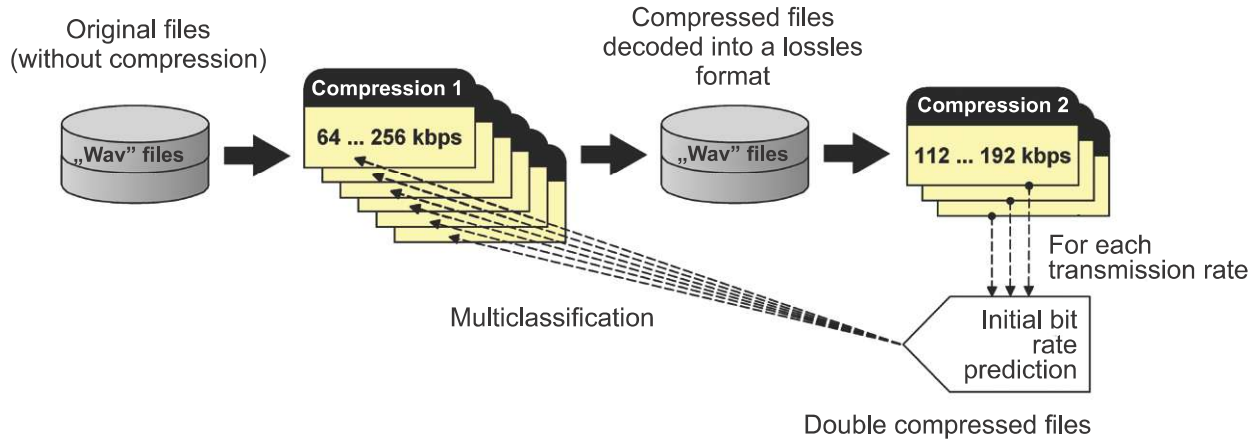

Fig. 2 Audio compression and data comparison scheme used in revealing the original quality of double compressed MP3 files.

traces of interference made to their content. The deliberate lowering of the quality of recording by re-compressing it at a lower rate than it was originally can also be expected.

In order to check the possibility of detecting interference involving transcoding MP3 files, the recordings from author's database were used, based on the schema shown in Figure 2. The source files, music fragments, were compressed using six bit rates, decoded into a lossless format, and then re-compressed, this time using three selected transmission rates: 112, 128 and 192 kbps. The MP3 files created isolated feature vectors (each consisting of 228 elements), one for each file. For the classification, linear discriminant analysis was used. In order to perform classifications, the feature vectors were divided into two equal sequences: training and testing. In this case, multiclassification of the recordings undergoing double compression consisted of predicting the bit rate used during the first compression. In tables 2-4, the initial bit rate predictions (in percentage) are presented, depending on the secondary bit rate, 112 kbps, (Table 2), 128 kbps (Table 3) and 192 kbps (Table 4) respectively. For example, all the transcoded files, from bit rates of 64 kbps and 96 kbps to 112 kbps were classified correctly, while the number of correct conversion from 256 kbps to 112 kbps amounted to 81.63% (see Table 2).

Analysis of these results shows that the prediction of the source rate in the recordings subjected to double compression is more accurate in the case of predicting lower values of the rate than those used in the second encoding. However, with regard to the recordings transcoded from a greater to lower rate, it is still possible to make the correct classification of the converted files.

Detection of double and triple compression

Predicting the source bit rate of MP3 files subjected to double compression is carried out under the assumption that the recording, after its fixation, was coded again. It is therefore necessary to make it possible to distinguish the original recordings

Table 2
Classification matrix of 112 kbps double compressed MP3 audio files

Rate predicted	Rate used during the first compression [kbps]					
	64	96	112	128	192	256
64	100	–	–	–	–	–
96	–	100	–	–	–	–
112	–	–	96.26	2.38	0.34	1.02
128	–	–	1.36	92.52	1.36	2.04
192	–	–	0.68	1.70	71.43	15.31
256	–	–	1.70	3.40	26.87	81.63

Table 3
Classification matrix of 128 kbps double compressed MP3 audio files

Rate predicted	Rate used during the first compression [kbps]					
	64	96	112	128	192	256
64	99.66	–	–	–	–	–
96	0.34	99.32	–	–	–	–
112	–	0.68	100	–	–	–
128	–	–	–	94.90	1.36	2.38
192	–	–	–	1.70	85.71	10.20
256	–	–	–	3.40	12.93	87.41

Table 4
Classification matrix of 192 kbps double compressed MP3 audio files

Rate predicted	Rate used during the first compression [kbps]					
	64	96	112	128	192	256
64	100	–	–	–	–	–
96	–	99.66	–	–	–	–
112	–	0.34	99.66	–	–	–
128	–	–	0.34	100	–	–
192	–	–	–	–	**81.97**	**17.35**
256	–	–	–	–	**18.03**	**82.65**

(subjected to single compression) from those which were re-compressed. In addition, the person who edited the recordings is usually guided by the need to modify the voice content, disregarding the specific settings of the audio files, among those, the compression data. Therefore, it may be expected that after editing, the recording will be fixed with the default set of options, or with the same selection of parameters as the ones that characterized the original recording.

Taking into account the above-mentioned considerations, it was decided to check the possibility of detecting the entries subjected to double compression. For this purpose, the database of recordings created by the author was used according to the schema shown in Figure 3. The fragments were compressed using six bit rates, decoded into a lossless format, then compressed again using the same three selected bit rate values (112, 128, and 192 kbps). Feature vectors were isolated from the created MP3 files which were then classified using linear discriminant analysis and the support vector machine with a kernel linear function.

Table 5
Binary classification of single and double compressed MP3 audio files using SVM algorithm and different initial bit rates

Bit rate during the first compression	Rate during the re-compression [kbps]		
	112	128	192
64	100	100	100
96	99.83	100	100
112	**98.64**	100	100
128	96.26	**96.09**	100
192	89.63	91.16	**89.80**
256	75.85	82.82	83.67

Table 6
Binary classification of single and double compressed MP3 audio files, using LDA algorithm and different initial bit rates

Bit rate during the first compression	Rate during the re-compression [kbps]		
	112	128	192
64	100	99.83	100
96	98.83	99.83	100
112	**98.81**	100	100
128	97.62	**98.30**	100
192	91.67	94.39	**87.93**
256	72.79	79.25	86.56

Regarding their classification, each of the feature vectors were divided into two equal sequences: training and testing. In this case, the binary classification was performed based on distinguishing recordings that underwent single and double compression,

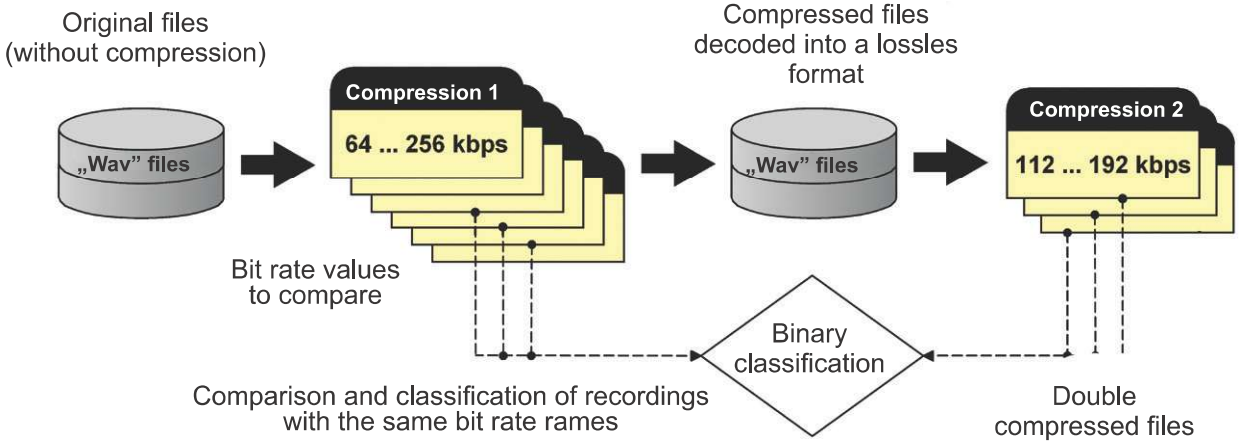

Fig. 3. Audio compression and data comparison scheme used in investigations involving the detection of double MP3 compression.

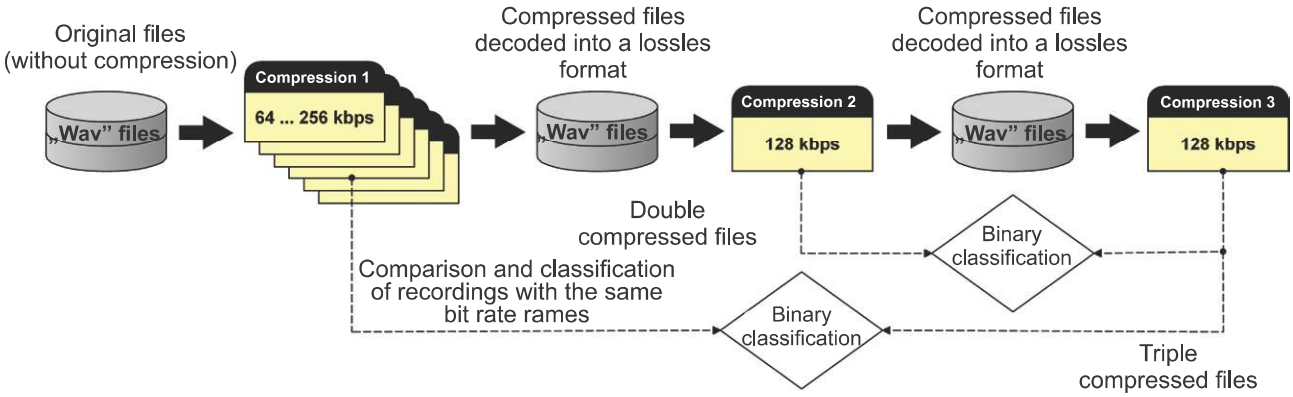

Fig. 4. Audio compression and data comparison scheme used in investigations involving the detection of triple MP3 compression.

wherein designated vectors based on files of the same parameters were compared, e.g. 96 kbps compressed files converted to 128 kbps compared with the recordings compressed using a 128 kbps bit rate. Shown in Tables 5 and 6 are the results of the double-compression detection (in percent) in relation to the secondary bit rate (respectively: 112 kbps, 128 kbps and 192 kbps). For example, during classification, using the SVM algorithm, of the converted recordings from the rate of 192 kbps to 128 kbps and their original counterparts (subjected to single compression of a bit rate of 128 kbps), 91.16% of the files were correctly associated with matching classes (see Table 5).

The person editing the recording may, however, want to do more than one additional compression, for example, to hide traces of modifications to the content of the recording. Therefore, it was necessary to study the possibility of detecting such modified recordings and determine how many times the relevant material was converted. As before, for this aim, the database recordings were used, performed according to the scheme shown in Figure 4. The sound track fragments were compressed using six bit rates, decoded to a lossless format, and then compressed twice with a 128 kbps bit rate. From the created MP3 files, the feature vectors were separated, one for each file. To carry out the classification, linear discriminant analysis and a support vector machine were used.

In the case of the performed classifications, the feature vectors were divided into two equal sequences: training and testing. In this case, the binary classification was based on distinguishing recordings that underwent single and triple compression and double and triple compression, meanwhile the feature vectors, determined on the basis of files of the same parameters, were compared. Table 7 presents the results of comparisons being made (in percent) in relation to the number of modifications made in the recordings. For example, during classification, by means of the LDA algorithm, recordings that underwent single compression (128 kbps) as well as converted files with the rate of 192 kbps to 128 kbps, and then again compressed (rate of 128 kbps), 99.32% of the recordings were properly assigned to the appropriate class (see Table 7).

On the basis of the obtained results, it can be seen that there is the possibility of detecting recordings

Table 7

Binary classification of single, double, triple and MP3 compressed audio files using different initial bit rates. Files were compressed for a second and third time with 128 kbps bit rate

Rate during the first compression	Compared files from various stages of compression			
	SVM		LDA	
	1 Compression – Compression 3	2 Compression – Compression 3	1 Compression – Compression 3	2 Compression – Compression 3
64	100	83.33	99.83	79.59
96	100	89.12	100	84.18
112	99.83	90.48	100	88.78
128	**95.92**	**77.38**	**99.15**	**75.51**
192	99.49	98.13	99.32	98.47
256	96.94	97.45	98.64	98.30

subjected to double and triple compression. Interestingly, the introduction of additional processing makes the triple-encoded recordings easier to distinguish from the original recordings (subjected to single compression) than is in the case of files encoded twice. Worse results were obtained in the case of comparing files subjected to double and triple compression. However, despite this, the experiments which were carried out show that it is possible to reproduce the type and order of modifications which affected the examined material. Important is the creation of appropriate models for the classification algorithms.

Detection of discontinuities in compressed recordings

Editing of an MP3 file, consisting of excision or insertion of a fragment, may be made after prior decoding of the audio recording to a lossless format. As mentioned, in the compression process, the input data is arranged in frames, and then is processed in such form. If montage has occurred, the existing frame order will have been affected. To restore the original layout, the appropriate transfer, identical for each of the next frame, from the place of performing the edition, should be made. The problem of measuring the transfer and using it to detect the interference in the tracks of the recording, subjected to compression and decoded to a lossless format, was earlier the topic of the author's conclusions and it will not be discussed in detail here [3: 20].

While carrying out research for the purpose of this study, the possibility of using the mentioned phenomenon in order to examine the integrity of records subjected to double and – in general – multiple compression was recognized. Violation of the frame grid, created in the process of lossy coding of a recording before the re-compression of tracks, will actually change the shape of the spectrum marked by the MP3 algorithm, as in the case illustrated in Figure 1. In light of the above, any operations that are performed by the encoder, and in particular, the estimation of the threshold of masking and quantization, depending on the estimated characteristics of masking, will be determined by the distribution of MDCT transformation bands.

With regard to the above considerations, it was decided to check the possibility of detection of double compression in the recordings. Files from the created database were used for this purpose according to the schema shown in Figure 5. Excerpts from the tracks were compressed using six rates, decoded to lossless format and then their copies were transferred in the circular buffer by 10 milliseconds. The files prepared in this way (with and without transfer) were also re-compressed using the three selected rates (112, 128, and 192 kbps). For the classification of feature vectors, separated from the MP3 files, as in the previous cases, linear analysis and support vector machine were used.

In the case of each of the carried out classifications, the feature vectors were divided into two equal sequences – testing and training. Binary classification consisted of distinguishing between the edited recordings (which is shifted by 10 ms frames) and those non-edited (without moving the frames), and the feature vectors marked based on files of the same parameters were compared. Tables 8 and 9 contain the results of the detection of discontinuities (in percent) depending on the secondary bit rates (respectively: 112 kbps, 128 kbps and 192 kbps). For example, during classification of the recordings, using the SVM algorithm (the first compression performed with the rate of 128 kbps) transferred in the circular buffer and unedited, which

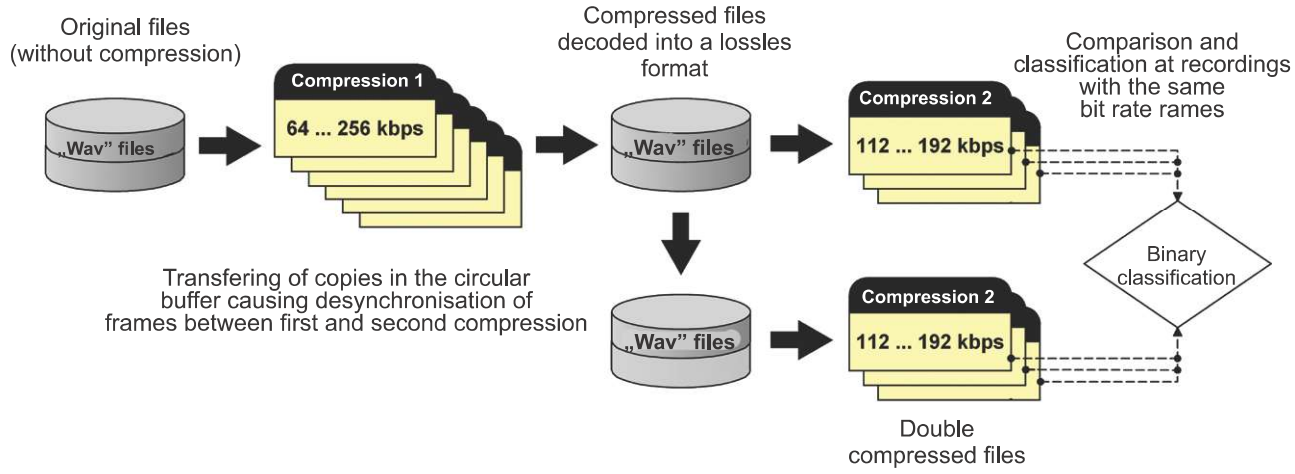

Fig. 5. Audio compression and data comparison scheme used in detection of montage in double compressed MP3 audio files.

Table 8
Binary classification using SVM algorithm and different initial bit rates for the detection of montage in double compressed recordings

Rate during the first compression	Rate during the re-compressing [kbps]		
	112	128	192
64	100	100	100
96	99.66	100	100
112	**98.13**	100	100
128	96.43	**96.26**	100
192	89.46	90.31	**92.69**
256	71.26	78.74	88.33

Table 9
Binary classification using LDA algorithm and different values of the initial bit rate involving montage detection in double compressed audio files

Rate during the first compression	Rate during the re-compression [kbps]		
	112	128	192
64	100	99.83	100
96	100	100	99.83
112	**99.32**	100	99.66
128	97.79	**98.81**	100
192	90.82	94.22	**92.18**
256	69.73	79.76	83.67

were then compressed again with the bit rate of 112 kbps, 96.43% of the recordings were properly assigned to the appropriate class (see table 8).

As in the case of detection of triple compression, the possibility of detecting discontinuities in these recordings, which after editing were compressed two more times, was also checked. The investigation was conducted in accordance with the scheme shown in Figure 5, performing additional decoding and a third compression (with a rate of 128 kbps). From the created MP3 files, feature vectors were separated, and for their classification linear discriminatory analysis and machine support vectors were also used. The binary classification consisted in distinguishing between edited and unedited recordings. Shown in table 10 are the results of the detection of discontinuities (in percent) depending on the classification algorithm used. For example, during classification of the recordings compressed with the help of SVM (first compression of 112 kbps) being transferred in a circular buffer and unedited, and then re-compressed twice with the throughput of 128 kbps,

Table 10
Binary classification involving the detection of montage in triple compressed audio files

Rate during the first compression	Rate during the re-compression [kbps]	
	SVM	LDA
64	100	99.83
96	100	100
112	99.66	100
128	**84.01**	**84.69**
192	90.48	89.63
256	64.63	64.29

99.66% of the recordings were properly assigned to their class (see table 10).

Analysing the obtained results, it can be seen that taking advantage of the occurrence of changing the distribution of MDCT spectral bands in the case of breaching the frame grid created during lossy encoding, it is possible to detect discontinuities in compressed recordings. In addition, the introduction of additional processing makes the identification of montage still possible, in particular, for the low bit rate values used during the first encoding.

Summary

Results of the research, carried out within the framework of the research project, show that it is possible to detect a variety of interference in lossily compressed MP3 audio files. Making use of available analyses, the author put forward his own algorithms for the initial bit rate prediction, detection of double and triple compression as well as the detection of montage in audio files repeatedly encoded and decoded. To check the effectiveness of the developed methods, a proprietary database of audio files was created, consisting of nearly one million MP3 files. Studies have shown that by properly prepared learning sequences, it is possible to create an algorithm which allows examination of the authenticity of recordings compressed in MP3. In addition, according to the definition of authentic recordings according to standard AES27-1996 [2], it can be noticed that in the framework of the research the analyses of recording originality is an effective attempt to reveal traces of montage.

As discussed in this paper, methods of examining the authenticity of lossily compressed recordings can be used to test MP3 as well as other compression algorithms, which use sets of DCT filters of either perfect or quasi-perfect construction, for example

AAC or Ogg Vorbis. It is also worth emphasizing that in connection with using, primarily, small samples of the MDCT spectrum for creation of the feature vector, the efficiency of the proposed methods cannot depend on the shape of the envelope of the spectral characteristics of the examined audio records, which undoubtedly is the case, among others, for methods of voice recognition of speakers and other solutions involving the analysis and classification of the contents of the recordings (including music recordings). In view of the above, the presented methods of examining the authenticity of the MP3 compressed recordings can also be used in relation to the actual evidential recordings.

Due to the increasing interest of contemporary forensic science in the authenticity examination of digital audio recordings, the continuous development and automation of new methods of research are necessary. Available solutions, such as the analysis of current frequency changes in the power grid, examining the changes in sound background using methods of time-frequency or assigning the lossy compresion recording parameters, as far as possible, should be used together. After all, one should bear in mind the fact that the lack of detection of locations of montage does not preclude the fact of the unauthorized recording modification. It may show, among other things, the use of yet unexplored recording or editing techniques, or the lack of available analytical methods.

Research carried out within the framework of the project entitled: “Design of empirical research and analysis of the materials concerning the specificity of methods of forensic science in the work of the special services of public order services” funded by the National Centre for Research and Development in the form of a grant no. 0023/R/ID3/2012/02.

Source

Fig. 1–5: author
Tab. 1–10: own elaboration, results published in [37]

Translation *Ronald Scott Henderson*