Automatic Recording Environment Identification Using Acoustic Features

Usman Amin Chaudhary¹, Hafiz Malik¹
¹University of Michigan-Dearborn, Dearborn, MI 48128, USA

Correspondence should be addressed to Usman A Chaudhary, Hafiz Malik ([uchaudha,hafiz]@umd.umich.edu)

ABSTRACT
Recording environment leaves its acoustic signature in the audio recording captured in it. For example, the persistence of sound, due to multiple reflections from various surfaces in a room, causes temporal and spectral smearing of the recorded sound. This distortion is referred to as audio reverberation time. The amount of reverberation depends on the geometry and composition of a recording location, the difference in the estimated acoustic signature can be used for recording environment identification. We describe a statistical framework based on maximum likelihood estimation to estimate acoustic signature from the audio recording and use it for automatic recording environment identification. To achieve these objectives, digital audio recording is analyzed first to estimate acoustic signature (in the form of reverberation time and variance of the background noise), and competitive neural network based clustering is then applied to the estimated acoustic signature for automatic recording location identification. We have also analyzed the impact of source-sensor directivity, microphone type, and learning rate of clustering algorithm on the identification accuracy of the proposed method.

1. INTRODUCTION
In this digital age, technologies allow digital media to be produced, altered, manipulated, and shared in ways that were beyond the imagination a few years ago. This fact poses serious challenges to forensic science. Today, whether it be a viral video of “pop corn with cell phone” posted on youtube [1] or Iranian missile test images appeared in cover stories [2], as a consequence, we can no longer take the authenticity of media objects for granted; digital technologies are the major contributing factor behind this paradigm shift. As digital technologies continue to
evolve it will become increasingly more important for the science of digital forensics to keep pace.

The past few years have witnessed significant advances in image forensics [3–11], on the other hand, techniques for audio forensics are less developed. Notable exceptions can be divided into three major categories:

1. The electrical network frequency (ENF) based framework which verifies integrity by comparing the extracted ENF with the reference ENF database [15, 16, 27]. These methods are effective against cut-and-paste (CAP) attacks, but complex electro-physical requirements of ENF-based approaches [15] make them ineffective for recordings captured using battery-powered devices.

2. Statistical pattern recognition based techniques [12,13,17–19,22–25] have been proposed for recording location and device identification. However, these methods are limited by their low accuracy and inability to uniquely map an audio recording to the source. We have also been addressing these limitations by modeling microphone induced nonlinearities in the recording and use them to map recording to the microphone used [31], and developing model driven approaches for automatic recording environment identification and forgery detection [20,32].

3. Techniques based on time domain analysis to determine authenticity of MP3 audio files and capture traces of double compression [35–37] by analyzing encoder frame offsets in time domain.

Major research areas in the field of audio forensics are considered to be speech recognition and speaker verification, localization and identification. Audio forensics focuses not only the direct speaker verification but also the recording environment identification [25]. This paper proposed a new approach which describes a detailed analysis of audio data to provide an evidence of the characteristics of the place where the recording was captured. Major contribution of this paper is to develop a statistical framework for automatic recording environment identification. Here we exploit specific artifacts introduced at the time of recording to authenticate an audio recording and recording environment identification, that is, indoor or outdoor. Audio reverberation is caused by the persistence of sound after the source has terminated. This persistence is due to the multiple reflections from various surfaces in a room. As such, differences in a room’s geometry and composition will lead to different amounts of reverberation time. There is a significant literature on modeling and estimating audio reverberation (see, for example, [33]). We describe how to model and estimate audio reverberation – this approach is a variant of that described in [28]. We have shown that reverberation can be reliably estimated and show its efficacy in simulated and recorded speech. In order to achieve automatic recording environment identification, competitive neural network (CNN) based classification is used. We have also analyzed the impact of source-sensor directivity, microphone type, and learning rate of clustering algorithm on the identification accuracy.

The rest of the paper is organized as follows: details of reverberation modeling and estimation are outlined in Section 2; a brief overview of the competitive neural network based clustering is provided in Section 2.3. Simulation results and performance analysis are discussed in Section 3. Finally the concluding remarks along with future research directions are presented in Section 4.

2. METHOD

2.1. Model of Sound Decay

The decay of an audio signal \( x(t) \) is modeled with a multiplicative decay and additive noise (Fig. 1):

\[
y(t) = d(t)x(t) + n(t),
\]

(1)

where,

\[
d(t) = \exp(-t/\tau).
\]

The decay parameter \( \tau \) embodies the extent of the reverberation, and can be estimated using a maximum likelihood estimator.

2.2. Maximum Likelihood Estimation

We assume that the signal \( x(t) \) is a sequence of \( N \) independently and identically-distributed (i.i.d) zero
The log-likelihood function, \( \ln(\mathcal{L}(\cdot)) \). This is achieved by setting the partial derivatives of \( \mathcal{L}(\cdot) \) equal to zero and solving for the desired \( \tau \).

\[
\frac{\partial \mathcal{L}}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^2} \sum_{k=0}^{N-1} \frac{y^2(k)}{\gamma^2(k)} \\
\frac{\partial \mathcal{L}}{\partial \tau} = -\sum_{k=0}^{N-1} k \tau^{-2k-1} \left( \frac{y^2(k)}{\sigma^2 \gamma^2(k)} - 1 \right) .
\]

For the purpose of numerical stability, the maximization is performed on \( \tilde{\tau} = \exp(-1/\tau) \). Although \( \sigma \) in Equation (7) can be solved for analytically, \( \tilde{\tau} \) in Equation (8) cannot. As such, an iterative nonlinear minimization is required. This minimization consists of two primary steps, one to estimate \( \sigma \) and one to estimate \( \tilde{\tau} \). In the first step \( \sigma \) is estimated by setting the partial derivative in Equation (7) equal to zero and solving for \( \sigma \), to yield:

\[
\sigma^2 = \frac{1}{N} \sum_{k=0}^{N-1} \frac{y^2(k)}{\gamma^2(k)} = \frac{1}{N} \sum_{k=0}^{N-1} \frac{y^2(k)}{\tau^2 + \sigma_n^2} .
\]

This solution requires an estimate of \( \sigma \), which is estimated from the noise floor following the decayed signal. This solution also requires an estimate of \( \tilde{\tau} \) which is initially estimated using Schroeder’s integration method [34]. In the second step, \( \tilde{\tau} \) is estimated by maximizing the log-likelihood function \( \mathcal{L}(\cdot) \) in Equation (6). This is performed using a standard gradient descent optimization, where the derivative of the objective function is given by Equation (8). These two steps are iteratively executed until the differences between consecutive estimates of \( \sigma \) and \( \tilde{\tau} \) are less than a specified threshold. In practice, this optimization is quite efficient, converging after only a few iterations.

2.3. Clustering

Data clustering is a complex optimization problem with applications ranging from vision and speech processing to data transmission and data storage in technical as well as in biological systems. Different clustering methods like visual clustering, K-means clustering, fuzzy clustering, entropy constrained vector quantization, or topological feature maps and competitive neural networks can be used to group estimated acoustic reverberation parameters, i.e., \( \hat{\tau} \) and \( \hat{n} \).
and $\sigma$, based on some distance measure. Visual clustering is probably the most efficient clustering methods to group estimated $\tau$ and $\sigma$ in low dimensional feature space, however it can become extremely cumbersome when facing large quantity of data points in the N-dimensional (where $N \geq 4$) feature space. In order to achieve, automatic recording environment identification, competitive neural network (CNN) based clustering is considered. The CNN-based clustering strategy provides explicit tradeoff between simplicity and precision of a data representation by jointly optimizing distortion errors and complexity costs. A maximum entropy estimation of the clustering cost function yields an optimal number of clusters, their positions, and their cluster probabilities. The CNN-based clustering was implemented using COMPACT 2.0 tool [38], a freeware downloaded from 1. Clustering performance is investigated various learning rates and number of iterations.

3. EXPERIMENTAL RESULTS

To test the effectiveness of the proposed framework, we analyzed synthetic data and real-world speech recordings captured in diverse set of recording environments like small offices, large offices, hallway, stairwell, restroom, atrium, staircase, and outdoor environments.

3.1. Acoustic Parameter Estimation using Synthetic Data

In the first experimental setting direct signal $x(t)$ is generated using our iid zero mean and normally distributed assumption with $N = 1024$ and with an assumed sampling rate of 512 samples/seconds. Shown in the top panel of Fig. 1 is a signal $x(t)$. Shown in the central panel is the exponential decay $d(t) = \exp(-t/\tau)$ with $\tau = 0.29$ seconds, and shown in the bottom panel is the resulting decayed signal $y(t)$ with additive noise as specified by Equation (1).

We generated 1000 random signals according to this model with values of $\tau \in [0.29, 0.88]$ seconds, and either with no noise ($\sigma_n = 0$), or with a $\sigma_n$ to yield an average signal-to-noise ratio of 26dB. As described in the previous section, the decay parameter $\tau$ was estimated from these signals. Shown in the left panel of Fig. 2 are the actual values of $\tau$ as a function of the estimated values ($\tau_{est}$) for the no noise case.

The average estimation error is 0.01 seconds with a standard deviation of 0.01. Shown in the right panel of Fig. 2 are the estimation results for the additive noise case. The average estimation error is 0.04 seconds with a standard deviation of 0.03. The handful of outliers have small values of $\tau$ (i.e., rapid decay) which leads to a signal where the noise dominates, thus leading to occasionally unreliable estimates.

3.2. Acoustic Parameter Estimation using Speech Recordings

In our second experiment, we recorded human speech in four different environments: (1) outdoors; (2) small office ($7' \times 11' \times 9'$); (3) large office ($15' \times 11' \times 9'$); and (4) stairwell. In each case, the same speaker read the opening paragraph of Charles Dickens’ Tale of Two Cities. The audio was recorded using a commercial-grade microphone. As described above, the reverberation was estimated from fourteen positions in each of the recorded audio segments. These were manually selected on the basis that the speech at these positions decayed to the noise floor, Fig. 3. Because there was considerable background noise in these recordings, each recording was initially pre-processed with a speech enhancement filter [30]. The mean (and standard deviation) estimate for the decay parameter $\tau$, in seconds, is: (1) outdoors: 0.049 (0.013); (2) small office: 0.062 (0.017); (3) large office: 0.083 (0.012); and (4) stairwell 0.203 (0.064). This difference is significant as confirmed by a one-way ANOVA ($F(3, 40) = 39.93$, $p \leq 0.000001$). Although individual estimates of $\tau$ are not sufficiently reliable to fully characterize a speaker’s environment, the running averages over even a short length of audio shows significant differences in the estimated decay parameter.

3.3. Automatic Recording Environment Identification

In our third experiment we recorded 108 samples of a hand-clap recording using 3 microphones: a commercial grade external microphone, a built-in microphone in Apple’s MacBook, and a built-in HP Compaq Laptop. These recordings were captured at 9 locations: three small offices, a large office, a hallway, a restroom, an atrium, a staircase, and an outdoor environment. The hand-clap recording (downloaded from 2) was played using a pair

1http://adios.tau.ac.il/compact/

2http://www.freesound.org/samplesViewSingle.php?id=345
Fig. 2: Estimation results for synthetically generated signals with no noise (left) and with 26dB of additive noise (right).

of commercial grade external speakers. In each environment four samples were recorded through each microphone. These recording were captured with mono audio channel and sampling rate of 16000 samples per second. Each recording was initially preprocessed with a speech enhancement filter [30]. For acoustic feature estimation, 15 decaying tails were manually selected from each clean recording. From the identified tails acoustic signature, i.e., $\tau$ and $\sigma_n$ was estimated using method discussed in sec:MLE.

The estimated acoustic signature from all recordings were first classified using visual clustering. Shown in Fig. 4 is the scatter plot of the estimated acoustic signatures from all nine recording environments recorded using external microphone. It can be observed that estimates from recordings captured in atrium and hallway cannot be classified even using visual clustering. These recording environments therefore were removed from further analysis. Shown in Fig. 5 is the classification of visual clustering of recordings captured using external microphone in the selected seven environments. It can be observed from Fig. 5 the outdoor environments, large office, small offices, staircase, and restrooms are clearly classified.

Although visual clustering did a decent job in grouping acoustic signatures based on density based similarity measure, however to achieve automatic recording environment identification competitive neural network based clustering was applied to the estimated acoustic signature from all recordings. Shown in Fig. 6 is the output of CNN clustering tool for recordings captured in seven environments using external microphone. These clustering results were obtained for 150 iterations and a learning rate of 0.5. It can be observed that the CNN based clustering yields five clusters. Actually, CNN clustering merged three acoustically similar environments, that is, identical small rooms of same size but different settings into a single cluster (cluster # 3 in Fig. 6). Third cluster, representing three small offices, was analyzed further using CNN-based clustering again yielded three distinct clusters as shown in Fig. 7.

We also investigated recording environment identification accuracy as a function of number of iterations and microphone type. Different number of iterations were tried to find the optimal clustering. Shown in Table 1 is the clustering performance as a function
**Fig. 4:** Visual clustering of all recording environments: three small offices, a large office, a hallway, a restroom, an atrium, a staircase, and an outdoor environment.

**Fig. 5:** Visual clustering of selected seven recording locations: three small offices, a large office, a restroom, an atrium, and an outdoor environment.

It can be observed from Table 1 that when the sample space consisted of seven selected recording environments the CNN clustering yielded three clusters for 100 iterations, five clusters for 150 iterations and 3 clusters for 200 iterations.

<table>
<thead>
<tr>
<th>Number of Iterations</th>
<th>Actual Number of Environments</th>
<th>Correctly Identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>150</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>200</td>
<td>7</td>
<td>3</td>
</tr>
</tbody>
</table>

**Table 1:** Clustering performance for *external microphone* as a function of number of iterations with learning rate 0.5.

of number of iterations for the recording captured in seven selected locations (three small offices, a large office, a restroom, an atrium, and an outdoor environment) using external microphone. It can be observed from Table 1 that when the sample space consisted of seven selected recording environments the CNN clustering yielded three clusters for 100 iterations, five clusters for 150 iterations and 3 clusters for 200 iterations.

<table>
<thead>
<tr>
<th>Number of Iterations</th>
<th>Actual Number of Environments</th>
<th>Correctly Identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>150</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>200</td>
<td>7</td>
<td>3</td>
</tr>
</tbody>
</table>

**Table 2:** Clustering performance for *HP Laptop built-in microphone* as a function of number of iterations with learning rate 0.5.
Table 3: Clustering performance for *MacBook built-in microphone* as a function of number of iterations with learning rate 0.5.

<table>
<thead>
<tr>
<th>Number of Iterations</th>
<th>Actual Number of Environments</th>
<th>Correctly Identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>150</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>200</td>
<td>7</td>
<td>5</td>
</tr>
</tbody>
</table>

It can be observed from Table 2 that clustering accuracy for recordings captured using for HP Laptop built-in microphone deteriorates as number of iterations increases, whereas, in case of MacBook clustering accuracy seems independent of number of iterations.

These simulation results shows that that clustering accuracy depends on two very important factors: 1) sensitivity of the microphone, based on the data collected using three microphones (two built-in and one external), the results for external microphone was much better then the built-in microphones; and 2) directivity of the microphone, it has been observed that directed towards the sound-source the results tend to be better.

4. DISCUSSION

This proposed a statistical framework for automatic recording environment identification using acoustic signature from audio recording. Acoustic reverberation is used for acoustic signature estimation and for automatic classification competitive clustering is considered. Simulation results presented in this paper indicates that the proposed framework is efficient for most of the recording environments. The propose methods fails to uniquely identify dynamic recording environments such as hallway and atrium.

We are currently investigating performance of the proposed system for speaker independent as well as speech independent recording environment identification. We are also focusing on developing framework for automatic decaying trail selection from audio recording. We also plan to develop a component based framework for audio forensic analysis which involves characterizing contributing sources (e.g. acoustic reverberation, background noise, mi-
Fig. 7: Competitive neural network based clustering of cluster number 3 representing small rooms

crophone noise, and transcoding artifacts) and estimating mixing processes that typifies the formation of digital audio recording.

5. REFERENCES


digital forgeries in color filter array interpolated
images,” IEEE Transactions on Signal Process-

“Digital Image Forensics via Intrinsic Finger-
prints,” IEEE Trans. on Information Forensics

The New Science of Forensic Phonetics,”
Plenum Publishing Corporation, ISBN:
0306434679.


[14] G. Defrance, L. Daudet, and J.-D Polack,
(2008), “Characterizing sound sources from
room-acoustical measurements,” International
Symposium on Room Acoustics.

[15] C. Grigoras, A. Cooper, and M. Michalek,
(2009), “Forensic Speech and Audio Analysis
Working Group - Best Practice Guidelines for
ENF Analysis in Forensic Authentication of
Digital Evidence,” European Network of Foren-
sic Science Institutes.

terion in forensic audio, video, computer and
telecommunication analysis,” Forensic Science

[17] D. Garcia-Romero, and C. Espy-Wilson,
(2010), “Speech forensics: Automatic acquisi-

[18] D. Garcia-Romero, and C. Espy-Wilson,
(2010), “Automatic acquisition device identifi-
Processing (ICASSP 2010), Dallas, TX.

[19] D. Garcia-Romero, and C. Espy-Wilson,
(2009), “Automatic acquisition device identifi-

Forensics using Background Noise,” IEEE Int.
Conf. on Multimedia and Expo 2010 (ICME10).

provement of information fusion-based audio
steganalysis,” Proc. Multimedia on Mobile
Devices 2010, Electronic Imaging Conference

[22] R. Buchholz, C. Kraetzer, and J. Dittmann,
(2010), “Microphone Classification Using
Fourier Coefficients,” Lecture Notes in Com-

[23] C. Kraetzer, M. Schott, and J. Dittmann,
(2009), “Unweighted Fusion in Microphone
Forensics using a Decision Tree and Linear Lo-
gistic Regression Models,” Proc. ACM Multi-

practical evaluation on microphone and envi-
ronment classification,” Proceedings of the 9th
workshop on Multimedia and Security.

“Verifier-tuple for audio-forensic to determine
speaker environment,” Proc. ACM Multi-

tion of Gunshots,” Signal Processing Applica-
tions for Public Security and Forensics.

[27] D.P. Nicolalde and J.A. Apolinario, (2009),
“Evaluating digital audio authenticity with
spectral distances and ENF phase change,”
IEEE International Conference on Acoustics,
Speech and Signal Processing, (ICASSP 2009).

tion of the Reverberation Time in Noisy Envi-
ronments,” Proceedings of International Work-
shop on Acoustic Echo and Noise Control.

“Prediction of energy decay in room im-
pulse responses simulated with an image-source
model,” Journal of Acoustic Society of Amer-


Fig. 3: Shown in the top panel is an audio signal recorded in a large office. The reverberation time was estimated from fourteen positions (shaded areas), each manually selected such that the speech decayed to the noise floor. Shown below are three sample segments revealing the form of the audio decay due to reverberation.