RECORDING ENVIRONMENT IDENTIFICATION USING ACOUSTIC REVERBERATION

Hafiz Malik and Hong Zhao

ECE Department, University of Michigan - Dearborn, Dearborn, MI 48128

ABSTRACT

Acoustic environment leaves its fingerprints in the audio recording captured in it. Acoustic reverberation and background noise are generally used to characterize an acoustic environment. Acoustic reverberation depends on the shape and the composition of a room, therefore, differences in the estimated reverberation can be used in a forensic and ballistic settings and acoustic environment identification (AEI). We describe a framework that uses acoustic reverberation to characterize recording environment and use it for AEI. Inverse filtering is used to estimate the reverberation component from audio recording. A 48-dimensional feature vector consisting of Mel-frequency Cepstral Coefficients and Logarithmic Mel-spectral Coefficients is used to capture traces of reverberation. A multi-class support vector machine (SVM) classifier is used for AEI. Experimental results show that the proposed system can successfully identify a recording environment for regular as well as blind AEI.

Index Terms— Audio Forensics

1. INTRODUCTION

Today, technologies allow digital media to be produced, altered, manipulated, and shared in ways that were beyond imagination few years ago. This technological age is affecting nearly every corner of our lives ranging from courts to media, politics, business, and science. Today, whether it be a viral video of “pop corn with cell phone” posted on youtube or images of “Iranian missile test” appeared in the news media, we no longer jump to a conclusion instantly. 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 existing state-of-the-art in digital audio forensics [1, 2, 3, 4] is capable of detecting a limited set of alterations. For example, electrical network frequency (ENF) based methods [1] are effective against cut-and-paste attacks. Statistical pattern recognition based techniques [2, 4] can be used for recording location and acquisition device identification. Model driven approaches have also been proposed [5, 6, 7] to estimate acoustic reverberation parameters estimation using maximum likelihood framework and use them for automatic acoustic environment identification (AEI).

Here we exploit specific artifacts introduced at the time of recording to authenticate an audio recording using acoustic reverberation. 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. There is a significant literature on modeling and estimating audio reverberation (see, for example, [8]) and blind de-reverberation (see, for example, [9]).

We describe how to model and estimate reverberant signal from audio recording using inverse filtering based on spectral subtraction – this approach is a variant of that described in [9]. Estimated reverberant signal is used for feature extraction. A 48-D feature vector consisting of Mel-frequency Cepstral Coefficients (MFCC) and Logarithmic Mel-spectral Coefficients (LMSC) is used to capture traces of reverberation. A multi-class support vector machine (SVM) is used for acoustic environment identification. We have shown that reverberant signal component can be reliably estimated using blind de-reverberation and used for the AEI. A low-dimensional feature space (e.g., 48-dimensional feature space) and high classification accuracy (e.g. > 90%) are the salient features of the proposed system.

1.1. Problem Modeling

Consider digital audio recording $y(t)$, which is a combination of several components, i.e., direct speech signal $s(t)$, acoustic environment distortion (consisting of reverberant signal $r(t)$ and background noise $η(t)$), microphone noise $η_{Mic}(t)$, encoding distortion $η_{E}(t)$, and transcoding artifacts $η_{Transc}(t)$. The direct signal at the microphone input can be expressed as $s(t) = h_{VT}(t) * u(t)$, where $h_{VT}$ represents vocal track impulse response, $u(t)$ represents excitation source.
passing through speech production model, and \( * \) denotes convolution operator. Shown in Fig. 1 is a simplified model for digital audio recording.

The combined effect of direct, reflected, and background noise at the input of the microphone can be expressed as \( y(t) = s(t) + h_{RIR}(t) \ast s(t) + \eta(t) \), where \( h_{RIR}(t) \) denotes room impulse response (RIR). If \( h_{Mic}(t) \) denotes microphone impulse response, then \( y(t) \) can be expressed as \( y(t) = h_{final}(t) \ast u(t) + h_{Mic}(t) \ast \eta(t) + \eta_{Mic}(t) + \eta_{TC}(t) \). where \( h_{final}(t) \) can be expressed as, \( h_{final}(t) = h_{VT}(t) \ast h_{RIR}(t) \ast h_{Mic}(t) \).

The proposed framework estimates acoustic reverbation signal from audio recording, \( y(t) \), and use it for AEI. What follows is an overview of how to model and estimate reverbation component and its application for AEI.

2. METHOD

The proposed system can be divided into two subsystems: 1) blind de-reverberation subsystem and 2) feature extraction and classification subsystem. Details of each subsystem are provided next.

2.1. Blind De-reverberation Subsystem

This subsystem estimates reverbation signal, \( h_{RIR}(t) \ast h_{Mic}(t) \ast s(t) \), from the audio recording, \( y(t) \) which embodies acoustic environment characterization. It is therefore reasonable to focus on \( h_{RIR}(t) \). The proposed audio recording model indicates that it is hard to estimate \( h_{RIR} \) directly from audio recording, as \( y(t) = s(t) + h_{RIR}(t) \ast h_{Mic}(t) \ast s(t) + h_{Mic}(t) \ast \eta(t) + \eta_{Mic}(t) + \eta_{TC}(t) \). To get rid of microphone impulse response and microphone and transcoding distortion, flat microphone impulse response and negligible distortions are assumed. This is a reasonable assumption, as we have shown in the Section 3 that identification performance is independent of the microphone used. The \( y(t) \) now can be expressed as \( y(t) = s(t) + r(t) = s(t) + h_{RIR}(t) \ast s(t) \), where \( s(t) \) represents direct sound component, also referred as dry signal and \( r(t) \) represents the reverbenter component.

The \( h_{RIR}(t) \) can be modeled by a finite impulse response (FIR) filter, provided that filter is of sufficient length \( L \). Shown in Fig. 2 is an example of a typical room impulse response \( h_{RIR}(t) \).

![Fig. 2. Shown is the block-based representation of room impulse response (RIR).](image)

The process of separating \( r(t) \) from \( y(t) \) is called de-reverberation. Blind de-reverberation is estimation of \( s(t) \) from single channel audio recording without exploiting knowledge about \( h_{RIR}(t) \). In this paper, a blind reverbant signal extraction method \( [9] \) is considered.

As shown in Fig. 2 the room impulse response, \( h_{RIR}(t) \), can be divided into \( K + 1 \) blocks, that is, \( h_{RIR,i}(t) : i = 0, 1, \ldots, K \) (with corresponding frequency domain representation \( H_{RIR,i}(\omega) : i = 0, 1, \ldots, K \)). Assuming that each filter block is of same length, say \( L \) units of time. The \( y(t) \), based on block-based FIR filtering, can be expressed as,

\[
y(t) = \sum_{i=0}^{K} s(t) \ast \delta(t-iL) \ast h_{RIR,i}(t).
\]

Likewise, the reverberation signal component, \( r(t) \), can be expressed as,

\[
r(t) = \sum_{i=1}^{K} s(t) \ast \delta(t-iL) \ast h_{RIR,i}(t),
\]

and \( r(t) \) in frequency domain can be expressed as,

\[
R(\omega) = \sum_{i=1}^{K} S(\omega)Z^{-iL}H_{RIR,i}(\omega).
\]

We know that, the effect of an FIR filter can be reversed using an appropriate infinite impulse response (IIR) filter. The dry signal \( s(t) \), therefore, can be recovered from \( y(t) \) using an appropriate IIR filter given FIR filter response is known. However, under blind de-reverberation it is not possible to measure or derive FIR filter response directly. To overcome this limitation, perceptually relevant estimate of FIR filter block is used. The perceptually relevant estimates, \( \tilde{H}_{RIR,i}(\omega) : i = 0, 1, \ldots, K \), are obtained using magnitude response of block-based FIR filter, i.e.,

\[
\tilde{H}_{RIR,i}(\omega) \approx |H_{RIR,i}(\omega)|^2.
\]

The IIR filter uses \( Y(\omega) \) to estimate \( \tilde{S}(\omega) \) and \( R(\omega) \) as,

\[
\tilde{S}(\omega) = Y(\omega) - \sum_{i=1}^{K} \tilde{S}_i(\omega)Z^{-iL}\tilde{H}_{RIR,i}(\omega),
\]

where \( \tilde{S}_i(\omega) \) is an estimate of the true value of \( S_i(\omega) \). Here current input block for IIR filter, \( Y_0(\omega) \), consists of \( s_0(t) \) convolved with \( h_{RIR,0}(t) \) plus previous blocks of dry signal \( s_i(t) \) convolved with \( h_{RIR,i}(t) \), \( i = 1, \ldots, K \). Assuming \( \tilde{H}_{RIR,0}(\omega) = 1 \), the current block estimate of the dry signal therefore it can be expressed as,

\[
\tilde{S}_0(\omega) = Y_0(\omega) - \sum_{i=1}^{K} \tilde{S}_i(\omega)\tilde{H}_{RIR,i}(\omega),
\]

and similarly estimate of the reverberant signal component can be expressed as,

\[
\tilde{R}_0(\omega) = \sum_{i=1}^{K} \tilde{S}_i(\omega)\tilde{H}_{RIR,i}(\omega).
\]

As indicated earlier, magnitude response is used for block-based impulse response estimation, same assumption is extended for reverberation component estimation, i.e.,
In practice, we can replace $\approx$ with $=$ which ensures that $\bar{R}_0(\omega)$ is always positive. Therefore, the process

$$|\bar{R}_0(\omega)|^2 = \sum_{i=1}^{K} |\bar{S}_i(\omega)|^2 |\bar{H}_{RIR, i}(\omega)|^2$$

will only remove energy from $Y_0(\omega)$ when estimating $\bar{R}(\omega)$, and $\bar{R}(\omega) < Y_0(\omega)$.

Gain vector for estimating $\bar{R}(\omega)$, $G_R(\omega)$ is computed as,

$$G_R(\omega) = \sum_{i=1}^{K} \frac{|S_i(\omega)|^2 |\bar{H}_{RIR, i}(\omega)|^2}{Y_0(\omega)}$$

Following temporal smoothing operation is applied to further refine the gain vector, $G_{R,T}(\omega) = G_{R,T-1}(\omega)(1 - \alpha(\omega)) + G_{R,T}(\omega)\beta(\omega)$, where $\tau$ denotes the current time frame of the process and $\alpha(\omega)$ ranges between 0 and 1. Finally, estimate of the reverberation signal component can be expressed as, $\bar{R}_0(\omega) = G_R(\omega)Y_0(\omega)$. This process is repeated for each frame of the input signal.

The proposed method is applied to audio recordings to test its effectiveness. Shown in Fig. 3 are the temporal plots of the test recording captured in a reverberant environment (top), estimated dry signal (middle), and estimated reverberation signal (bottom). It can be observed from Fig. 3 that the blind de-reverberation scheme is effective in separating dry and reverberation components from the input audio.

Fig. 3. Shown in the top panel is the plot of the test recording $y(t)$, in the middle panel is the estimated $s(t)$, and in the bottom panel is the estimated $r(t)$.

### 2.2. Feature Extraction and Classification

The estimated reverberant signal, $r(t)$, is used for feature extraction. Sehr et al. [10] have shown that Mel-frequency Cepstral Coefficients (MFCC) and Logarithmic Mel-spectral Coefficients (LMSC) are effective for speech recognition [10]. The proposed scheme also uses MFCC and LMSC to capture traces of acoustic reverberation. A relatively small (i.e. 48-dimensional) feature space is used to characterize an acoustic environment. The 48-D feature vector is obtained by concatenating a 24-D MFSC and 24-D LMSC vectors. Shown in Fig. 4 is the flowchart of features extraction subsystem. A multi-class SVM classifier based on radial basis kernel function is used for AEI.

![Flowchart of feature extraction subsystem](image)

**Fig. 4.** Shown is the flowchart of feature extraction substage.

### 3. PERFORMANCE EVALUATION

Effectiveness of the proposed system is evaluated using a data set consisting human speech. Details for data set used, experimental setup, and experimental results are provided next.

#### 3.1. Data Set

Performance of the proposed scheme is evaluate using a data set consisting of 284 speech recordings. We recorded human speech of two speakers, a male and a female, in nine acoustic environments. Details of each acoustic environment is listed in Table 1. The data set is recorded using four commercial-grade microphones (see Table 2 for microphones type used) with 44.1 kHz sampling rate and 16 bits/sample resolution. The reverberation signal is estimated from each input recording using method discussed in Section 2.

<table>
<thead>
<tr>
<th>Environment</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_1$ &amp; $E_2$</td>
<td>Small Office 1 &amp; 3, (7’ × 11’ × 9’)</td>
</tr>
<tr>
<td>$E_3$</td>
<td>Large Office, (9’ × 10’ × 15’)</td>
</tr>
<tr>
<td>$E_4$ &amp; $E_5$</td>
<td>Restroom 1 &amp; 2</td>
</tr>
<tr>
<td>$E_6$ &amp; $E_7$</td>
<td>Hallway 1 &amp; 2</td>
</tr>
<tr>
<td>$E_8$ &amp; $E_9$</td>
<td>Outdoors 1 &amp; 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Microphone</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_1$ &amp; $M_2$</td>
<td>Cond. Mic. (BEHRINGER ECM8000)</td>
</tr>
<tr>
<td>$M_3$</td>
<td>Left Built-in Mic. of ZOOM R16 Recorder</td>
</tr>
<tr>
<td>$M_4$</td>
<td>Right Built-in Mic. of ZOOM R16 Recorder</td>
</tr>
</tbody>
</table>

### 3.2. Experimental Setup

The reverberant component is extracted from input speech signal. To this end, input audio is pre-emphasized according to $r(t) = r(t) - p \times r(t-1)$ with $p = 0.97$ before $r(t)$ estimation. The estimated $r(t)$ is then decomposed into overlapping frames each of 25 msec. duration and 10 msec. frame advancing. For each audio frame, a 48-D feature vector is extracted. The 48-D feature vector, averaged over all frames, is used for training and testing of the SVM classifier.

For classification, a multi-class SVM [11] with radial basis kernel function is used. To begin with, we randomly selected 50% of recordings from each category for training.
The rest 50% are used to verify performance our proposed scheme. The optimal parameters for the classifier are determined using grid search technique with five-fold cross-validation on training data. Each experiment is repeated 10 times. Classification accuracy, averaged over all runs, is used for performance evaluation.

3.3. Experimental Results

In the first experiment, we investigated microphone dependent AEI performance. To this end, SVM classifier was trained and tested using feature vectors extracted from recordings captured using same microphone type. The average classification accuracies for microphones $M_1$, $M_2$, $M_3$, and $M_4$ are 94%, 92%, 93%, and 92%, respectively. These results indicate that the AEI is independent of the microphone type.

In the second experiment, we investigated microphone independent blind AEI performance evaluation. To this end, SVM classifier was trained using feature vectors extracted from recordings captured with microphone $M_1$ and tested for feature vectors extracted from recordings captured with microphone $M_2$. Likewise, to investigate performance degradation due to microphone type variation, the SVM classifier was trained using recordings captured with microphone $M_1$ and tested for $M_3$. A negligible ($<1\%$) performance loss due to microphone variation was noted for this experiment. This observation also justifies flat microphone impulse response assumption used for reverberation modeling.

In the third experiment, we investigated performance of the proposed scheme for blind AEI. To this end, SVM classifier was trained using feature vectors extracted from recordings captured in the acoustic environments $E_1$, $E_4$, $E_6$, and $E_8$ and tested for feature vectors extracted from recordings captured in $E_2$, $E_5$, $E_7$, and $E_9$. Shown in the Table 3 is the average classification accuracies.

<table>
<thead>
<tr>
<th>True Class Label</th>
<th>Predicted Class Label</th>
<th>$E_1$</th>
<th>$E_4$</th>
<th>$E_6$</th>
<th>$E_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_2$</td>
<td>93%</td>
<td>1%</td>
<td>3%</td>
<td>3%</td>
<td></td>
</tr>
<tr>
<td>$E_5$</td>
<td>11%</td>
<td>89%</td>
<td>0%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>$E_7$</td>
<td>1%</td>
<td>34%</td>
<td>65%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>$E_9$</td>
<td>0%</td>
<td>1%</td>
<td>0%</td>
<td>99%</td>
<td></td>
</tr>
</tbody>
</table>

It can be observed from Table 3 that the proposed scheme is capable of blindly identifying all acoustic environments correctly with relatively high accuracy, except environment $E_7$. Relatively, large misclassification of $E_7$ (hallway) to $E_4$ (restroom) can be attributed to the fact that a hallway is a very complex acoustic environment. As both the restroom and the hallway are concrete structures, therefore, misclassification $E_7$ to $E_4$ is understandable.

In the final experiment, we investigated performance gain due de-reverberation. To this end, SVM classifier was trained and tested using feature vectors extracted from the unprocessed speech recordings (from $y(t)$). The average classification accuracies for with (and without) de-reverberation based identification systems for microphone $M_1$, $M_2$, $M_3$, and $M_4$ are 94%(84%) 92%(86%), 93%(86%), and 92%(86%), respectively. These results indicate that de-reverberation does improve classification accuracy.

4. CONCLUSION

We have described how audio reverberation can be modeled, estimated, and used for the AEI. The proposed scheme uses low dimensional feature space to characterize acoustic environment signature. We have shown effectiveness of the de-reverberation based AEI on a data set consisting of 284 speech recordings captured using 4 microphone in 9 environments. Experimental results indicate that the proposed scheme improves classification accuracy and is also applicable for blind AEI. We expect this approach to be a useful forensic tool when used in conjunction with other techniques that measure microphone characteristics, background noise, etc.

5. REFERENCES


