Feature-Based Classification for Audio Bootlegs Detection

P. Bestagini [1], M. Zanoni [2], L. Albonico [3], A. Paganini [4], A. Sarti [5], S. Tubaro [6]

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano
Piazza Leonardo da Vinci 32, 20133 Milano, Italy
{[1] bestagini/[2] zanoni/[5] sarti/[6] tubaro}@elet.polimi.it
{[3] luca1.albonico/[4] andrea3.paganini}@mail.polimi.it

***Abstract*—In the past few years, thanks to the increasing availability of multimedia sharing platforms, the online availability of user generated content has incredibly grown. However, since media sharing is often not well regulated, copyright infringement cases may occur. One classic example is the pirate distribution of audio bootlegs, i.e., concerts illegally recorded using portable devices. In order to guarantee copyrights and avoid the sharing of such illicit material, in this paper we propose an automatic audio bootleg detector. This can be used to analyze audio data in bulk, in order to filter out from a database the audio tracks recorded, e.g., by fans during a live performance. To this purpose, we propose to use a set of acoustic features to characterize audio bootlegs, justified by theoretical foundations. Then, we train a binary classifier that operates on this set of features to discriminate between: i) audio tracks recorded at either concerts, clubs, or theaters; ii) legally distributed live performances professionally mixed and edited. In order to validate our system, we tested it on a dataset of more than 250 audio excerpts considering different musical genres and different kinds of music performances. The results achieved are promising, showing a high bootleg detection accuracy.**

I. Introduction

Nowadays, with the rapid proliferation of inexpensive hardware devices that enable the acquisition of high-quality audio-visual data, the possibility to generate multimedia objects is within everyone's reach. Modern smart-phones and digital camcorders are, indeed, often equipped with small microphones or microphone arrays that allow high-quality recordings. Furthermore, the increasing availability of multimedia sharing platforms has encouraged the widespread diffusion of audio-visual content on the Internet. However, when not strictly regulated, media sharing often gives rise to copyright infringement cases and legal issues. Indeed, it is customary to find online illegal copies of copyrighted audio-visual material. A typical example is the distribution on websites such as Youtube of concerts recorded from the audience without permission. Another common scenario is the distribution of movies illegally captured at theaters before their official DVD release date.

In order to detect illegal or inauthentic audio-visual content, the multimedia forensic community has proposed a series of detectors targeting different kinds of media (i.e., images, video, and audio) over the years. Nonetheless, the most contributions in the literature focus on still-images, [1][2], and few effective solutions have been proposed for video [3] and audio [4] content. With a specific regard to audio, the media that we consider in this study, we can group the algorithms in two broad categories: i) methods that digitally sign audio content during the acquisition step using watermarking techniques [5][6]; ii) blind methods based on detecting traces left by processing operations [7][8].

In this paper we face the problem of audio bootlegs detection, and to this end we present a detector that belongs to the latter category, i.e., blind methods that do not need any control on the acquisition step. Notice that the definition of bootleg that we consider in this study slightly differs from the one that music and movie production commonly refer to. In the production industry, in fact, the term is used to point out all the media content not officially released. However, this definition is purely related to laws and it is only based on copyright treaties. From the audio forensics perspective, we are more interested in detecting audio bootlegs as audio recordings captured with portable (e.g., hand-held) devices and possibly re-distributed. According to this definition, audio bootlegs we consider in this study are live concerts captured by the audience (e.g., some fans) at a concert hall or in a club, and also audio excerpts re-captured with any device (e.g., the audio track of a movie in a theater).

In the recent forensic literature, the detection of re-captured material (e.g., photographs of printed pictures) is a problem of actual interest discussed for both images [9] and videos [10][11]. However, when it comes to audio, to the best of our knowledge, no specific bootleg or re-capture detectors have been presented yet. Methods dealing with similar problems are: i) those addressing room estimation given a recorded track; ii) those aiming at estimating the model of the acquisition device used to capture the audio track. As far as the first category is concerned, in [12] the size of the room is estimated exploiting reverberation cues. In [13], a feature-based analysis is conducted to discriminate between specific kinds of room. As fa as the second category is concerned, in [14] authors proposed a method for device estimation using speech recordings. Whereas, in [15] the problem of microphone classification is studied from different perspectives and different microphone models. More recently, in [16], the

WIFS'2013, November 18-21, 2013, Guangzhou, China.
ISBN 978-1-4673-5593-3

authors have shown how to apply microphones classification methods using also portable devices (e.g., smartphones).

In this study we address the problem of audio bootleg detection from a different perspective that relies on the fact that humans have the ability to distinguish between audio tracks captured with portable devices and audio tracks that come from professional editing and mixing stages. This perceptual ability is based on acoustic cues (descriptors) we use to characterize sounds. Therefore, the contribution of this paper is twofold: i) we analyze music content in order to discover the best characterizing set of descriptors; ii) we develop a classification algorithm based on these features, in order to blindly detect audio bootlegs.

The rest of this paper is structured as follows. In Section II we describe the bootleg generation process and we explain which features are useful to characterize it. In Section III we introduce the detection algorithm explaining each step of it. In Section IV we present the results obtained on an audio dataset to validate our method. Finally, in Section V we draw our conclusions on this study.

II. Audio bootleg characterization

The ability of humans to describe and classify sounds and music has been subject of studies in many disciplines including psychology, sociology, acoustics, signal processing and music information retrieval. Although an exhaustive knowledge of the perceptual mechanisms involved in the human decision process is still not reached, many studies show how our attitude to discriminate and isolate sounds is highly related to many simple acoustic and structural cues [17][18].

In this paper we are interested in understanding which are the possible cues used to discriminate between bootleg and official professional produced music content. Indeed, in doing so, we can define a set of audio descriptors (i.e., features) for bootleg characterization that can be used to solve the detection problem. The features that we select come from those extensively used in the music information retrieval field and exhaustively explained in [17][19]. For this reason we are more focused on explaining why the feature that we select can be used as cues to solve the bootleg detection problem, rather than reporting a formal definition of all of them.

To this purpose, let us consider the process of bootleg acquisition. Typically portable devices are used for the recording, and the acquisition process chain is composed by several phases: Digital-to-Analog conversion performed by loudspeakers; sound propagation in a reverberant environments; possible addition of audience and ambient noise; Analog-to-Digital conversion performed by the recording engine. Each of these steps tends to alter the original content by introducing characteristic artifacts that we can exploit as fingerprints for bootleg detection. These effects can be broadly split into four categories:

- *Inharmonic Distortion:* alteration of the audio source typically due to Digital-to-Analog and Analog-to-Digital conversions.

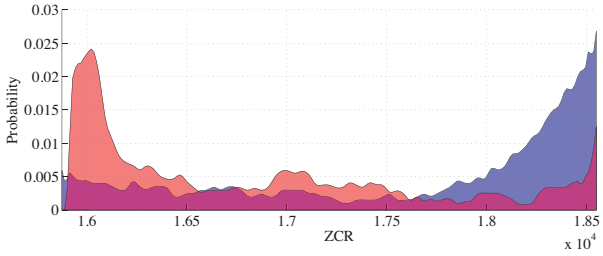


Fig. 1: ZCR values distribution for bootleg (blue) and official professionally produced live (red) when a high-frequency band is analyzed. Notice that this feature is highly discriminant.

- *Harmonic Distortion:* audio source rendered through a large set of loudspeakers in an uncontrolled environment can be affected by disturbs that tend to slightly alter the harmonic content of the audio source.
- *Loudness Saturation:* audio source rendered at high volumes and captured using non professional recording engines can affect the audio source by a compression of the waveform to the volume upper bound. This cause a saturation effect to the sound.
- *Background Noise:* environmental audio activities such as speech and scream alter the sound source adding noisy components.

In the following we show which features can be used to capture these traces.

To measure *Inharmonic Distortion* and *Background Noise*, features that can be used are Zero Crossing Rate (ZCR), Spectral Entropy, Flatness, and Spectral Irregularity. ZCR is defined as the normalized frequency at which the audio signal $x(n)$ crosses the zero axis. Fig. 1 depicts the distribution of the ZCR values computed on a set of songs belonging to the two classes of analysis, when high-frequency components are analyzed. The figure outlines the ZCR highly discriminant attitude for bootleg detection issues.

The most of the following features are computed through spectral analysis and in particular they derive from the magnitude spectrum. Magnitude spectrum is defined as the module of the Fast Fourier Transform (FFT). Particularly, given $x(n)$ a mono dimensional audio signal and $X(k)$ its FFT, the magnitude spectrum is formalized as

$$S(k) = |X(k)| = (Re(X(k))^2 + Im(X(k))^2)^{\frac{1}{2}}, \quad (1)$$

where $Re(X(k))$ is the real component, $Im(X(k))$ is the imaginary component of the FFT, and k is the frequency bin. Spectral analysis are performed through short term windowing techniques. Hence, magnitude spectrum and features are computed for each window (frame).

Spectral Entropy and Flatness features are measures of the similarity between the spectral magnitude of the signal and a flat spectrum (i.e., the spectrum of a white noise signal). Whereas, since noisy signals tend to exhibit a weak correlation in the spectrum of successive temporal frame of analysis, Spectral Irregularity feature is used to capture the variation

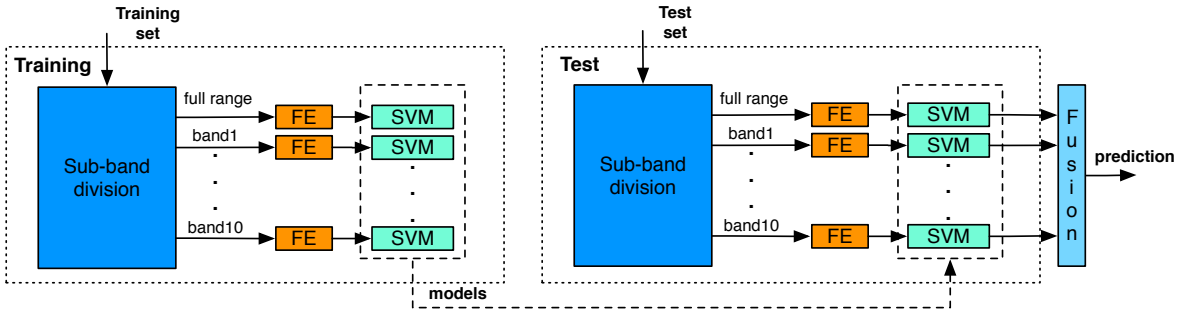


Fig. 2: Overall scheme of the proposed method. The methods is composed by a training phase, a test phase and a fusion phase. In both training and test phase Feature Extraction (**FE**) are computed and a battery of Support Vector Machine (**SVMs**) are used for classification.

of the successive peaks of the spectrum, and it is defined as

$$F_{IR} = \frac{\sum_{k=1}^{K} (S_l(k) + S_l(k+1))^2}{\sum_{k=1}^{K} S_l(k)^2}, \quad (2)$$

where $S_l(k)$ is the magnitude spectrum at the l-th frame and the k-th frequency bin.

In order to provide a characterization of *Harmonic Distortion*, we also consider Spectral Inharmonicity and Chromagram features. Spectral Inharmonicity measures the divergence of the partials with respect to the purely harmonic signal, and it is defined as

$$F_{SH} = \frac{2}{f_0} \frac{\sum_{h=1}^{H} |f_h - h f_0| \, (S_l(k_h))^2}{\sum_{h=1}^{H} (S_l(k_h))^2}, \quad (3)$$

where f_0 is the fundamental frequency, f_h is the estimated h-th partial, hf_0 is h-th harmonic of f_0 and k_h is the frequency bin of f_h.

The Chromagram is a representation of the spectrum in the logarithmic scale projected onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave.

Loudness Saturation is an inherent flaw in the accuracy of the recording medium. Non professional recordings performed in uncontrolled environments tend to produce frequent saturations. In order to capture this artifact, Root-Mean-Square (RMS) is considered.

Although the selected set of features can be very discriminant, as stated before, part of the human discriminant process is still not well understood. For this reason, since this process is mainly related to timbral characteristics, we include basilar timbral features (*Basic Descriptors*) to the set: Spectral Centroid, Spectral Spread, Spectral Kurtosis, Spectral Skewness, Spectral Flux, and Brightness. We also include MFCCs, which are proved to be very effective in many classification applications, since they are closely related to the human auditory system [13].

For the sake of clarity Table I reports the full list of features used in this study, grouped according to their characteristics.

TABLE I: List of acoustic and structural features used to characterize music content and relative issue to capture.

Characteristics	Features
Inharmonic Distortion	Spectral Irregularity, Spectral Entropy, Flatness, ZCR
Harmonic Distortion	Inharmonicity, Chromagram
Loudness Saturation	Spectral Irregularity, Spectral Entropy, Flatness, ZCR, RMS
Background Noise	Spectral Irregularity, Spectral Entropy, Flatness, ZCR
Basic Descriptors	Spectral Centroid, Spectral Spread, Spectral Kurtosis, Spectral Skewness, Spectral Flux, Spectral Rolloff, Brightness, MFCC

III. Methodology

Bootleg detection issue mainly consists in performing a discrimination between bootlegs ($\mathtt{B}$) and official professional productions ($\bar{\mathtt{B}}$). Hence, machine learning binary classification paradigms can be used for the purpose. In this study we adopt Support Vector Machines (SVMs) resulted to be very effective in several sound and music classification applications [20].

As described so far in this study, human attitude to music classification is mainly based on acoustic cues and it is performed through spectral analysis. Due to the intrinsic characteristic of the sound source to analyze, and due to the characteristic of our auditory system, some spectral frequency are more informative than others. To the purpose a multiband analysis is adopted.

The overall scheme of the method is depicted in Fig. 2. Each song is filtered through a Butterworth filter bank (Fig. 3) in order to split the spectrum in 10 bands as shown in Table II. For formalization purposes we refer to the full range case (i.e., the whole spectrum) as *band0*.

For each of the 11 bands (i.e., from *band0* to *band10*) a SVM model is trained independently. SVMs take as input a set of training pairs $\langle \mathbf{d}_i^b; y_i \rangle$, where $\mathbf{d}_i^b \in \mathbb{R}^D$ is the feature vector composed by D elements representing the b-th band of the i-th audio track, and $y_i \in \{\mathtt{B}; \bar{\mathtt{B}}\}$ is the class the track belongs to. During training, the surface in $\mathbb{R}^D$ that maximizes the margin between the two classes is sought. This surface serves in the test phase as a decision boundary between the

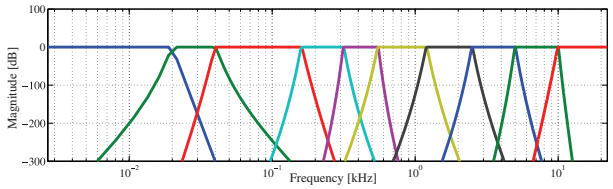


Fig. 3: Magnitude response of the filter bank.

TABLE II: Frequency boundaries used in multiband analysis as parameters for the Butterworth filters. The Table also shows the frequency boundaries for the full range analysis. In this case the whole spectrum is considered.

Frequency Boundaries [Hz]	ID
0-22000	full range (*band0*)
0-60	*band1*
60-230	*band2*
230-500	*band3*
500-1000	*band4*
1000-2000	*band5*
2000-4000	*band6*
4000-8000	*band7*
8000-12000	*band8*
12000-16000	*band9*
16000-22100	*band10*

two classes, splitting the D-dimensional space in two parts (see Fig. 4). Considering a single SVM (i.e., decision taken on a single band), an audio track whose $\mathbf{d}_i^b$ falls from one side of the surface is classified as belonging to one class, otherwise to the other. Notice that since some features are meaningful only for the full range (e.g., MFCCs), these are extracted only for *band0*. Moreover, since some features are not informative for all the bands, feature selection methods can be applied. To this end, in this study we used ReliefF feature selection algorithm that resulted to be very effective in music classification applications in the literature [21].

In order to take advantage from the multiband analysis, a prediction value is generated for each of the 11 bands, and a soft fusion technique is applied to combine them. More specifically, prediction values for each band are expressed as the signed distances δ_i^b between $\mathbf{d}_i^b$ and the separating surface. Notice that the sign of the distance, defined as $\text{sign}(\delta_i^b)$, reveals from which side of the surface the point lies (i.e., to which category it belongs to). Once δ_i^b are computed, they are aggregated as

$$\Delta_i = \sum_{b \in \mathcal{B}} \delta_i^b, \tag{4}$$

where $\mathcal{B}$ is a subset of bands. Indeed, in multiband approach to classification it is important to select only bands that maximize the information contribution. Adding non informative features can introduce some noise and downgrade the overall performances. For this reason only a subset $\mathcal{B}$ of the 11 bands that maximizes the overall accuracy in the training stage is considered in the fusion stage. The final decision $\mathtt{B}^*$ is then taken on the sign of Δ_i as

$$\mathtt{B}^* = \begin{cases} \mathtt{B} & \text{if} \quad \Delta_i > 0, \\ \bar{\mathtt{B}} & \text{if} \quad \Delta_i < 0. \end{cases} \tag{5}$$

Notice that the values δ_i^b can be summed without any normalization if features have been normalized beforehand.

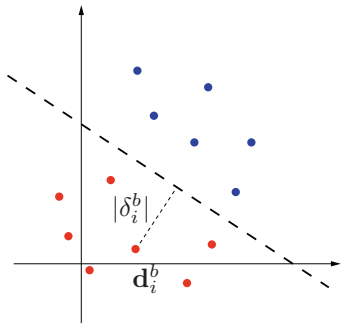


Fig. 4: Two-dimensional representation of a SVM for the b-th band. Point $\mathbf{d}_i^b$ is the representation of the b-th band of the i-th song. Its distance from the class separating surface is $|\delta_i^b|$.

IV. Experimental Results and Evaluations

With the intent to validate the developed detector, a dataset of songs has been conveniently collected. The dataset is composed by 260 audio 60s-long excerpts, all sampled at 44100 Hz. The excerpts have been extracted from 130 songs. Since classification techniques require a large dataset to be effective and since collecting a reliable dataset is a hard task, two excerpts have been extracted from each track. In order to maximize the differences between chunks of the same piece, one has been extracted at the beginning and one in the middle of the song. The dataset is balanced over the two classes: 130 excerpts are related to live concerts officially distributed in stores and 130 to bootlegs. Songs have been selected to cover a big set of music genres (e.g., rock, classic, jazz, etc.) and recording conditions (e.g., noisy concert halls, small clubs, etc.)[1]. Notice that we avoided the trivial case of comparing bootlegs to album versions of songs since tracks belonging to the latter category are easy to be distinguished from live performances.

Features have been computed using a Hanning windowing technique. As window size we used 1024 samples for all the features except for Spectral Irregularity, Inharmonicity, and Chromagram, for which we set the window size to 16384 samples. Windows were overlapped using 50% as hopsize. A compact representation for each feature have been obtained by averaging values for each song. The result is a feature vector $\mathbf{d}_i^b \in \mathbb{R}^D$ per song (i) per band (b), where D is the number of features. The features have been extracted from each track using the MIR toolbox [19]. The SVM classification was performed using the LIBSVM [22] implementation with radial kernel.

To estimate the robustness of the detector we performed a set of tests varying the cardinality of the training set. For each training set size N_{Tr}, we performed 10 different realizations randomly selecting N_{Tr} songs, and using the remaining N_{Te} songs as the test set. We tested the algorithm using the whole feature set. Fig. 5 shows the average accuracy of the system for each tested case. Error bars also show the standard deviation computed on the 10 different realizations. The evaluation

[1]The complete list of songs with audio examples is available at http://home.deib.polimi.it/bestagini/audio_bootleg/audio_bootleg.html

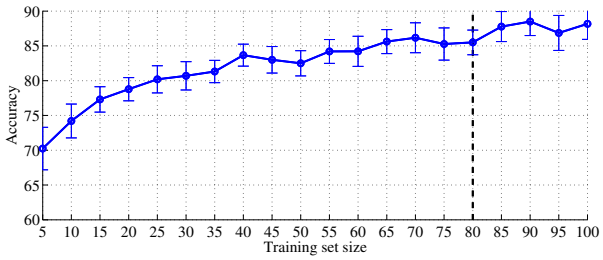


Fig. 5: Effect of the training set size on the accuracy. The black dashed line indicates the training set size (i.e., 80 excerpts) that we choose for the other experiments as a good compromise between size and accuracy. Error bars show the standard deviation computed on 10 different realization of training and test set.

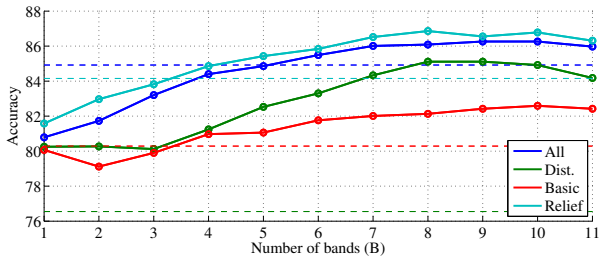


Fig. 6: Accuracy of the detector varying the number of bands selected by the fusion algorithm for each group of features (solid line). The accuracy is the average value computed over the 100 realizations. Dashed lines represent the accuracy value achieved in the full range case (i.e., using only *band0*).

resulted very promising. Indeed, the accuracy is prominent even for small cardinality. Moreover, using more than 35 excerpts for training, the accuracy is always higher than 80%. For the further evaluations, we decided to adopt 80 as the cardinality of the training set (as highlighted by the dashed black line in Fig. 5) as a good compromise between the set size and the achieved accuracy.

Since we proposed features capturing different acoustic characteristics, we are also interested in studying the discriminatory effect of different feature sets. For this reason, we tested the classification algorithm using either the whole set of features, or subsets of them. More specifically, we defined five groups:

- *All*: all the features;
- *Distortion*: RMS, ZCR, Flatness, Spectral Entropy, Spectral Irregularity, Inharmonicity, and Chromagram;
- *Basic*: Brightness, Spectral Rolloff, Spectral Flux, Spectral Centroid, Spectral Spread, Spectral Kurtosis, Spectral Skewness, MFCC;
- *MFCC*: MFCC only;
- *ReliefF*: features selected by the ReliefF algorithm.

These groups have also been used to test the effect of the fusion algorithm (except the *MFCC* group for which the band division is meaningless). In particular, for each group, we computed the accuracy for each band on the training set, and we ranked the bands according to this accuracy value. We then fixed the number $B \in [1; 11]$ of bands to consider. We applied the fusion method to the most significant B bands, according to the ranking obtained on the training set accuracy. Fig. 6 shows the effect on the accuracy as B increases for different groups of features. Results are reported as average value on 100 realizations of training and test set.

As shown in Table III and in Fig. 6, the band fusion method (solid lines) outperforms the full range case (*band0* - dashed lines). The results are mainly influenced by the fact that in the perceptual classification process most of the discriminant information lies in some specific frequency bands. Those bands are selected by the fusion method, by eclipsing those that can introduce noisy information. This effect can be seen in Fig. 6, where in some cases the add of the 11-th band downgrades the performances. From Fig. 6, we can also retrieve that when fusion is applied to a single band, results are different from the ones achieved in the full range case.

TABLE III: Comparative percentage values of accuracy, Precision, Recall, and F-measure when the detector makes use of different groups of features averaged over the 100 realizations. Results are shown for both the full range case (*band0*), and when fusion (*fusion*) is applied (considering the set of bands that maximize the accuracy).

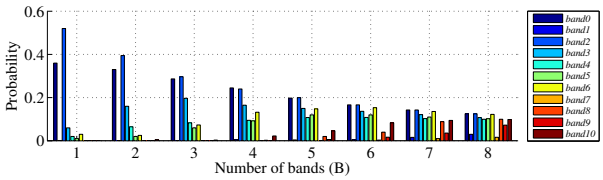
Features	Bands	Accuracy	Precision	Recall	F-measure
All	*band0*	84.92%	83.97%	86.32%	85.13%
	fusion	86.26%	85.38%	87.51%	86.43%
Distortion	*band0*	76.55%	75.83%	77.94%	76.87%
	fusion	85.11%	83.50%	87.51%	85.46%
Basic	*band0*	80.29%	77.34%	85.68%	81.30%
	fusion	82.59%	80.77%	85.54%	83.09%
MFCC	band0	79.15%	75.87%	85.50%	80.39%
ReliefF	*band0*	84.15%	82.89%	86.06%	84.45%
	fusion	**86.86%**	**85.70%**	**88.48%**	**87.07%**

This is due to the fact that the ranking of bands is suboptimal, since it has been performed in the training phase.

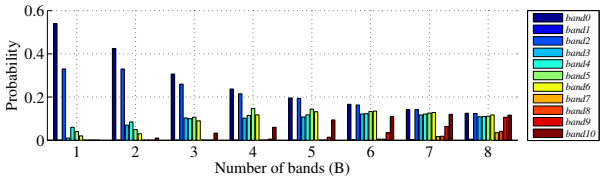
In order to analyze the contribute of each band, Fig. 7(a) shows the histograms of the bands selected over the 100 realizations. In figure, histograms are presented by varying the number of the first B ranked bands considered in the fusion process. The figure is relative to the case where *All* feature set is used. Notice that in general the most informative bands are *band0* and *band2*, whereas the less significant ones are *band7* and *band8*. Indeed, bootlegs are often recorded in the audience. As a consequence they suffer, with respect to the official versions, of the alteration derived by: i) the reverberant environment, which tends to alter the sound source by boosting low frequencies (as *band2*); ii) the emphasis to low frequencies provided by the sound engineers "punching the live".

Since less informative features can produce noise in the classification process, bands with both informative and non-informative features risk to be discarded by the fusion algorithm. This is the case of the *band10*. The use of ReliefF helps to solve the problem. The effect is shown in Fig. 7(b), where the application of the feature selection algorithm tends to gain the importance of the *band10*. As the results, the most informative bands become *band0*, *band2*, and *band10*. This is plausible, since the recording engine used in the bootleg acquisition chain is often a low-quality device that tends to apply a high attenuation (quality downgrade) on both low (*band2*) and high (*band10*) frequencies.

The best results (over 86% of accuracy) are obtained combining ReliefF and the band fusion method, since they are able to select the most discriminant bands and features for each band. Moreover, the analysis of most selected features



(a) *All*



(b) *ReliefF*

Fig. 7: Histograms of the most selected bands when the fusion method is applied as the number of bands B increases for both the *All* features case (a) and when ReliefF is applied (b).

confirms that the human ability to recognize bootlegs mainly relies on timbral cues. The most frequently selected features (more than the 80% of the times on the 100 realizations) are: RMS; ZCR; Spectral Entropy; Spectral Rolloff; Spectral Flux; Spectral Centroid; Spectral Skewness; Spectral Kurtosis. Among these, RMS and Spectral Flux are the two most discriminant features (i.e., selected the 98% of the times). Flux, in fact, captures the smoothing effect on the evolution of the spectrum over the time, due to the long reverberations typical of concert halls and clubs.

A further confirmation of the predominance of timbral descriptors can be proved considering the experiment with the *band0* (the only one in which MFCCs and Chromagram are extracted). In this case, indeed, MFCCs have been selected the 88% of the times, while chromagram only the 71%.

V. Conclusions

In this paper we presented a supervised algorithm to detect audio bootlegs. To develop such a detector, we first defined a set of audio features that are able to characterize the traces left during the bootleg acquisition process from the perspective of the human auditory system. Then, we trained a set of SVMs classifiers using these features extracted from different frequency bands for each audio track. In order to decide if an audio track is a bootleg, the outputs of the SVMs have been aggregated with a soft-assignment-based fusion technique. Results are promising, showing a detection accuracy of over 86%.

This algorithm can in principle be used also as a general audio re-capture detector. This make the developed tool interesting also for other scenarios. As an example, it can be jointly used with a video re-capture detector, to detect illegally re-captured movies. From this point of view, this detector may pave the way for multimodal audio-visual forensic analyses.

Acknowledgment

This work was supported by the REWIND Project funded by the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme (FP7) of the European Commission, under FET-Open grant number: 268478.

References

[1] H. Farid, "Exposing digital forgeries in scientific images," in *MM&Sec '06: Proceedings of the 8th workshop on Multimedia and security*, 2006.

[2] A. Piva, "An overview on image forensics," *ISRN Signal Processing*, vol. 2013, p. 22, 2013.

[3] S. Milani, M. Fontani, P. Bestagini, M. Barni, A. Piva, M. Tagliasacchi, and S. Tubaro, "An overview on video forensics," *APSIPA Transactions on Signal and Information Processing*, vol. 1, p. e2, 2012.

[4] S. Gupta, S. Cho, and C.-C. Kuo, "Current developments and future trends in audio authentication," *IEEE MultiMedia*, vol. 19, pp. 50–59, 2012.

[5] R. Olanrewaju and O. Khalifa, "Digital audio watermarking; techniques and applications," in *2012 International Conference on Computer and Communication Engineering (ICCCE)*, 2012.

[6] X.-M. Chena, M. Arnolda, P. Baum, and G. Doerr, "AC-3 bit stream watermarking," in *2012 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2012.

[7] S. Hicsonmez, H. Sencar, and I. Avcibas, "Audio codec identification through payload sampling," in *2011 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2011.

[8] C. Dittmar, K. Hildebrand, D. Gaertner, M. Winges, F. Muller, and P. Aichroth, "Audio forensics meets Music Information Retrieval - A toolbox for inspection of music plagiarism," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, 2012.

[9] H. Cao and A. Kot, "Identification of recaptured photographs on LCD screens," in *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010.

[10] M. Visentini-Scarzanella and P. L. Dragotti, "Video jitter analysis for automatic bootleg detection," in *2012 IEEE 14th International Workshop on Multimedia Signal Processing (MMSP)*, 2012.

[11] P. Bestagini, M. Visentini-Scarzanella, M. Tagliasacchi, P. Dragotti, and S. Tubaro, "Video recapture detection based on ghosting artifact analysis," in *2013 IEEE International Conference on Image Processing (ICIP)*, 2013.

[12] H. Malik and H. Farid, "Audio forensics from acoustic reverberation," in *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010.

[13] N. Peters, H. Lei, and G. Friedland, "Name that room: room identification using acoustic features in a recording," in *Proceedings of the 20th ACM international conference on Multimedia*, 2012.

[14] D. Garcia-Romero and C. Espy-Wilson, "Automatic acquisition device identification from speech recordings," in *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010.

[15] C. Krätzer, K. Qian, M. Schott, and J. Dittmann, "A context model for microphone forensics and its application in evaluations," in *SPIE Conference on Media Watermarking, Security, and Forensics*, 2011.

[16] L. Cuccovillo, S. Mann, M. Tagliasacchi, and P. Aichroth, "Audio tampering detection via microphone classification," in *2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP)*, 2013.

[17] H.-G. Kim, N. Moreau, and T. Sikora, *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. John Wiley & Sons, 2005.

[18] M. Zanoni, D. Ciminieri, A. Sarti, and S. Tubaro, "Searching for dominant high-level features for music information retrieval," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, 2012.

[19] O. Lartillot and P. Toiviainen, "MIR in matlab (ii): A toolbox for musical feature extraction from audio," in *2007 International Society for Music Information Retrieval conference (ISMIR)*, 2007.

[20] A. Lerch, *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*, 2012, Ed. Hoboken: Wiley-IEEE Press, 2012.

[21] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 448–457, 2008.

[22] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.