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 time. 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

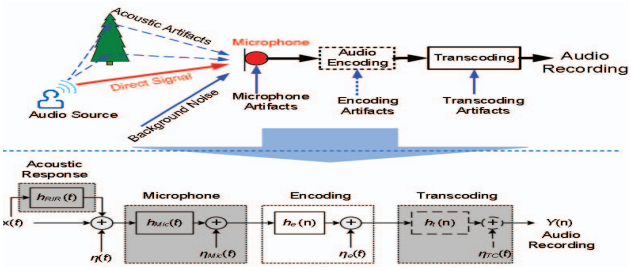

Fig. 1. A simplified digital audio recording diagram.

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* $\eta(t)$), microphone noise $\eta_{Mic}(t)$, encoding distortion $\eta_e(t)$, and transcoding artifacts $\eta_{TC}(t)$. 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 signals, and background noise at the input of the microphone can be expressed as, $\widetilde{x}(t) = s(t) + h_{RIR}(t) * s(t) + \eta(t)$, where $h_{RIR}(t)$ denotes room impulse response (RIR). If $h_{Mic}(t)$ denotes microphone impulse response, then $y(t)$ can be expressed as $y(t) = h_{final}(t) * u(t) + h_{Mic}(t) * \eta(t) + \eta_{Mic}(t) + \eta_{TC}(t)$, where $h_{final}(t)$ can be expressed as, $h_{final}(t) = h_{VT}(t) * h_{RIR}(t) * h_{Mic}(t)$.

The proposed framework estimates acoustic reverberation signal from audio recording, $y(t)$, and use it for AEI. What follows is an overview of how to model and estimate reverberation 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 reverberation signal, $h_{RIR}(t) * h_{Mic}(t) * s(t)$, 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 $y(t) = s(t) + h_{RIR}(t) * h_{Mic}(t) * s(t) + h_{Mic}(t) * \eta(t) + \eta_{Mic}(t) + \eta_{TC}(t)$. 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 $y(t) = s(t) + r(t) = s(t) + h_{RIR}(t) * s(t)$, where $s(t)$ represents direct sound component, also referred as dry signal and $r(t)$ represents the reverberant component. The reverberant signal $r(t)$ can expressed as $r(t) = s(t) * h_{RIR}(t)$.

The $h_{RIR}(t)$ can be modeled by a finite impulse response (FIR) filter, provided that filter is of sufficient length [9]. Shown in Fig. 2 is an example of a typical room impulse response $h_{RIR}(t)$.

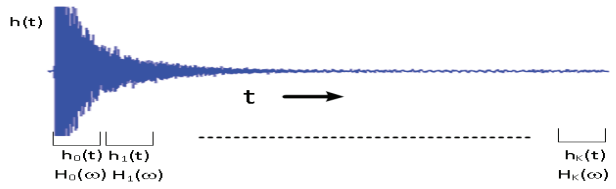

Fig. 2. Shown is the block-based representation of room impulse response (RIR).

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 reverberant 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, $h_{RIR,\,i}(t) : i = 0, 1, \cdots, K$ (with corresponding frequency domain representation $H_{RIR,\,i}(\omega) : i = 0, 1, \cdots, K$). 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,

$$y(t) = \sum_{i=0}^{K} s(t) * \delta(t - iL) * h_{RIR,\,i}(t). \tag{1}$$

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

$$r(t) = \sum_{i=1}^{K} s(t) * \delta(t - iL) * h_{RIR,\,i}(t), \tag{2}$$

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

$$R(\omega) = \sum_{i=1}^{K} S(\omega) Z^{-iL} H_{RIR,\,i}(\omega). \tag{3}$$

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, $\widetilde{H}_{RIR,\,i}(\omega) : i = 0, 1, \cdots, K$, are obtained using magnitude response of block-based FIR filter, i.e,

$$\widetilde{H}_{RIR,\,i}(\omega) \approx \mid H_{RIR,\,i}(\omega) \mid^2 . \tag{4}$$

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

$$\widetilde{S}(\omega) = \frac{Y(\omega) - \sum_{i=1}^{K} \widetilde{S}_i(\omega) Z^{-iL} \widetilde{H}_{RIR,\,i}(\omega)}{\widetilde{H}_{RIR,\,i}(\omega)}, \tag{5}$$

where $\widetilde{S}_i(\omega)$ 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 the signal $s_i(t)$ convolved with $h_{RIR,\,i}(t), i = 1, c \cdots, K$. Assuming $\widetilde{H}_{RIR,\,0}(\omega) = 1$, the current block estimate of the dry signal therefore it can be expressed as,

$$\widetilde{S}_0(\omega) = Y_0(\omega) - \sum_{i=1}^{K} \widetilde{S}_i(\omega) \widetilde{H}_{RIR,\,i}(\omega), \tag{6}$$

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

$$\widetilde{R}_0(\omega) = \sum_{i=1}^{K} \widetilde{S}_i(\omega) \widetilde{H}_{RIR,\,i}(\omega). \tag{7}$$

As indicated earlier, that magnitude response is used for block-based impulse response estimation, same assumption is extended for reverberation component estimation, i.e.,

$$| \widetilde{R}_0(\omega) |^2 = \sum_{i=1}^{K} | \widetilde{S}_i(\omega)\widetilde{H}_{RIR,\,i}(\omega) |^2 \quad (8)$$

$$\cong \sum_{i=1}^{K} | \widetilde{S}_i(\omega) |^2 | \widetilde{H}_{RIR,\,i}(\omega) |^2 \quad (9)$$

In practice, we can replace $\cong$ with $=$ which ensures that $\widetilde{R}_0(\omega)$ is always positive. Therefore, the process

$$| \widetilde{R}_0(\omega) |^2 = | Y_0(\omega) |^2 - | \widetilde{S}_0(\omega) |^2 \quad (10)$$

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

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

$$G_R(\omega) = \frac{\sum_{i=1}^{K} | S(\omega) |^2 \ | \widetilde{H}_{RIR,\,i}(\omega) |^2}{| \widetilde{Y}_0(\omega) |^2}. \quad (11)$$

Following temporal smoothing operation is applied to further refine the gain vector, $G_{R,\tau}(\omega) = G_{R,\tau-1}(\omega)(1 - \alpha(\omega)) + G_{R,\tau}(\omega)\beta(\omega)$, where τ 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, $\widetilde{R}_0(\omega) = G_R(\omega)Y_0(\omega)$. 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.

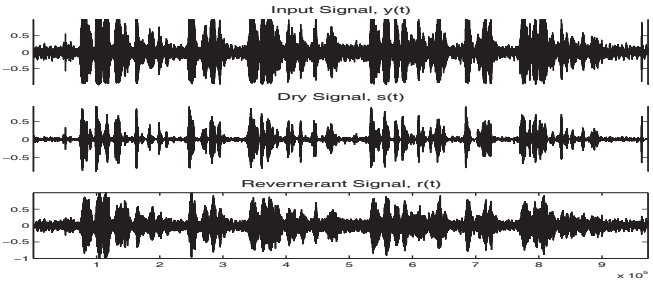

Fig. 3. Shown in the top panel is the plot of the test recording $y(t)$, in the middle panel is the estimated $s(t)$, and in the bottom panel is the estimated $r(t)$

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.

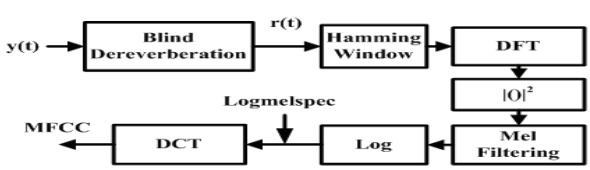

Fig. 4. Shown is the flowchart of feature extraction substage.

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 1. **Acoustic Environments Considered**

Environment	Description
E_1 & E_2	Small Office 1 & 3, $(7' \times 11' \times 9')$
E_3	Large Office, $(9' \times 10' \times 15')$
E_4 & E_5	Restroom 1 & 2
E_6 & E_7	Hallway 1 & 2
E_8 & E_9	Outdoors 1 & 2

Table 2. **Microphones Used**

Microphone	Description
M_1 & M_2	Cond. Mic. (BEHRINGER ECM8000)
M_3	Left Built-in Mic. of ZOOM R16 Recorder
M_4	Right Built-in Mic. of ZOOM R16 Recorder

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 3. Classification performance for blind identification

True Class Label	**Predicted Class Label**			
	E_1	E_4	E_6	E_8
E_2	93%	1%	3%	3%
E_5	11%	89%	0%	0%
E_7	1%	34%	65%	0%
E_9	0%	1%	0%	99%

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

[1] C. Grigoras, A. Cooper, and M. Michalek, "Forensic speech and audio analysis working group - best practice guidelines for enf analysis in forensic authentication of digital evidence," *European Network of Forensic Science Institutes*.

[2] D. Garcia-Romero and C. Espy-Wilson, "Speech forensics: Automatic acquisition device identification," *J. Acoust. Soc. Am.*, vol. 127(3), pp. 2044–2044, 2010.

[3] D. Nicolalde and J. Apolinario, "Evaluating digital audio authenticity with spectral distances and ENF phase change," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, 2009.

[4] C. Kraetzer, M. Schott, and J. Dittmann, "Unweighted fusion in microphone forensics using a decision tree and linear logistic regression models," in *Proc. ACM Multimedia and Security Workshop*, 2009.

[5] H. Malik and H. Farid, "Audio forensics from acoustic reverberation," in *Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, Dallas, TX, 2010.

[6] S. Ikram and H. Malik, "Digital audio forensics using background noise," in *Proc. IEEE Int. Conference on Multimedia and Expo 2010 (ICME10)*, 2010.

[7] U. Chaudhary and H. Malik, "Automatic recording environment identification using acoustic features," in *AES 129th Convention*, 2010.

[8] R. Ratnam, D. Jones, B. Wheeler, W. O. Jr., C. Lansing, and A. Feng, "Blind estimation of reverberation time," *J. of Acoustic Society of America*, vol. 5, no. 114, pp. 2877–2892, 2003.

[9] G. Soulodre, "About this dereverberation business: A method for extracting reverberation from audio signals," in *AES 129th Convention*, 2010.

[10] A. Sehr, R. Maas, and W. Kellermann, "Reverberation model-based decoding in the logmelspec domain for robust distant-talking speech recognitionr," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18(7), pp. 1676–1691, 2010.

[11] C. Chang and C. Lin. (2001) Libsvm: A library for support vector machines. [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/libsvm