FORENSIC AND ANTI-FORENSIC ANALYSIS OF INDOOR/OUTDOOR CLASSIFIERS BASED ON ACOUSTIC CLUES

M. Mascia, A. Canclini, F. Antonacci, M. Tagliasacchi, A. Sarti, S. Tubaro

Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano, 20133 Milano, Italy

ABSTRACT

This paper addresses the problem of identifying the class of the environment where an audio recording was taken. We focus on distinguishing between indoor and outdoor speech recordings, and we propose a set of classifiers that provide a support for the forensic analyst in verifying the authenticity of audio content. The classifiers rely on acoustic clues extracted from the reverberant signal, namely the reverberation time (RT_{60}) and MFCC/LMSC feature vectors. We conducted several experiments, aimed at analyzing the algorithms from both the forensic and anti-forensic perspective. To do so, we devised a methodology for manipulating the signals in order to pretend that outdoor contents were recorded indoor, and vice-versa. Experimental results confirm the effectiveness of the proposed methods, which achieve high classification accuracy. The anti-forensics analysis reveals that attacks have moderate success rates, and severely depend from the classification algorithm adopted by the analyst.

Index Terms— audio forensics, anti-forensics, audio authentication, acoustic environment identification

1. INTRODUCTION

The increased availability of multimedia digital objects made them more and more adopted in courts of law, and other official venues, as evidence for establishing facts. For this reason, in the last few years, audio have received a great deal of attention from the multimedia forensics community [1]. Audio forensics includes authentication, aimed at verifying whether audio content is pristine; and audio tampering detection, which focuses on detecting and localizing malicious manipulations aimed at hiding or altering relevant information. Concerning tampering detection, many works rely on the extraction of the Electric Network Frequency (ENF) from the signal under analysis, whose phase discontinuities reveal potential attacks [2]. Other approaches aim at identifying the recording device, analyzing the footprints left by different microphones [3], possibly in conjunction with the ENF signal [4]. An overview of methods for audio authentication is reported in [5]. For instance, in [2] the authenticity of audio contents is verified by comparing the extracted ENF with a reference database. In [6], a system for audio bootleg (illegal versions of copyrighted objects) identification is proposed.

An important aspect of authentication is related to reverberation, which can be seen as the trace left by the environment on the acquired signal (called "roomprints" in [7]). The information carried by reverberation can be useful for certifying that a recording was made in a particular room and, moreover, to identify the room where the recording took place. Many works in the literature address these problems, with particular focus on speech recordings. In [8], a SVM (Support Vector Machine) is adopted to classify the environment, on the basis of MFCC (Mel-Frequency Cepstral Coefficients) and LMSC (Log Mel-Frequency Spectral Coefficients) acoustic features. Authors in [9] accomplish the identification task by using a GMM (Gaussian Mixture Model) classifier based on MFCCs. In [10] audio steganalysis features are used for jointly determining the microphone used and the environment. In [11], MFCCs are matched against those extracted from a reference database, in order to classify the environments according to their volume.

Differently from the aforementioned approaches, in this work we address the problem without relying on a database of reference room impulse responses (as done in [11]), neither limiting the identification to a set of candidate environments (as in [8, 9, 10]). Instead, we consider the problem of distinguishing speech signals recorded in indoor and outdoor locations. As noticed in [12], this task is helpful for verifying the coherence of the audio and video tracks in a recorded video file. The problem is addressed by exploiting two kinds of acoustic features: the reverberation time RT_{60} (i.e., the time taken for an impulsive sound to drop $60\,\mathrm{dB}$ below its initial intensity) and the MFCC/LMSC feature vectors. Outdoor locations are typically characterized by short RT_{60} values; conversely, indoor environments usually exhibit higher reverberation times, which implicitly provide a rough information about the volume of the enclosure. Unfortunately, the sole knowledge of the RT_{60} is not sufficient to identify the class of an acoustic environment. In fact, we can trust only on blind estimates of RT_{60}, which may be affected by errors. Moreover, some outdoor places may exhibit reverberation times higher or comparable to those of low-reverberant indoor locations. For these reasons, we also consider the MFCC/LMSC feature sets extracted from an estimate of the

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

reverberant component of the signal, which revealed to be robust for acoustic environment identification purposes [8].

In this paper we propose and compare a set of classification algorithms, which differently combine RT_{60} estimates and the MFCC/LMSC features. The goal of the work is twofold: on one hand we aim at assessing the effectiveness of the proposed algorithms from the forensic analyst perspective, thus investigating to what extent it is possible for an analyst to accurately discriminate between indoor and outdoor recordings. On the other hand, we are interested in considering an anti-forensic scenario, where the recordings are maliciously altered in order to deceive the analyst. To do so, we first design an anti-forensic methodology for manipulating the audio content; then we test the proposed classifiers over a set of manipulated audio objects. The experimental results confirmed the effectiveness of the classifiers in absence of an adversary. The anti-forensic analysis made possible to identify the algorithms that minimize the success rate of the attack, keeping it moderate.

2. THEORETICAL BACKGROUND

This section describes the theoretical background at the basis of the adopted classification algorithms. We first summarize the state-of-the-art algorithms for the blind estimation of the reverberation time. Then, we focus on the problem of obtaining the reverberant component from those signals, from which MFCC/LMSC features are extracted.

2.1. Blind estimation of the reverberation time

Signal model The algorithms described in this paragraph rely on the conventional signal model $x^{\mathrm{R}}(t) = x^{\mathrm{D}}(t) * h(t)$, where $x^{\mathrm{R}}(t)$ represents the reverberant speech signal; $x^{\mathrm{D}}(t)$ is the dry (anechoic) speech signal; and $h(t)$ is the impulse response of the environment. For $t \geq L_0$, L_0 being the time sample corresponding to the delay caused by the propagation from the speaker to the microphone, the response is modeled as an exponentially damped Gaussian white noise process [13]. Such model is given by the function $d(t) = a^t v(t) = e^{-t/\tau} v(t)$ where $v(t), t \geq 0$, is a sequence of i.i.d. random variables drawn from the normal distribution $\mathcal{N}(0,1)$. The constant τ describes the decay, and is proportional to the reverberation time through the relationship $\mathrm{RT}_{60} = (3 \cdot \ln 10)\tau$. The goal of the algorithms is, therefore, to estimate τ for inferring the reverberation time RT_{60} .

Algorithms Ratnam et al. [13] proposed to estimate the parameter a using a maximum likelihood approach. The speech signal $x^{\mathrm{R}}(t)$ is segmented into short frames and, for each frame, an estimate of a is obtained through a two-step iterative optimization. On the base of such estimates, an order-statistics filtering [13] is adopted for selecting the frames that include a free decay, and discarding all the others. The sequence of estimates of a relative to the retained frames is then processed with a median filter, followed by a temporal smoothing filtering. The last value of a in the sequence finally leads to the estimate of the RT_{60} .

Löllman et al. [14] proposed a more computationally efficient variant of Ratnam's method. In a preliminary stage, the reverberant signal $x^{\mathrm{R}}(t)$ is downsampled to operate a preselection of the frames candidate to contain free sound decays. Then, the algorithm follows the same steps as in [13] to obtain a sequence of estimates of the parameter a. The final RT_{60} value is obtained after a recursive smoothing of the sequence, aimed at reducing the variance of the estimation.

2.2. Blind estimation of the reverberant signal

In this paragraph we outline the algorithm proposed in [8] for extracting the reverberant component from a speech recording. The room impulse response $h(t)$ is segmented in $K+1$ blocks of length L, denoted as $\bar{h}_0(t), \dots \bar{h}_L(t)$. It is assumed that the first block contains only the direct path, i.e., $\bar{h}_0(t) = \delta(t - L_0)$, so that

$$x^{\mathrm{R}}(t) = x^{\mathrm{D}}(t - L_0) + r(t)\,, \quad r(t) = \sum_{n=1}^{K} x^{\mathrm{D}}(t - nL) * \bar{h}_n(t)\,,$$

where $r(t)$ indicates the reverberant signal to be estimated.

A short-time analysis of the input signal is performed, and we indicate with $X_m^{\mathrm{R}}(\omega)$ and $X_m^{\mathrm{D}}(\omega)$ the Short-Time Fourier Transforms (STFTs) of $x^{\mathrm{R}}(t)$ and $x^{\mathrm{D}}(t)$, respectively, where m is the temporal index. Moreover, we refer to $\bar{H}_n(\omega)$ as the Fourier transform of $\bar{h}_n(t)$. In a first stage, the algorithm provides a set of perceptual relevant estimates of the actual responses $\bar{H}_n(\omega)$, $n = 0, \dots K$ (see [8] for details), which are employed to compute an estimate $|\hat{X}_m^{\mathrm{D}}(\omega)|^2$ of the power spectrum of the dry signal. The magnitude of the reverberant signal is then estimated through spectral subtraction, i.e. $|\hat{R}_m(\omega)|^2 = |X_m^{\mathrm{R}}(\omega)|^2 - |\hat{X}_m^{\mathrm{D}}(\omega)|^2$. The phase component of $X_m^{\mathrm{R}}(\omega)$ is used for approximating that of $\hat{R}_m(\omega)$, and the reverberant signal $\hat{r}(t)$ is obtained through inverse STFT of $\hat{R}_m(\omega)$. As in [8], the signal $\hat{r}(t)$ is finally used for extracting the MFCC/LMSC features, which represent one of the inputs of the indoor/outdoor classifiers.

3. CLASSIFICATION SYSTEMS

In this section we detail the algorithms for indoor/outdoor classification. We start from those based on estimates of the RT_{60} . Then, we describe the method based on the MFCC/LMSC feature sets. Finally we consider two fusion techniques that differently combine the RT_{60} estimates and the extracted features.

3.1. Threshold-based classification

The simplest classification system can be derived by thresholding the estimated RT_{60} values, considering the fact that small reverberation times generally correspond to outdoor environments, while higher values are representative of indoor locations. Denoting with $\mathrm{RT}_{60}(\cdot)$ the function that estimates the reverberation time, the signal x^{R} is classified as *outdoor* if $\mathrm{RT}_{60}(x^{\mathrm{R}}) < \bar{T}$, and *indoor* otherwise. The discriminant threshold is selected as the value $\bar{T}$ that maximizes the accuracy of the classifier.

In this work we consider two threshold-based classifiers, derived from the two methods for the estimation of the RT_{60} discussed in Section 2.1.

3.2. Feature-based classification

In this paragraph we derive a classification system based on acoustic features extracted from the reverberant component $r(t)$, estimated as described in Section 2.2. Following the same approach as in [8], the 26 MFCC and 26 LMSC feature sets are computed from $r(t)$, and then concatenated to obtain a 52 elements global feature vector. To reduce the overfitting, we adopt the scalar feature selection technique described in [15], and the most effective features are then employed for building the classifier. As in [8], we consider a SVM with a radial basis kernel function to perform the classification.

3.3. Fusion-based algorithms

Fusion methods allow to integrate different kinds of knowledge or data in a new classifier, and are widely adopted to improve the effectiveness and the robustness of the classification. We consider fusion at feature and at measurement level.

Feature-level fusion With fusion at feature level, different features originally collected for different classifiers are put together and used to train a new classifier. A novel feature vector is obtained by concatenating the MFCC/LMSC features and the two RT_{60} estimates provided by the Ratnam' and Löllmann' methods (Section 2.1). The feature selection stage revealed the presence of the two RT_{60} estimates among the 5 most relevant features, thus confirming the discriminative power of such values. As before, the classification is accomplished by a SVM with radial basis kernel function, considering the most significant features.

Measurement-level fusion Fusion at measurement level is performed by mixing the outcome of various classifiers, to produce a final decision for the classification. In this paper we consider the plurality vote method described in [16]: given the outputs of M classifiers $(C_1(x), \ldots, C_M(x))$, the classification of the dataset entry x is obtained as $C(x) = \mathrm{mode}\ \ [C_1(x), \ldots C_M(x)]$. The resultant novel classifiers considers the outcomes of $M = 3$ systems: the two threshold-based classifiers (Section 3.1) and the feature-based method (Section 3.2).

4. DATASET DESCRIPTION

In this section we describe the dataset used for assessing the performance of the proposed indoor/outdoor classification systems. The dataset were built starting from 14 dry speech recordings[1] from different speakers: 6 males, 6 females, 2 children. The signals were recorded in anechoic conditions, and sampled at $48\,\mathrm{kHz}$. A total number of 196 reverberant speech signals were obtained through the convolution of each

[1]TSP Database, www-mmsp.ece.mcgill.ca/Documents/Data/

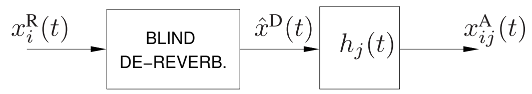


Fig. 1: General scheme for anti-forentics manipulations.

dry signal with 14 impulse responses relative to different acoustic environments. Each signal $x^{\mathrm{D}}(t)$ generates the set of reverberated signals $x_i^{\mathrm{R}}(t) = x^{\mathrm{D}}(t) * h_i(t)$, where $h_i(t)$ is the impulse response of the ith environment, $i = 1, \ldots 14$. We considered 7 indoor and 7 outdoor responses [2], coming from locations such as variously sized rooms, open spaces, churches, and parking lots. The considered impulse responses are provided along with the corresponding reverberation time RT_{60} , which ranges from $1.08\,\mathrm{s}$ to $7.65\,\mathrm{s}$ for the indoor case; and from $0.54\,\mathrm{s}$ to $1.58\,\mathrm{s}$ for the outdoor case. The resulting dataset contains 98 representatives for each of the two classes, for a total of 196 reverberant speech signals.

In order to test the robustness of the classification systems against anti-forensic attacks, we altered the dataset following the scheme in Fig. 1. In a first stage, the original dataset signals $x_i^{\mathrm{R}}(t)$ were dereverberated by subtracting the reverberant components $\hat{r}_i(t)$ estimated using the technique discussed in Section 2.2. More specifically, an estimate of the dry signal is given by $\hat{x}^{\mathrm{D}}(t) = x_i^{\mathrm{R}}(t) - \hat{r}_i(t)$. This operation represents the attempt of deleting the traces left by the original environment. After that, we need to simulate the effect of a new environment (different from the original one). To do so, we obtain the attacked signal as $x_{ij}^{\mathrm{A}}(t) = \hat{x}^{\mathrm{D}}(t) * h_j(t)$, where $h_j(t)$ $(j \neq i)$ is the impulse response of the new environment. We considered two different manipulations:

1) **outdoor-to-indoor attack:** the 98 signals of the outdoor class were dereverberated and then convolved with the 7 indoor impulse responses, obtaining a total number of 686 attacks. A further set of 686 attacked signals was obtained by skipping the dereverberation;

2) **indoor-to-outdoor attack:** the 98 signals of the indoor class were dereverberated and then convolved with the 7 outdoor responses, leading to 686 altered signals. We obtained a further set of 98 attacked signals by skipping the convolution.

5. EXPERIMENTAL RESULTS

In this section we assess the performance of the classification systems. We first consider the perspective of the forensic analyst without the presence of an adversary, testing the proposed classification systems on the 196 signals in the dataset; then, we analyze the anti-forensic scenario, assessing the performance of the classifiers on the dataset manipulated by an adversary. We will refer to the classifiers as: RAT (classifier based on the Ratnam's estimator); LOLL (classifier based on the Löllmann's estimator); FEAT (classifier based on the

[2]Open air library, www.openairlib.net; Noise collector, /www.freesound.org/people/NoiseCollector/packs/7917; Acoustic mirror impulses www.sonycreativesoftware.com/download/impulses

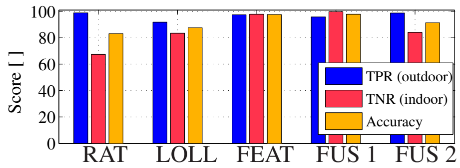
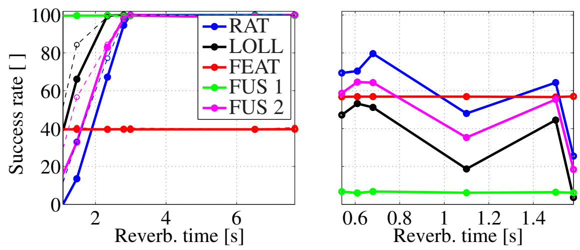


Fig. 2: Performance of the classifiers on non-attacked signals.



(a) outdoor-to-indoor

(b) indoor-to-outdoor

Fig. 3: Success rate of the attacks.

MFCC/LMSC feature vector); FUS 1 (classifier performing fusion at the feature level) and FUS 2 (classifier based on fusion at measurement level).

Forensic analysis The classification systems were evaluated on the $N = 196$ speech signals of the dataset, using a random sub-sampling validation. The dataset was randomly split into a sub-set of $N - K$ elements over which the classifier was trained; and a sub-set of K elements, used as validation data; we selected $K = 14$. The sub-sampling has been repeated for 1000 realizations. For each realization we computed the True Positive Rate (TPR) and the True Negative Rate (TNR), i.e. the percentage of correct classifications of the outdoor and indoor representatives, respectively. We also computed the accuracy, i.e., the percentage of correct classifications. The average results are shown in Fig. 2. All the systems exhibit high TPRs and TNRs, whose respective average values are 86.5% and 96.5%. Among the threshold-based methods (RAT and LOLL), the latter achieves the best accuracy (87.6%). The feature-based classifier (FEAT) outperforms the two former algorithms, approaching an accuracy of 97.6%. Fusion at feature level, (FUS 1), is the most effective strategy, leading to a further increase of the accuracy (97.8%). Fusion at measurement level (FUS 2) improves the accuracy of threshold-based classifiers, but is less effective than feature-level fusion.

Anti-forensic analysis We are now interested in evaluating the effect, on the proposed classifiers, of malicious attacks to the speech recordings. The testing was performed separately on the two manipulated datasets (outdoor-to-indoor and indoor-to-outdoor attacks) described in Section 4, following a leave-one-out approach. In particular, for each attacked signal to be classified, the classifiers were trained on the original (non-manipulated) dataset, from which the corresponding non-attacked signal had been removed.

The results of the outdoor-to-indoor attacks are reported in Fig. 3-(a), which shows the success rate (i.e., the percentage of wrong classifications) as a function of the RT_{60} of the impulse response used for altering the signals. As described in Section 4, we distinguish two versions of the same attack: one following the complete manipulation scheme (continuous lines in Fig. 3-(a)); the other skipping the dereverberation step (dashed lines). We first observe that, for most of the algorithms (RAT, LOL, and FUS 2), the success rate of the attack depends on the reverberation time. In these cases, the attack reveals to be successful only for $\mathrm{RT}_{60} \geq 3\,\mathrm{s}$. Below this threshold, the forensic analyst still have chances of correctly identifying the class of the signal, especially when the RAT and LOLL techniques are adopted. The FUS 1 and FEAT algorithms present a different behavior, being constant for all the considered RT_{60} values. However, they represent opposite situations: while for FUS 1 the attack is always successful ($\approx 100\%$), the FEAT method reveals to be the least attackable algorithm ($\approx 40\%$ success rate). Comparing the two approaches for accomplishing the attack, we generally observe a slight increase of the success rate when the dereverberation stage is bypassed. This can be explained since, in these cases, the attacked signals are characterized by the effect of two impulse responses (the original and the additional ones), leading to longer reverberant tails.

We now focus on the indoor-to-outdoor attack. As before, we consider the two versions of the manipulated dataset described in Section 4: one obtained by duly following the scheme in Fig. 1; and the other obtained bypassing the convolution operation. The results relative to the first manipulated dataset are reported in Fig. 3-(b). We observe that, as in the outdoor-to-indoor case, we can define two groups of classifiers: RAT, LOLL, and FUS 2, whose behaviors depend on the reverberation time; FEAT and FUS 1, which maintain a constant success rate. However, in this scenario FUS 1 represents the least attackable algorithm, with a very low success rate ($< 7\%$); while FEAT turns out to be more attackable, with a success rate of $\approx 59\%$. It is worth noticing that, in general, the indoor-to-outdoor attack is more challenging than the converse type of manipulation. This is not unexpected, since blind dereverberation represents a hard task, and it does not provide a complete removal of the reverberant tail. For this reason, the success rate tends to diminish as the RT_{60} of the new impulse response increases (particularly for RAT, LOLL, and FUS 2). Indeed, the residual indoor reverberation left by the non-ideal dereverberation is mixed with that of the new outdoor response, causing the attacked signals to be more reverberating than expected. Lastly, we analyze the effect of the convolution stage in performing the indoor-to-outdoor attack. When the convolution step is bypassed, we observe an average increase of the attack success rate of about 7%. In other words, bypassing the convolution facilitates the attacker, as it partially mitigates the non-ideality of dereverberation.

Finally, we consider an experiment aimed at resembling a real scenario, where the analyst has no a-priori informa-

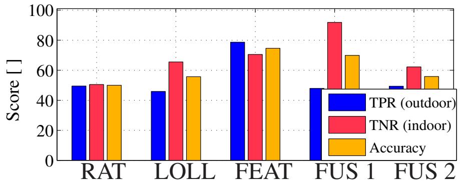


Fig. 4: Performance of the classifiers in presence of attacks.

tion about the type of the attack (if any) conducted on the signals. The classifiers were tested over a test set containing both unaltered and attacked signals, performing a random sub-sampling validation. The classifiers were trained over $N - K = 196 - 14$ unaltered signals. Then, they were tested over a mixed set, composed of the remaining K unaltered signals and K attacked ones, randomly chosen from the two manipulated datasets. The procedure has been repeated 1000 times. The results are summarized in Fig. 4. We observe that the FEAT classifier achieves the best accuracy (74.6%), followed by FUS 1 (69.9%). FEAT is well balanced between TPR and TNR; FUS 1 exhibits the highest TNR (91.8%), but a very low TPR (47.9%). It is worth noticing that this scenario is very conservative for the analyst perspective, since pristine and attacked signals are equally distributed in the dataset. In practice, we expect only a small percentage of audio content to be manipulated, since this can successfully accomplished only by experts who have signal processing expertise and a good knowledge of the classifiers used. However, even this difficult scenario guarantees high accuracy, making the FEAT and FUS 1 algorithms useful for authentication purposes.

Discussion In the light of the results reported in the previous paragraphs, we can draw some final considerations considering the two scenarios under analysis. From the analyst perspective, the best choice is the FEAT algorithm, which guarantees high accuracy and a moderate robustness against attacks. FUS 1, despite its high accuracy, exhibits a poor TPR ($< 50\%$). This is due to the fact that outdoor-to-indoor attacks are always successful in this case. Therefore, solely in case of some a-priori knowledge about the kind of attack (i.e., if only indoor-to-outdoor attacks are expected), FUS 1 allows the analyst to prevail over the adversary with very high probability. From the anti-forensic perspective, the adversary can assume that the FEAT algorithm is the most likely to be adopted by the analyst. In this case, the success rate is $\approx 40\%$ for the outdoor-to-indoor attack. For the indoor-to-outdoor case, the attacker has more chances ($\approx 59\%$) to cheat the analyst when the manipulation is performed considering both the dereverberation and convolution stages in Fig. 1.

6. CONCLUSIONS

We presented a forensic and anti-forensic analysis of several indoor/outdoor speech classifiers. The experimental evaluation has proved the effectiveness of the proposed systems in classifying unmodified signals. The anti-forensic analysis revealed that the success rate of the attack is dependent from the specific classifier adopted by the analyst. A detailed analysis of the results allowed us to identify the algorithm that jointly maximizes the accuracy and minimizes the success rate of both indoor-to-outdoor and outdoor-to-indoor attacks. The future developments will focus on investigating how the outputs of the classifiers could be fused to improve the overall robustness against malicious manipulations of data.

REFERENCES

[1] R. C. Maher, "Overview of audio forensics.," in *Intelligent Multimedia Analysis for Security Applications*. Springer, 2010.

[2] S. Mann, L. Cuccovillo, P. Aichroth, and C. Dittmar, "Combining enf phase discontinuity checking and temporal pattern matching for audio tampering detection.," in *GI-Jahrestagung*, 2013.

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

[4] L. Cuccovillo, S. Mann, P. Aichroth, M. Tagliasacchi, and C. Dittmar, "Blind microphone analysis and stable tone phase analysis for audio tampering detection," in *Audio Engineering Society Convention*, 2013.

[5] C. Grigoras, "Applications of enf criterion in forensic audio, video, computer and telecommunication analysis.," *Forensic Science Iternational Journal*, 2007.

[6] P. Bestagini, M. Zanoni, L. Albonico, A. Paganini, A. Sarti, and S. Tubaro, "Feature-based classification for audio bootlegs detection," in *IEEE International Workshop on Information Forensics and Security, WIFS*, 2013.

[7] A. H. Moore, M. Brookes, and P. A. Naylor, "Roomprints for forensic audio applications," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA*, 2013.

[8] H. Malik and H. Zhao, "Recording environment identification using acoustic reverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2012.

[9] N. Peters, H. Lei, and G. Friedland, "Name that room: Room identification using acoustic features in a recording," in *ACM Multimedia Conference*, 2012.

[10] C. Kraetzer, A. Oermann, J. Dittmann, and A. Lang, "Digital audio forensics: A first practical evaluation on microphone and environment classification," in *Workshop on Multimedia & Security*, 2007.

[11] N. Shabtai, B. Rafaely, and Y. Zigel, "Room volume classification from reverberant speech," in *International Workshop on Acoustic Echo and Noise Control, IWAENC*, 2010.

[12] S. Milani, P. Piazza, P. Bestagini, and S. Tubaro, "Audio tampering detection using multimodal features," in *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, 2014.

[13] R. Ratnam, D. L. Jones, B. C. Wheeler, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *Journal of the Acoustical Society of America*, 2003.

[14] H. W. Löllman, E. Yilmax, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *International Workshop on Acoustic Echo and Noise Control, IWAENC*, 2008.

[15] S. Theodoridis and K. Koutroumbas, *Pattern Recognition, Fourth Edition*, Academic Press, 2008.

[16] F. Moreno-Seco, J. M. I. Quereda, P. J. P. de Len, and L. Mic, "Comparison of classifier fusion methods for classification in pattern recognition tasks," in *Structural, Syntactic, and Statistical Pattern Recognition*. Springer, 2006.