Security Informatics
a SpringerOpen Journal

RESEARCH Open Access

Acoustic environment identification using unsupervised learning

Hafiz Malik[1*] and Hasan Mahmood[2]

Abstract

Acoustic environment leaves its characteristic signature in the audio recording captured in it. The acoustic environment signature can be modeled using acoustic reverberations and background noise. Acoustic reverberation depends on the geometry and composition of the recording location. The proposed scheme uses similarity in the estimated acoustic signature for acoustic environment identification (AEI). We describe a parametric model to realize acoustic reverberation, and a statistical framework based on maximum likelihood estimation is used to estimate the model parameters. The density-based clustering is used for automatic AEI using estimated acoustic parameters. Performance of the proposed framework is evaluated for two data sets consisting of hand-clapping and speech recordings made in a diverse set of acoustic environments using three microphones. Impact of the microphone type variation, frequency, and clustering accuracy and efficiency on the performance of the proposed method is investigated. Performance of the proposed method is also compared with the existing state-of-the-art (SoA) for AEI.

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 a set of *Iranian missile test images* release to international news media [2], 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 important for the science of digital forensics to keep pace.

The past few years have witnessed significant advances in image forensics [3], on the other hand, techniques for digital audio forensics are relatively less developed. An overview of the existing audio forensics methods can be found in [4] and references. Existing audio forensics methods based on signal characteristics can be broadly divided into the following categories:

1. The electrical network frequency (ENF) based framework that verifies integrity by comparing the extracted ENF with the reference ENF datebase [5-8]. These methods are effective against *cut-and-paste* (CAP) attacks, but complex electro-physical requirements of ENF-based approaches [5] make them ineffective for recordings made using battery-powered devices.
2. Statistical pattern recognition based techniques [9-17] 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.
3. Model driven approaches [18-24] have been proposed to address limitations of statistical learning based methods. These methods use mathematical models to realize artifacts due to acoustic reverberations [18-22] and distortions due to microphone nonlinearities [23]. Performance of model driven approaches depends on accuracy of the assumed model and reliability of the model parameter estimation method used.
4. Time-domain analysis based methods [25-29] have also been proposed to determine authenticity of digital audio recordings by capturing traces of lossy compression using encoder frame offsets in time domain [25-27] or detecting traces of "*butt-splicing*" in the digital recording using higher-order time-differences and correlation analysis [28].

*Correspondence: hafiz@umich.edu
[1]Electrical and Computer Engineering Department, University of Michigan - Dearborn, Dearborn, MI 48128, USA
Full list of author information is available at the end of the article

5. Spectral analysis based techniques [24,30] have been proposed for good quality audio recordings. H. Farid in [24] modeled the splicing process as a nonlinear operation and used bispectral analysis framework to capture traces of audio splicing. Similarly, C. Grigoras [30] proposed audio forensics framework based on statistical analysis such as long-term average spectrum histogram (LTASH) to detect traces of audio (re)compression, assess compression generation, and discriminate between different audio compression algorithms.

The research in the field of audio forensics can be broadly divided into the following major focus areas: (i) speech recognition that aims at producing readable text from human speech, especially from ambiguous utterances, (ii) speaker verification that compares a known voice to an unknown voice to determine the identity of the unknown voice, (iv) speaker localization that uses acoustic environment features such as reverberations and background noise to determine speaker location and acoustic environment, and (iv) speaker identification that performs comparison of similarities and differences of elements of speech such as bandwidths, fundamental frequency, prosody, vowel formant trajectory, occlusive, fricatives, pitch striations, formant energy, breath patterns, nasal resonance, coupling, and any special speech pathology of the speaker. Audio forensics focuses not only the direct speaker verification but also the recording environment identification [16] which can be used to determine the underlying facts about the evidentiary recording and to provide authoritative answers to questions, such as [31]:

- Is an evidentiary recording "*original*" or was it created by splicing multiple recordings together?
- What are the types and locations of forgeries, if there are any, in an evidentiary recording?
- Was the evidentiary recording made at location L, as claimed?
- Is the auditory scene in the evidentiary recording original or was it digitally altered to deceive the listener?

The acoustic environment identification (AEI) therefore has a wide range of applications ranging from audio recording integrity authentication to real-time crime acoustic space localization/identification. For instance, consider a scenario where a police call center receives an emergency call from a victim being harassed or chased by an offender. Under such crime situations it is very common that the harassed persons are unable to provide any relevant information about their actual location. The acoustic signals in the audio recording can be used to determine the acoustic space (i.e. car, street, neighborhood, living room, bath room, bed room, kitchen, etc.) of the crime scene. Similarly, for gun shooting cases, the sound of the firearms in the recording can be used to obtain important information about the crime scene such as weapon type.

The focus of this paper is acoustic environment identification (AEI) from evidentiary recording which has applications in the area of audio forensics, distant speech recognition, speaker localization. In the context of audio forensics, consider a test audio recording, obtained by *splicing* sections from one or multiple audio recordings made at different locations. When such spliced audio is used as evidence in the court of law its integrity must be verified. As doctored evidence can be used to fake the person, acoustic environment, event, auditory scene, acoustic environment, etc. in the evidentiary recording which might lead to serious consequence. It is therefore critical to authenticate the integrity of digital evidence.

This paper presents a model driven framework based on parametric modeling of late reverberations, parameter estimation using maximum likelihood estimation, and density-based clustering to determine where the recording was made. Motivation behind considering acoustic artifacts for AEI and audio forensic applications is that existing audio forensic analysis methods, e.g., ENF-based methods [5-8] and recording device identification based methods [11-14] cannot withstand *lossy compress attack*, e.g., MP3 compression. In our recent work [18,32], we have shown that acoustic reverberations can survive lossy compression attack, which is one of the motivations behind considering acoustic artifacts in an audio recording for AEI and digital audio forensic applications.

The major contribution of this paper is to develop a statistical framework for automatic AEI and its applications to digital audio forensics. Here, we exploit specific artifacts introduced at the time of recording to authenticate an audio recording and AEI, that is, to determine where the recording was made. 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 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 [34]. In this paper, we have shown that reverberation can be reliably estimated. In addition, effectiveness of the proposed method is evaluated for recorded audio and speech datasets. Moreover, to achieve automatic recording environment identification, density-based clustering is used. Performance of the proposed framework has also been evaluated for microphone type, frequency, and clustering accuracy and efficiency. Effectiveness of the proposed scheme has also

been evaluated using human speech recordings. Performance of the proposed method is also compared with the existing state-of-the-art (SoA) for AEI.

The rest of the paper is organized as follows: details of reverberation acoustic environment artifacts modeling and estimation are provided in Section 'Proposed method'; a brief overview of the density-based clustering is described in Section 'Automatic acoustic environment identification (AEI) using cluster analysis'; experimental results and performance analysis are provided in Section 'Experimental results'; and concluding remarks along with future research directions are discussed in Section 'Conclusion'.

Proposed method

Parametric modeling of acoustic environment artifacts

Consider a recorded response of an acoustic environment to an impulsive sound source "*a hand-clap*" shown in the Figure 1. It can be observed from Figure 1 that the recorded response can be divided into two non-overlapping segments: (i) strong early reflections (also known as early reverberations), and (ii) decaying reverberant tail or late reverberations. The early reflections are assumed to occur between the arrival of the *direct signal* and t_{ref} ms thereafter; whereas the late reverberations are occurring after t_{ref} ms, a typical value for $t_{ref} \in [50 - 100]$ ms [35]. Early reverberations depend on distance between the source and the receiver (e.g. microphone), directivity of source and receiver pair, etc. The late reverberations, on the other hand, depend on acoustic environment characteristics, e.g., enclosure geometry, surface area, surface material absorption coefficient, and so on. In this paper, We focus on late reverberations for the acoustic environment identification task.

We begin with a model for the late reverberations of acoustic activities in an acoustic environment (the dense reflections that follow the early reflections). The late reverberations are a result of multiple reflections, arriving at the receiver in *random order*, with successive reflections being damped based on the arrival time, that is, reflection amplitude is damped to a greater degree if they arrive (at the receiver) later in time. The assumption of randomness is very important to the development of a statistical model used for reverberation modeling and estimation. It has been demonstrated [36] that when a burst of white noise is radiated into a test enclosure, the phase and amplitudes of the normal modes are random in the instant preceding the cessation of the sound. This generates random decaying output of the enclosure following sound cessation, even if repeated trials were conducted with the same source and receiver geometry.

To validate these claims, that is, (i) output of the enclosure following sound cessation is random, and (ii) decaying tails of the repeated trials are uncorrelated, we computed a cross-correlation function between two non-overlapping segments of same decaying tail. In addition, we also computed cross-correlation between two identical segments of two decaying tails of same "hand-clap" recorded at two different time instances. Shown in Figure 2 is the plot of cross-correlation function of two non-overlapping segments (35 msec. apart) of a decaying tail of same "hand-clap" recorded using microphone *Mic*1 in a restroom (a highly reverberant environment, E_5). And, shown in Figure 3 is the plot of cross-correlation function of two time-aligned segments of decaying tails of identical "*hand-clap*" recordings made at two different time instances with *Mic*1 in E_5 (i.e., by playing the same "hand-clap" recording twice).

It can be observed from Figures 1, 2 and 3 that the assumed model for late reverberations is reasonably accurate. Based on these observations, reverberant decaying tail envelope can modeled using an exponential with a single (deterministic) parameter, decay rate τ. As demonstrated in Figure 3 that late reverberations are uncorrelated, the reverberant or dense tail is therefore modeled using an exponentially damped uncorrelated noise sequence obeying Gaussian distribution. More specifically, the decay of an audio signal $x[n]$ is

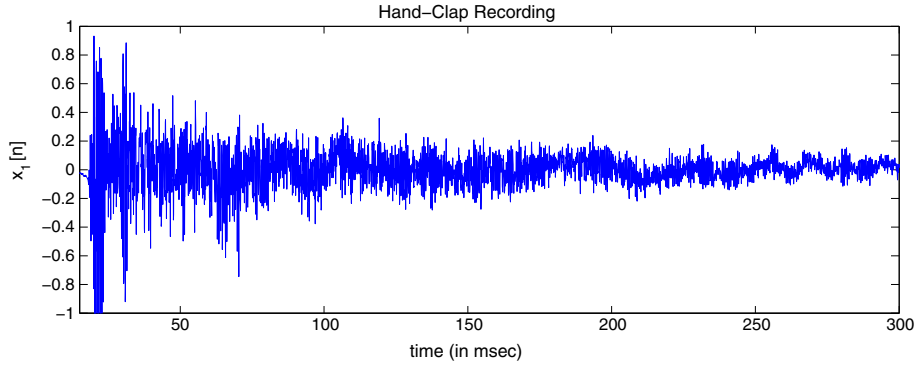

Figure 1 Shown is the plot of a "hand-clap" recording made with microphone *Mic*1 in a reverberant acoustic environment *E*5.

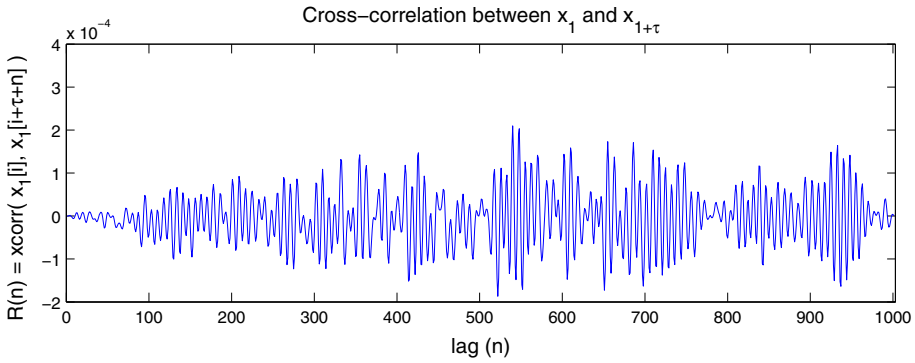

Figure 2 Shown is the plot of cross-correlation function between two non-overlapping segments (35 msec. apart) of decaying tail of a "hand-clap", recorded using microphone *Mic*1 in a reverberant environment *E*5.

modeled with a multiplicative decay and additive noise (see Figure 4):

$$y[n] = d[n]\,x[n] + \eta[n]\,, \tag{1}$$

where,

$$d[n] = \exp[-n/\tau]\,. \tag{2}$$

The decay parameter τ embodies the extent of the reverberation, and can be estimated using a maximum likelihood estimator.

It is important to mention that the proposed time-domain model does not include the direct sound or early reflections and it is accurate only during free decay, that is, when the sound source is not active. To capture traces of acoustic environment, decay rate of the reverberant tail is estimated from the exponentially decaying envelop.

Parameter estimation using maximum likelihood estimation

We assume that the signal $x[n]$ is a sequence of N independently and identically-distributed (*iid*) zero mean and normally distributed random variables with variance σ^2. We also assume that this signal is uncorrelated to the noise $\eta[n]$ which is also a sequence of N *iid* zero mean and normally distributed random variables with variance $\sigma_\eta^2 = \rho \times \sigma^2$, where ρ is a real-valued positive constant representing the signal to noise ratio (SNR). With these assumptions, the observed signal $y[n]$ is a random variable with a probability density function given by:

$$P_{y[n]}(k) = \frac{1}{\sqrt{2\pi\sigma^2\gamma^2[n]}} \cdot \exp\left(-\frac{k^2}{2\sigma^2\gamma^2[n]}\right), \tag{3}$$

where

$$\gamma[n] = \sqrt{\exp(-2n/\tau) + \rho^{-1}}. \tag{4}$$

The likelihood function is then given by:

$$\begin{aligned} Ł(y,\sigma,\gamma) = {} & \frac{1}{(2\pi\sigma^2)^{N/2}\prod_{k=0}^{N-1}\gamma(k)} \cdot \exp \\ & \times \left(-\frac{1}{2\sigma^2}\sum_{k=0}^{N-1}\frac{y^2(k)}{\gamma^2(k)}\right). \end{aligned} \tag{5}$$

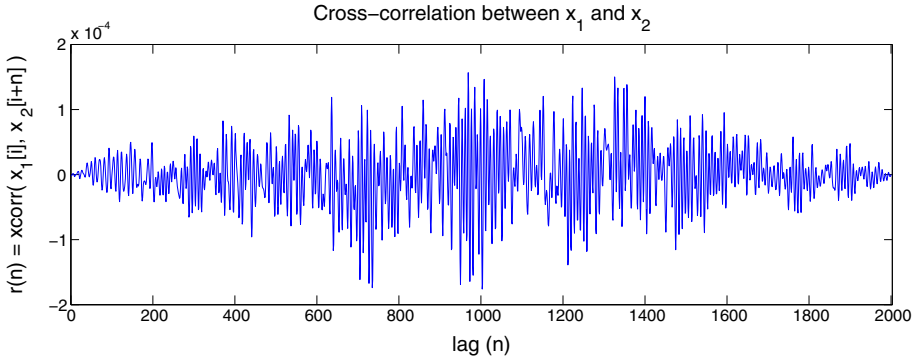

Figure 3 Shown is the plot of cross-correlation function between time-aligned segments of decaying tails of identical "*hand-clap*" recordings made at two different time instances using *Mic*1 in E_5.

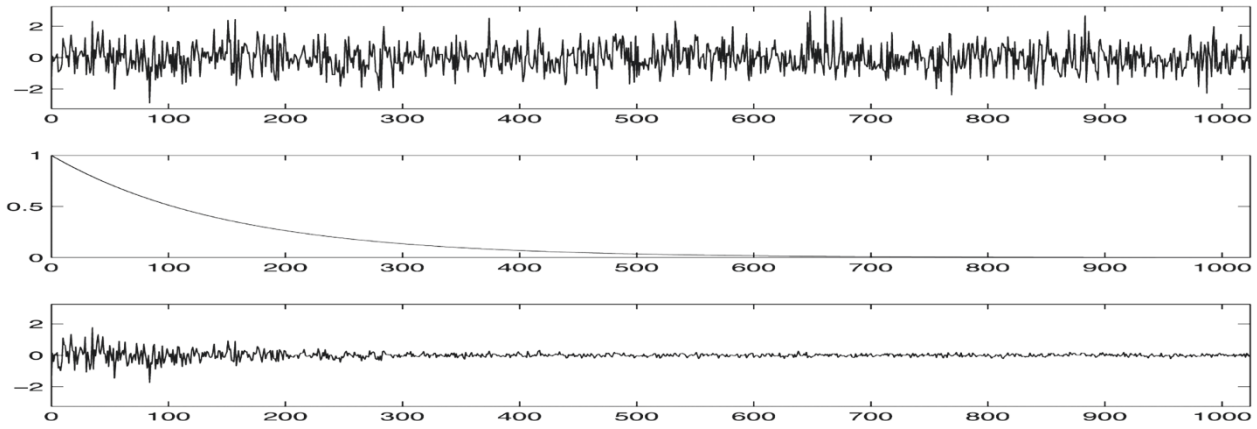

Figure 4 Shown from top to bottom are: a signal *x*[*n*]; the exponential decay *d*[*n*]; and the resulting decayed signal *y*[*n*] with additive noise.

The log-likelihood function, $\ln(L(\cdot))$, is:

$$\mathcal{L}(y,\sigma,\gamma) = -\frac{N}{2}\ln(2\pi\sigma^2) - \sum_{k=0}^{N-1}\ln(\gamma(k)) - \frac{1}{2\sigma^2}\sum_{k=0}^{N-1}\frac{y^2(k)}{\gamma^2(k)}. \tag{6}$$

The decay parameter τ is estimated by maximizing the log-likelihood function $\mathcal{L}(\cdot)$. This is achieved by setting the partial derivatives of $\mathcal{L}(\cdot)$ equal to zero and solving for the desired τ. For the purpose of numerical stability, the maximization is performed on $\tilde{\tau} = \exp(-1/\tau)$ instead of τ.

$$\frac{\partial\mathcal{L}}{\partial\sigma} = -\frac{N}{\sigma} + \frac{1}{\rho\sigma}\sum_{k=0}^{N-1}\frac{1}{\gamma^2(k)} + \frac{1}{\sigma^3}\sum_{k=0}^{N-1}\frac{y^2(k)\tilde{\tau}^{2k}}{\gamma^4(k)} \tag{7}$$

$$\frac{\partial\mathcal{L}}{\partial\tilde{\tau}} = \sum_{k=0}^{N-1}\frac{k\tilde{\tau}^{2k-1}}{\gamma^2(k)}\left(\frac{y^2(k)}{\sigma^2\gamma^2(k)} - 1\right). \tag{8}$$

It can be observed from Eq. (7) and Eq. (8) that both the σ in Eq. (7) and $\tilde{\tau}$ in Eq. (8) cannot be solved for analytically. As such, an iterative non-linear minimization is required which is computationally inefficient and sometime does not convergence. To get around this issue, high signal to noise ratio (SNR) is assumed in the selected decaying tail region, i.e., $\sigma \gg \sigma_\eta$ or $0 < \rho \ll 1$. This is a realistic assumption, especially, when audio recording is made in a relatively quiet environment and/or it is pre-processed for speech enhancement. Experimental results presented here are based on audio recordings made in quiet acoustic environments and are pre-processed with a speech enhancement filter [37]. With moderate SNR assumption, the Eq. (7) and Eq. (8) can be rewritten as:

$$\frac{\partial\mathcal{L}}{\partial\sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{k=0}^{N-1}\frac{y^2(k)}{\tilde{\gamma}^2(k)} \tag{9}$$

$$\frac{\partial\mathcal{L}}{\partial\tilde{\tau}} = \sum_{k=0}^{N-1}\frac{k\tilde{\tau}^{2k-1}}{\tilde{\gamma}^2(k)}\left(\frac{y^2(k)}{\sigma^2\tilde{\gamma}^2(k)} - 1\right). \tag{10}$$

where,

$$\tilde{\gamma}[k] = \tilde{\tau}^k. \tag{11}$$

Although σ in Eq. (9) can be solved for analytically, $\tilde{\tau}$ in Eq. (10) still cannot. As such, an iterative non-linear minimization is required. This minimization consists of two primary steps, first to estimate σ and second to estimate $\tilde{\tau}$. In the first step σ is estimated by setting the partial derivative in Eq. (9) equal to zero and solving for σ, to yield:

$$\sigma^2 = \frac{1}{N}\sum_{k=0}^{N-1}\frac{y^2(k)}{\tilde{\gamma}^2(k)} = \frac{1}{N}\sum_{k=0}^{N-1}\frac{y^2(k)}{\tilde{\tau}^{2k}}. \tag{12}$$

This solution requires an estimates of σ_η and $\tilde{\tau}$. The $\tilde{\tau}$ is initially estimated using Schroeder's integration method [38]. In the second step, $\tilde{\tau}$ is estimated by maximizing the log-likelihood function $\mathcal{L}(\cdot)$ in Eq. (6). This is performed using a standard gradient descent optimization, where the derivative of the objective function is given by Eq. (10). These two steps are iteratively executed until the differences between consecutive estimates of σ and $\tilde{\tau}$ are less than a specified threshold. In practice, this optimization is quite efficient, converging after only a few iterations.

Automatic Acoustic Environment Identification (AEI) using cluster analysis

The similarity of the estimated acoustic reverberation parameters, τ and σ, from selected segments of a given audio recording can be used for both forensic analysis and acoustic environment identification (AEI). For example, a small (resp. large) distance in the estimated reverberation parameters from a test recording indicates relatively consistent (resp. inconsistent) acoustic environment. In addition, similarity in the estimated reverberation parameters from two different recordings indicates that these recordings were made in acoustically identical environments and vice versa.

Cluster analysis, *an unsupervised classification framework*, is used to determine acoustic environment similarity in the test audio recording using acoustic parameters. For automatic AEI, density based clustering is considered. More specifically, ***D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise* (DBSCAN) [39,40], a density based clustering technique, is used to label audio recordings into acoustically similar groups (or clusters) based on estimated acoustic parameters. Motivation behind considering DBSCAN is that it can efficiently handle outliers in the data, it can find clusters of arbitrary (or non-convex) shapes, and it does not require prior knowledge of the number of clusters in the data. In addition, density-based clustering handles regions of varying densities more efficiently than commonly uses methods such as K-means, K-mediods, etc.

The DBSCAN uses center-based framework to estimate density for a particular point in the data set. More specifically, it counts the number of points in radius, ϵ, around point p. This center-based framework labels a given point, p, as (i) a *core point*, (ii) a *border point*, or (iii) a *noise point*. The core, border, and noise points are defined as:

- **Definition 1:** A point, p is a *core point* if $\mid \{x \mid d(x,p) \leq \epsilon\} \mid \geq MinPts$, where *MinPts* denotes minimum number points and $d(x,p)$ denotes the Euclidian distance between of point x and p. The core points makes the interior of a cluster.
- **Definition 2:** A point, p is a *border point* if $\mid \{x \mid d(x,p) \leq \epsilon\} \mid < MinPts$ but is in the neighborhood of a core point.
- **Definition 3:** A point is a *noise point* if it is neither a core point nor a border point.

Given the definitions of core, border, and noise points the DBSCAN algorithm can be described as follows. Label data points as core, border, and noise points and remove noise points. Clustering is then performed by assigning same cluster label to any two core points that are within ϵ-distance. Likewise, any border point within ϵ-distance from a core point is also assigned the same cluster label as the core point.

Experimental results

To test effectiveness of the proposed framework, we analyzed simple recordings, e.g., hand-clap recordings and relatively complex recordings, e.g., speech recordings made in a diverse set of recording environments including small offices, a large office, hallway, staircase, restroom, atrium, and outdoor environments. These recordings were made using three microphones.

Dataset and experimental settings

Two datasets consisting of audio recordings are used for performance evaluation of the proposed method.

- The first dataset used for performance evaluation consists of 120 hand-clap recordings made using three microphones: (i) **Mic1: a built-in HP Compaq Laptop**, (ii) **Mic2: a built-in microphone in Apple's MacBook**, and (iii) **Mic3: a commercial grade external microphone**. These recordings were made in ten acoustically different environments: three *small offices* ($E1 - E3$), an *atrium* $E4$, a *restroom* $E5$, a *hallway* $E6$, two *outdoors* $E7$&$E8$, a *large office* $E9$, and *stairs* $E10$. The hand-clap recording (downloaded from http://www.freesound.org/samplesViewSingle.php?id=345) was played using a pair of commercial grade external speakers. In each recording environment three samples were made through each microphone while keeping the distance between a pair of speakers and the microphone same. These recording were made with mono audio channel and a sampling rate of 16000 samples per second.
- The second dataset used for performance evaluation of the proposed method consists of 60 speech recordings. We recorded human speech of three speakers (two males and a female) in four different recording environments: *outdoors* E_1; a *small office* ($7' \times 11' \times 9'$) E_2; *stairs* E_3; and a *restroom* ($15' \times 11' \times 9'$) E_4. In each recording environment, each speaker read five different texts (each consisting of couple of short sentences) while keeping the distance between the speaker and the microphone same, as a result, a total of 60 audio recordings were made using a commercial-grade external microphone.

Each recording was initially pre-processed with a speech enhancement filter [37]. For acoustic parameter estimation, decaying tails were manually selected from each clean recording. From the selected tails acoustic reverberation parameters, i.e., τ and σ were estimated using method discussed in 'Parameter estimation using maximum likelihood estimation'.

Clustering performance is evaluated using clustering *purity*, *efficiency*, and *Jaccard* scores [40]. These clustering assessment measures are defined as follows:

$$Purity = \frac{f_{11}}{f_{11} + f_{10}} \tag{13}$$

$$Efficiency = \frac{f_{11}}{f_{11} + f_{01}} \tag{14}$$

$$Jaccard = \frac{f_{11}}{f_{11} + f_{10} + f_{01}} \tag{15}$$

where f_{11} is the number of pairs that are labeled correctly, f_{10} is the number of pairs that are labeled together in the true data, but not in the predicted labels, and f_{01} is the number of pairs that are labeled together during clustering but are not in the true labels.

Automatic AEI: *hand-clap recording*

The goal of the first experiment is to tested performance of the proposed framework for AEI using hand-clap recordings. In this experiment, the reverberation parameters, τ and σ^2, were estimated from selected decaying tails in each of the recording in the first dataset using the method discussed in Section 'Parameter estimation using maximum likelihood estimation'. The decaying tails were manually selected. The noise floor criterion is used for manual tail selection, that is, tail on-set starts at the peak hand-clap energy level and ends at the position where it has decayed to the noise floor. Selected decaying tails are used to estimate reverberation parameters.

Shown in Figure 5 is the scatter plot of estimated reverberation parameters, τ (in msec) and variance σ^2 (in $\log \sigma^2$), from the hand-clap recordings made with Mic1 (the built-in HP Compaq Laptop). Shown in the left panel of Figure 5 is the true acoustic environment labels for the estimated parameters and shown in the right panel of Figure 5 is the predicted acoustic environment labels using DBSCAN-based clustering. We iteratively refined the clustering parameters to partition the input data into at least seven clusters. Shown in Figure 5 are the clustering results obtained with clustering parameters $\epsilon = 1.8$ and $MinPt = 3$.

The learned clustering parameters, *MinPts* and ϵ, from recordings made with Mic1 are used for predicting cluster labels for recordings made using Mic2 and Mic3. Shown in the top panel of Figure 6 is the scatter plot of estimated parameters from hand-clap recordings made with Mic2 (built-in microphone on Apple's MacBook) along with true acoustic environments and shown in the bottom panel of Figure 6 is the predicted acoustic environment labels using DBSCAN-based clustering.

And, shown in the top panel of Figure 7 is the scatter plot of estimated parameters from hand-clap recordings made with Mic3 (external microphone) along with true acoustic environments and shown in the bottom panel of Figure 7 is the predicted acoustic environment labels using DBSCAN-based clustering.

Following observations can be made from Figures 5, 6 and 7:

- It can be observed from Figure 5 that clustering process has accurately predicted environment labels (or cluster IDs (CIDs)) for all acoustic environments except small offices ($E1 - E3$) where it has predicted same environment label, e.g., $CID1$. As all three small offices are structurally identical and the only difference between them is their furniture settings therefore acoustic characteristics are these environments are expected to be very close, the true labels in the left panel of Figure 5 confirms it. In addition, it also indicates that Mic1 is relatively less sensitive to small variations in the acoustic environment therefore forensic analyst should be careful when using such microphones for audio forensic applications.
- Secondly, Figure 6 shows that clustering process has accurately predicted environment labels for acoustic environments $E1$, $E2$, $E3$ and $E10$; whereas, it has assigned two separate labels $CID5$ & $CID6$ to $E5$, same label $CID4$ to $E4$, $E7$ and $E8$, and same label $CID7$ to $E8$ and $E9$. It indicates that Mic2 is insensitive in less reverberant environments and it is relatively more sensitive to highly reverberant environments. Findings of Figure 6 also suggest that Mic2 is not a good choice to differentiate between acoustically similar environments such as outdoors and atrium, and large office and hallway.
- Thirdly, Figure 7 shows that clustering process has accurately predicted environment labels for all acoustic environments with two exceptions, that is, (i) two labels, e.g., $CID7$ & $CID8$, for $E6$, and (ii) miss classification of few data points of $E4$. In addition, clustering process has also assigned '*noise*' label to data points of acoustic environments $E3$, $E4$, $E6$ and $E10$. It can also be observed from Figure 7 that estimated parameters for Mic3 exhibit larger variance than the other two microphones used. The larger variance of Mic3 indicates that it exhibits relatively higher sensitivity (see Section 'Performance evaluation: *microphone variation*' for more discussion on microphone sensitivity).
- Finally, Figures 5, 6 and 7 indicate that Mic1 and Mic3 exhibit relatively higher accuracy than Mic2. We have also learned through extensive analysis that prediction performance of the proposed method can be improved by learning microphone dependent clustering parameters, that is, learning microphone

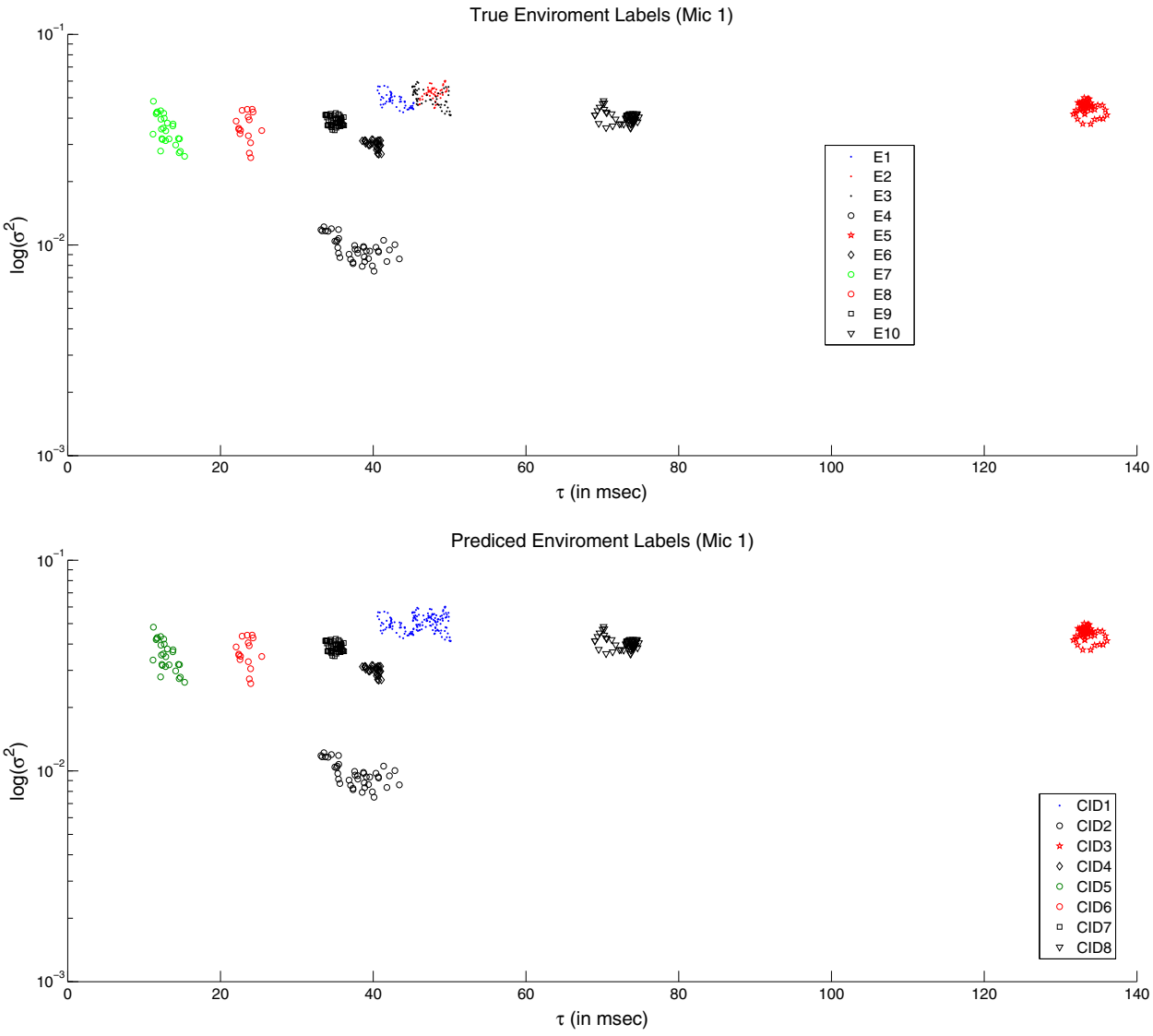

Figure 5 Shown in the top panel is the scatter plot of the estimated parameters from the hand-clap recordings made with Mic1 with true acoustic environment labels and shown in the bottom panel is the scatter plot of the estimated parameters with predicted environment labels.

specific clustering parameters and use them for environment prediction for recordings made.

To quantify microphone specific performance of the proposed method, AEI accuracy is measure is used. To this end, AEI accuracy is measured in terms of clustering purity, efficiently, and Jaccard scores defined in Equations (13-15). Shown in Table 1 is the microphone specific clustering performance.

It can be observed from Table 1 that Mic3 exhibits higher AEI accuracy than the other two microphones and Mic2 exhibits the lowest AEI accuracy than the other two microphones. Higher accuracy of Mic3 can be attributed to it better sensitivity and lower sensitivity of Mic2 resulted in lower AEI accuracy.

Performance evaluation: *microphone variation*

The aim of the second experiment is to investigate the impact of microphone type on the accuracy of the estimated parameters. To this end, we compared reverberation parameters estimated from recordings made in a given acoustic environment simultaneously using all three microphones. We have observed through this analysis (it can also be observed from Figures 5, 6 and 7) that microphone sensitivity to an acoustic activity does influence estimated acoustic parameters. For example, estimated τ,

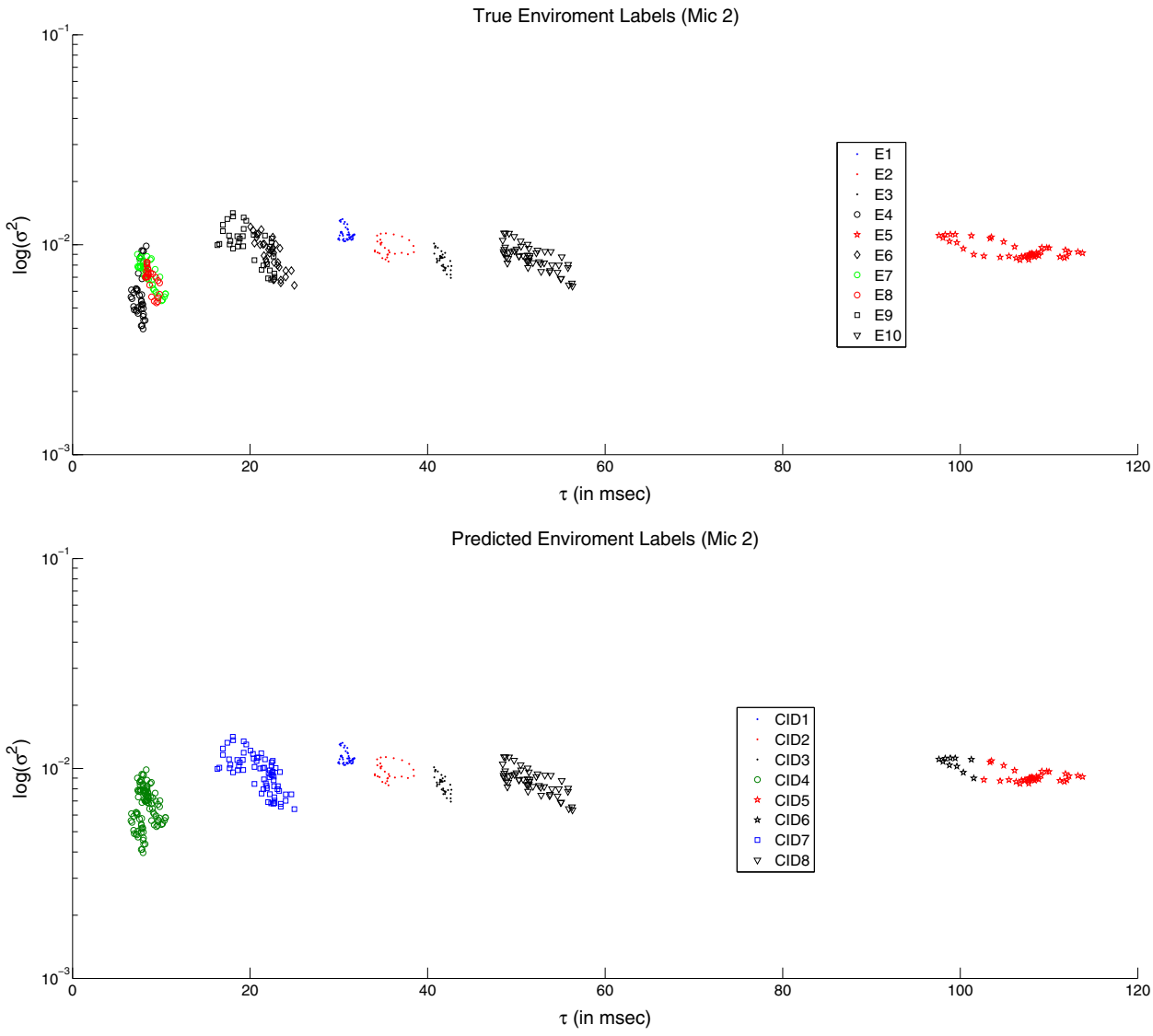

Figure 6 **Shown in the top panel is the scatter plot of the estimated parameters from the hand-clap recordings made through Mic2 with true acoustic environment labels and shown in the bottom panel are the scatter plot of the estimated parameters with predicted environment labels.**

for each acoustic environment, for Mic2 has significantly lower mean values, ($\mu_\tau = \frac{\sum_{i=1}^{n}\tau_i}{n}$), than μ_τ for Mic1 and Mic3. Similarly, μ_σ and standard deviation (std) of estimated σ^2, $\sigma_{\sigma^2} = \sqrt{(\frac{1}{n}\sum_1^n(\sigma_i^2 - \mu_{\sigma^2}))}$, where μ_{σ^2} is the mean value of sequence σ^2 of length n, for Mic2 is relatively larger than σ_{σ^2}s for remaining two microphones. To highlight this fact, we compared estimated parameters from recordings made in a given acoustic environment with all three microphones. Shown in Figure 8 are the scatter plots of the estimated τ and $\log(\sigma^2)$ for acoustic environment $E1$.

It can be observed from Figure 8 that the μ_τ and σ_τ for Mic3 is significantly larger than the μ_τ and σ_τ for Mic2. Similarly, μ_{σ^2} value for Mic3 is also larger than the other two microphones. This observation can be explained using the fact that Mic3 is an external microphone, therefore, it is expected to exhibit better sensitive to acoustic activities and ambient noise than the built-in laptop microphones.

To investigate the microphone response variations further, we selected two acoustic environments: (i) a less reverberant environment (outdoors), and (ii) a highly reverberant environment (restroom). Shown in the left panel of Figure 9 are the scatter plots of the estimated τ and $\log(\sigma^2)$ for all three microphones for acoustic environments $E7$ and shown in the right panel are the scatter

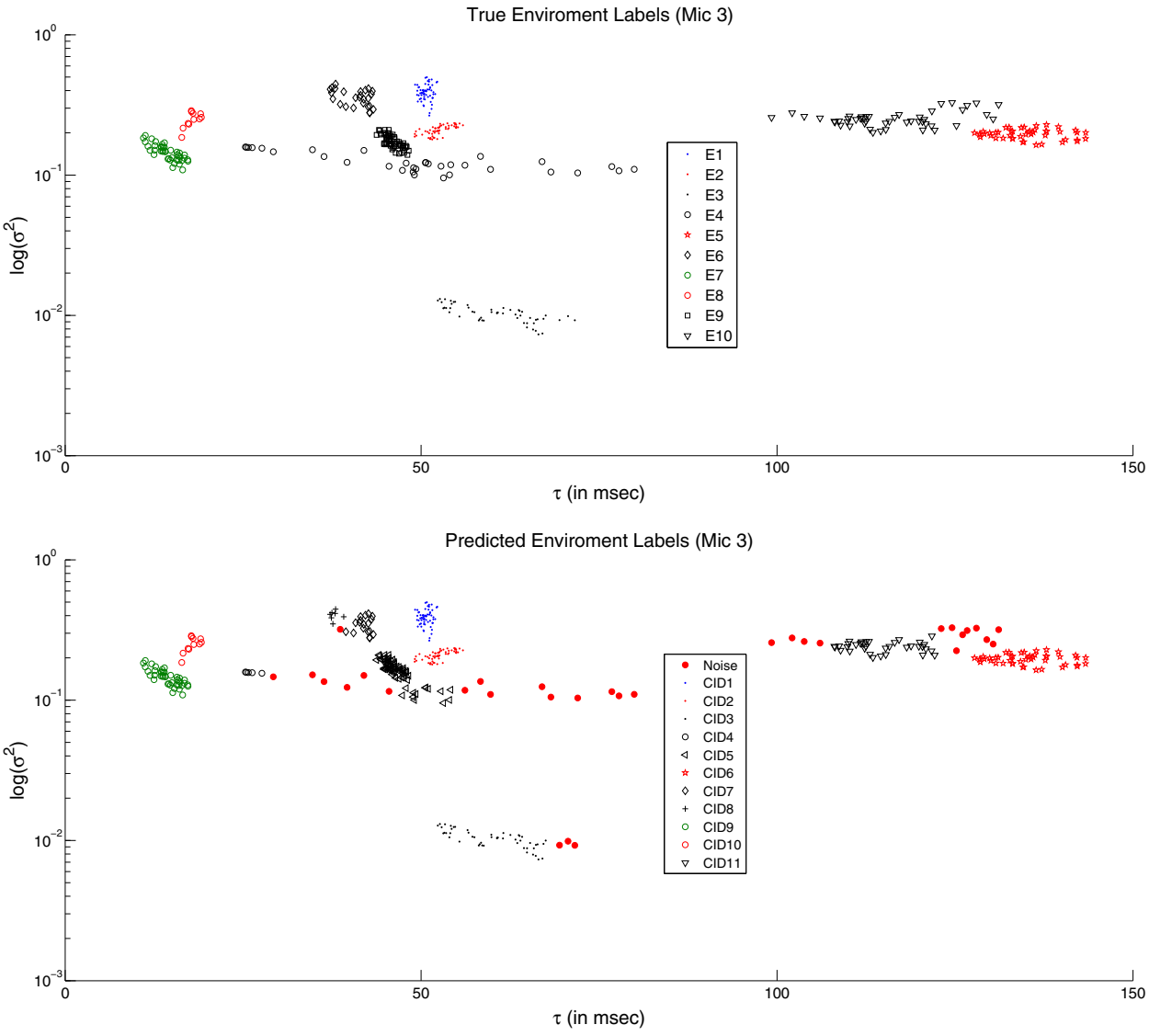

Figure 7 **Shown in the top panel is the scatter plot of the estimated parameters from the hand-clap recordings made with Mic3 with true acoustic environment labels and shown in the bottom panel is the scatter plot of the estimated parameters with predicted environment labels.**

Table 1 Shown AEI performances in terms of clustering accuracy for external, built-in HP, and built-in MacBook microphones

Microphone type	Purity score	Efficiency score	Jaccard score
Mic1: Built-in (HP)	75.17	75.17	60.22
Mic2: Built-in (MacBook)	73.64	71.80	58.22
Mic3: External	87.93	94.84	83.91

plots of the estimated parameters for all three microphones for acoustic environments *E*5.

Following observations can be made from Figures 8 and 9:

1. For external microphone: The μ_τ, μ_{σ^2}, σ_τ, and σ_{σ^2} for external microphone (as expected) is higher than the built-in microphones. This indicates that external exhibits relatively more sensitivity than the other two microphones.
2. For Mic1: The σ_τ and σ_{σ^2} for Mic1 is the lowest among all three microphones which makes it more suitable for forensics applications.

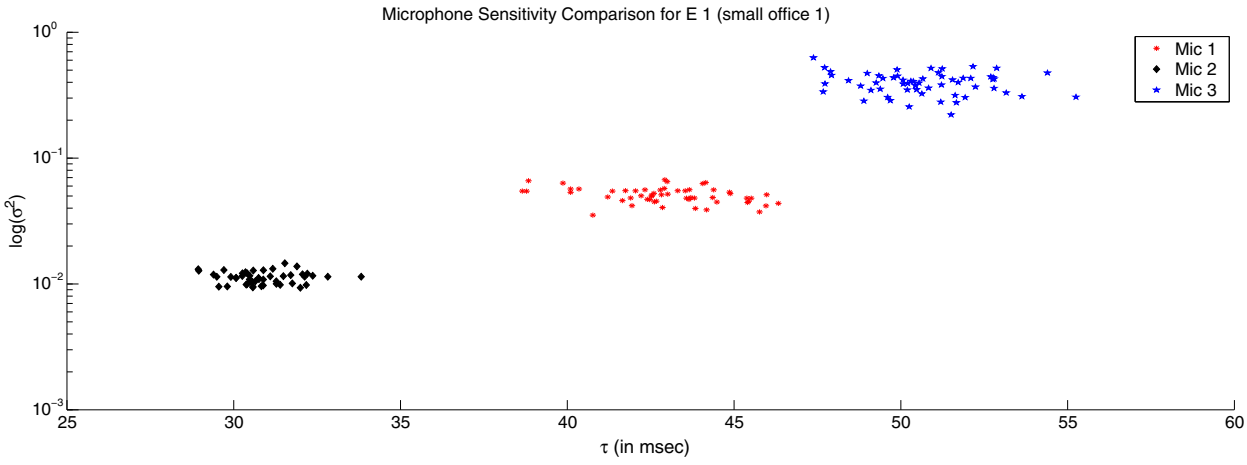

Figure 8 Shown is the scatter plots of the estimated parameters from the hand-clap recordings made with all microphones in the small-office 1.

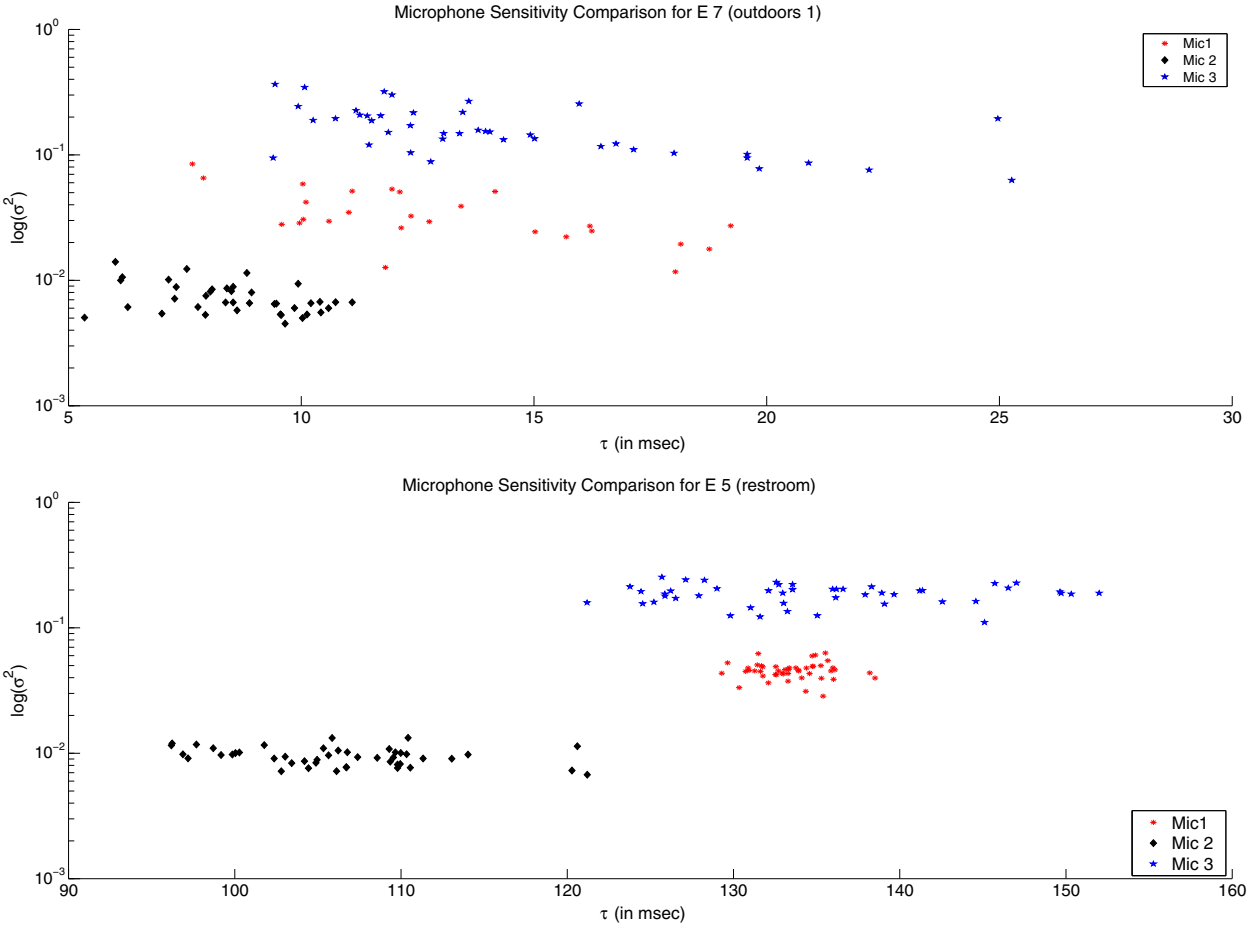

Figure 9 Shown in the top panel are the scatter plots of the estimated parameters from recordings made in acoustic environment *E7* and shown in the bottom panel are the scatter plots of the estimated parameters from recordings made environment *E5*.

3. For Mic2: The The μ_τ and μ_{σ^2} for Mic2 is lowest whereas variance is comparable to the external microphone for reverberant environments, e.g., $E1$ and $E5$.

To quantify variations in the estimated parameters, we computed mean and standard deviation (std) for each environment. Shown in Table 2 are the mean(std) of estimated parameters, τ and σ^2, for all microphones and all acoustic environments.

It can be observed from Table 2 that for all acoustic environments the μ_τ values of estimated τ for the external microphone and the built-in HP microphone are relatively close; whereas, the μ_τ for Mic2 are significantly lower than the other two microphones. This was a surprising observation, as recordings used were collected simultaneously using all three microphones, a small variation in the estimated parameters is understandable but a significant variation came as a surprise to us. Further investigation on recordings captured with Mic2 revealed that it has the lowest sensitivity among all three microphones used for data collection.

It can also be observed from Table 2 that the built-in MacBook microphone (Mic2) is relatively insensitive as compare to remaining two microphones. In addition, reliability of the estimated parameters decreases for complex acoustic structures such as atrium and stairs. This is due to the fact that due to low sensitivity Mic2 is unable to pick weak late reverberations and background noise which resulted in lower variance of the estimated parameters.

Finally, the external microphone is relatively more unreliable in complex environments than the built-in microphones which is reflected by a relatively large variance of the estimated τ for these environments. This is not a surprising observation as due to higher sensitivity, the Mic3 is expected to pick weak late reverberations mixed with background noise hence relatively higher variance of the estimated parameters.

In addition, as observed from Table 2 estimated τ is relatively higher for external microphone than the built-in microphones, this observation suggests that estimated reverberation parameter depends on microphone directivity and sensitivity. Therefore, for AEI and audio splicing detection performance microphones with superior directivity and sensitivity should be considered.

Impact of the frequency on estimated parameters

The goal of our third experiment is to investigate the impact of frequency on estimated parameters (e.g. τ and $\log(\sigma^2)$). To this end, each audio recording is decomposed

Table 2 Shown in third, fourth, and fifth columns are the mean(std) of estimated acoustic parameters from audio recordings made with the built-in HP, built-in MacBook, and external microphones, respectively

Environments		Microphones		
		Mic1: Built-in (HP)	**Mic2: Built-in(MacBook)**	**Mic3: External**
Small office1	τ mean(std)	42.95(1.46)	30.86(0.53)	50.68(0.74)
	$\sigma_n$mean(std)	0.05(0.004)	0.01(0.001)	0.39((0.052)
Small office2	τ mean(std)	47.93(0.99)	35.60(1.29)	52.50(1.7)
	σmean(std)	0.05(0.004)	0.01(0.001)	0.21(0.017)
Small office3	τ mean(std)	47.40(1.82)	41.65(0.56)	60.60(5.37)
	σmean(std)	0.05 (0.005)	0.01(0.001)	0.01(0.001)
Atrium	τ mean(std)	**37.75(2.77)**	**7.56(0.45)**	**49.38(15.24)**
	σmean(std)	0.01(0.001)	0.01(0.002)	0.12(0.02)
Restroom	τ mean(std)	133.48(1.01)	106.28(4.41)	135.05(4.38)
	σmean(std)	0.04(0.003)	0.01(0.001)	0.2(0.016)
Hallway	τ mean(std)	40.18(0.75)	22.33(1.22)	40.77(2.14)
	σmean(std)	0.03(0.001)	0.01(0.002)	0.36(0.048)
Outdoors1	τ mean(std)	12.93(1.17)	8.58(0.92)	14.36(1.74)
	σmean(std)	0.04(0.006)	0.01(0.001)	0.15(0.019)
Outdoors2	τ mean(std)	23.44(0.89)	8.81(0.54)	17.83 (0.92)
	σmean(std)	0.04(0.006)	0.01(0.001)	0.25(0.031)
Large office	τ mean(std)	35.03(0.72)	19.84(2.12)	46.0 (1.11)
	σmean(std)	0.04(0.002)	0.01(0.002)	0.17(0.018)
Stairs	τ mean(std)	**72.82(1.74)**	**51.73(2.42)**	**116.11 (7.51)**
	σmean(std)	0.04(0.003)	0.01(0.001)	0.25(0.031)

Bold: Observations for these two environments exhibit relatively large variance.

into four subband signals with equal frequency bands, that is, $sb_1 : 0 \leq f_{sb1} \leq 2$ kHz, $sb_2 : 2001 < f_{sb2} \leq 4$ kHz, $sb_3 : 4001 < f_{sb3} \leq 6$ kHz, and $sb_4 : 8001 < f_{sb4} \leq 8$ kHz, using wavelet packet decomposition. Reverberation parameters are then estimated from each subband signal using method discussed in Section 'Parameter estimation using maximum likelihood estimation'. Estimated parameters from recordings made in environments $E1$, $E5$, and $E7$ with all three microphones are shown in Figure 10.

Following observations can be made from Figure 10:

1. *Irrespective of the microphone type or acoustic environment*, the μ_{σ^2} decreases for higher subbands, i.e., sb_3 and sb_4.
2. *For all microphones and all selected acoustic environments*, the μ_{σ^2} for sb_1 and sb_2 (resp. sb_3 and sb_4) are relatively higher (resp. lower) than the μ_{σ^2} from original recordings.
3. *For all microphones and for moderately-to-highly reverberant environments* (e.g. $E5$ and $E1$), the μ_τ decreases for sb_3 and sb_4 and does not change for outdoors environment ($E7$). Whereas, for sb_1 and sb_2, the μ_τ do not change significantly (except Mic2 where μ_τ for sb_2 also decreases).
4. *For all microphones and all environments*, σ_τ for all subbands are larger than the σ_τ estimated from original recordings. This not a surprising observation as τ estimated from subband signals is using roughly one-fourth of the samples of the original recordings which can be translated into relatively less reliable estimates than the original recording.

Automatic AEI: *human speech recording*

The goal of the fourth experiment to evaluate performance of the proposed framework using speech recordings. To this end, second dataset consisting of 60 speech recordings of three speakers (a female and two male speakers) made in four acoustically different environments: (i) outdoors; (ii) a small office; (iii) stairs; and (iv) a restroom, with a commercial grade external microphone.

Acoustic reverberation parameters are estimated from manually selected decaying tails from each clean recording using method discussed in 'Parameter estimation using maximum likelihood estimation'. The DBSCAN-based clustering method is used for automatic AEI using estimated acoustic reverberation parameters, i.e., τ and σ. Shown in Figure 11 are the scatter plots of estimated reverberation parameters τ in msec. and σ with predicted environment labels for all speakers in all four acoustic environments.

It can be observed from Figure 11 that the proposed framework is capable of correctly predicting environment labels for speech recordings with very high accuracy. In addition, for each acoustic environment, the estimated τ exhibits relatively large spread compared with τ estimated from hand-clap recordings. Relatively large spread of the estimated τ for speech data can be attributed to the characterization of the speech signal and the decaying tail selection process. For example, in case of hand-clap recordings, decaying tail selection is very accurate as there is no overlapping from previous hand-clap instances, therefore, no interference, as a result reasonably consistent τ estimates from hand-clap recordings is expected. In case of speech recordings, on the other hand, for the voiced regions the previous phoneme utterance is likely to overlap with the following phoneme utterance, which causes interference in the selected decaying tails. Moreover, as the interference due to previous phoneme utterance is random in nature for real-world speech recordings. The τ estimated from decaying tails extracted from speech recordings is therefore expected to exhibit relatively larger spread than the hand-clap recordings, Figures 5, 6, 7, 8, 9, 10 and 11 support this argument.

Performance comparison with existing state-of-the-art

The aim of the final experiment to compare performance of the proposed framework with Hong's statistical learning-based method [18]. Speech recording dataset is used for this experiment.

For AEI using Hong's method, the reverberant component, $r(t)$, is extracted from each input speech signal using method discuss in the paper [18]. The resulting reverberant component is then pre-emphasized according to $r(t) = r(t) - p \times r(t-1)$ with $p = 0.97$. The estimated reverberant signal $r(t)$ is then decomposed into overlapping frames of length 25 ms with a frame shift of 10 ms, which resulted in 150 segments for each environment and a total of 600 segments for all four environments. For each segment, a Hamming window based 512-point DFT is computed, which is used to compute a 24-dimensional melspec coefficient vector. A 24-D *logarithmic melspec coefficient* (LMSC) vector is obtained by calculating the natural logarithm of the melspec coefficient vector; and a 24-D *mel-frequency cepstral coefficients* (MFCC) vector is obtained by computing DCT of the LMSC vector. For each segment, a 48-D feature vector is obtained by concatenating 24-D MFCC and 24-D LMSC vectors. The final 48-dimensional feature vector, averaged it over all frames is used for training and testing of the support vector machines (SVM) classifier.

For classification, a multi-class SVM trained with *radial basis kernel function* is used. The SVM tool downloaded from [41] was used for training and testing. To begin with, we randomly selected 50% of recordings from each category for training. The rest 50% are used to verify performance the proposed scheme. The optimal parameters for the classifier are determined using grid search

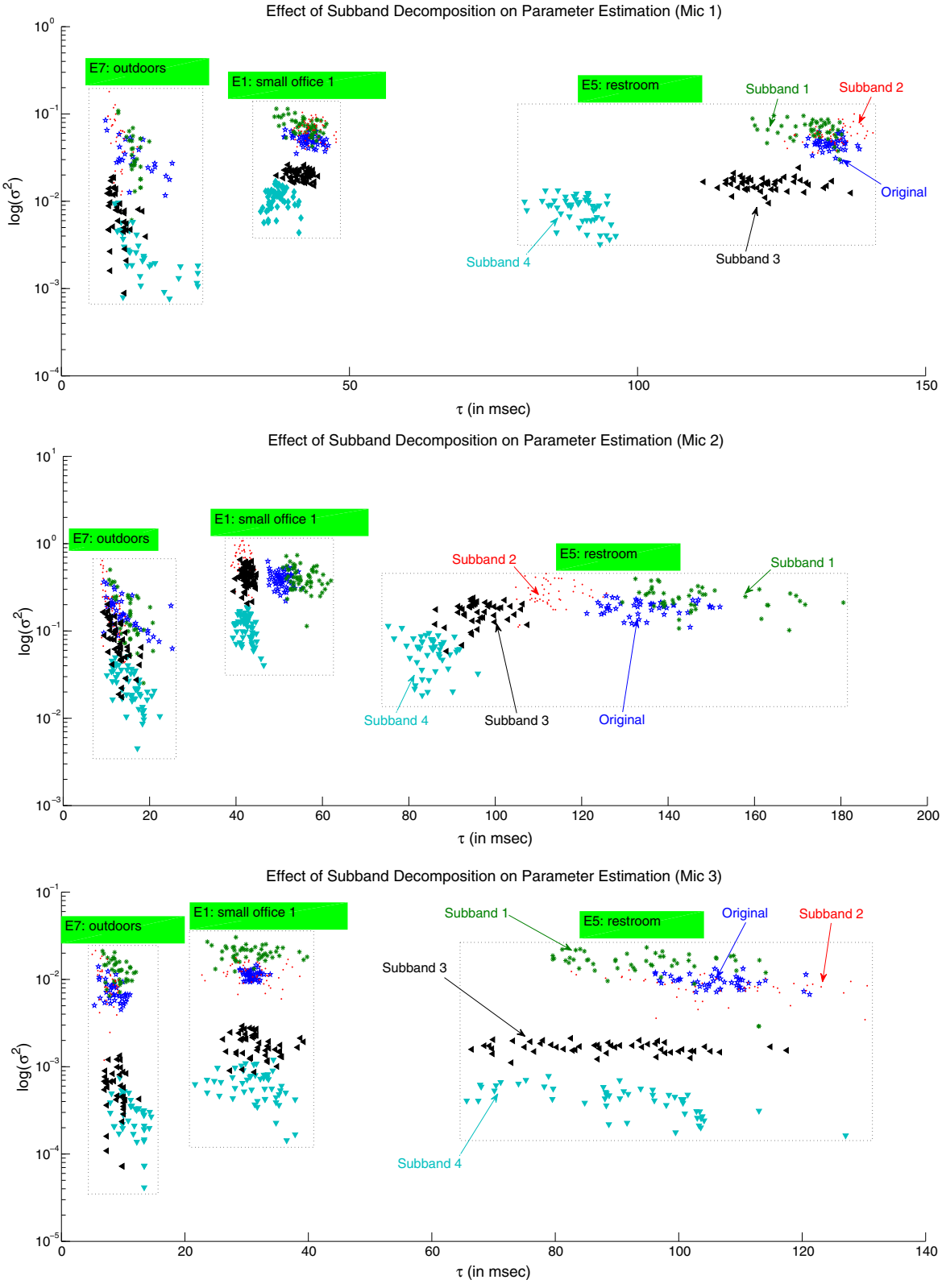

Figure 10 Shown in the top panel are the scatter plots of the estimated acoustic parameters from recordings made with Mic1 in *E1*, *E5* and *E7*; shown in the middle panel are the scatter plots of the estimated parameters from recordings made with Mic2 in the selected environments; and, shown in the bottom panel are the scatter plots of the estimated parameters from recordings made with Mic3.

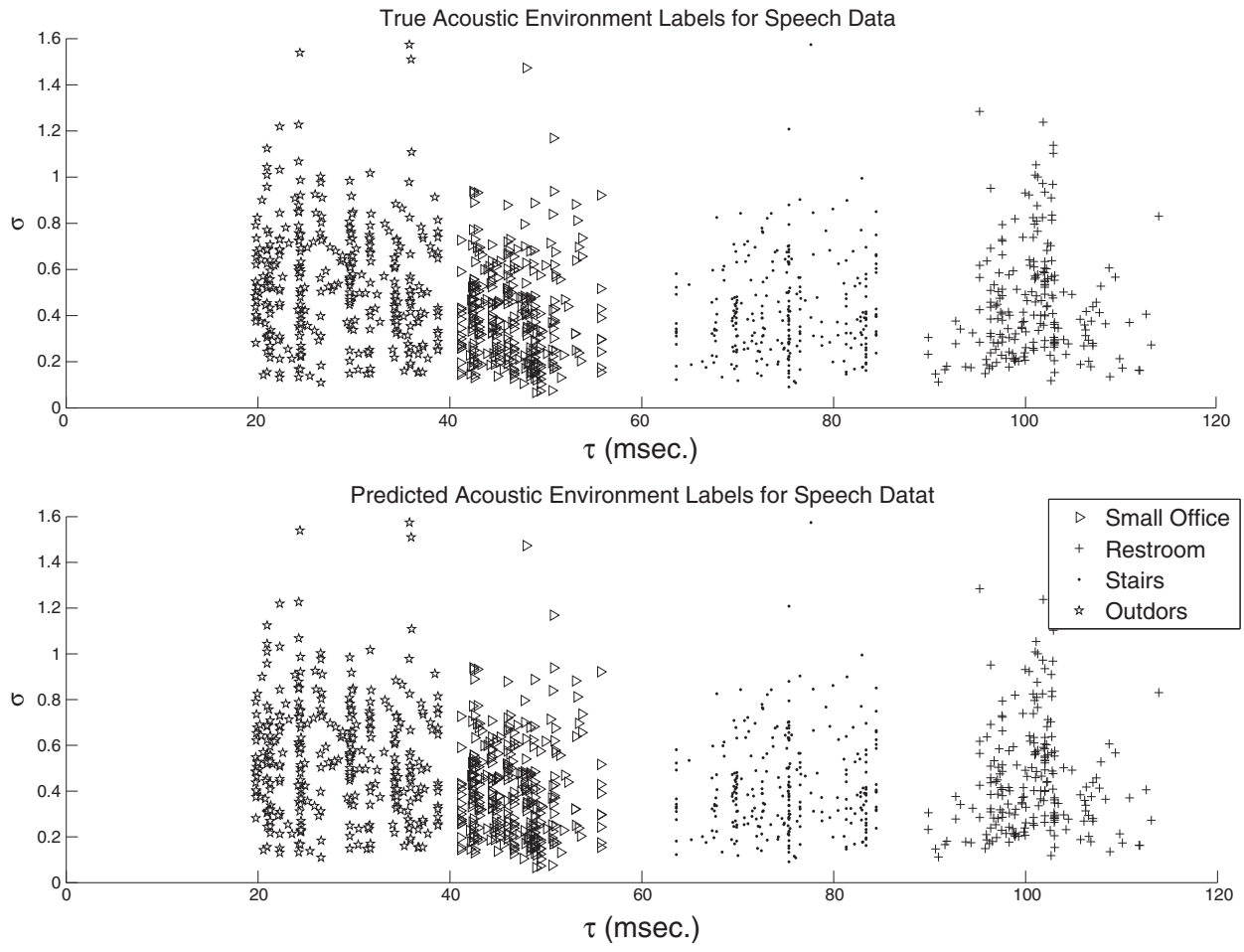

Figure 11 **Shown in the top panel are the scatter plots of the estimated parameters from speech recordings with true acoustic environment labels and shown in the bottom panel are the scatter plots of the estimated parameters with predicted environment labels using DBSCAN-based clustering.**

technique with five-fold cross-validation on training data. Shown in the Table 3 are the acoustic environment classification performance for Hong's method using speech dataset.

Shown in the Table 4 are the acoustic environment classification performance for the proposed method.

It can be observed from Table 3 Hong's learning-based method [18] achieves an average AEI accuracy around 94%; whereas the proposed scheme achieves perfect AEI accuracy, i.e., 100%, for the same dataset. This comparison indicates that the proposed scheme performs relatively better than the selected Hong's method. It is important to mention that the AEI results shown in Table 4 are obtained using manually selected decaying tails from speech recordings, whereas Hong's method does not require any user input for AEI. We have also observed

Table 3 Acoustic environment classification performance of the Hong's [18] scheme

True class	Predicted class label			
Label	*Outdoors*	*Small of fice*	*Stairs*	*Restroom*
Outdoors	88%	12%	0%	0%
Small of fice	10%	90%	0%	0%
Stairs	0%	0%	98%	2%
Restroom	0%	0%	1%	99%

Table 4 Classification performance of the proposed scheme

True class	Predicted class label			
Label	*Outdoors*	*Small of fice*	*Stairs*	*Restroom*
Outdoors	100%	0%	0%	0%
Small of fice	0%	100%	0%	0%
Stairs	0%	0%	100%	0%
Restroom	0%	0%	0%	100%

that when decaying tails are automatically selected using automatic tail selection method discussed in [42], AEI performance of the proposed method deteriorated around $< 3\%$, which is still better than the Hong's method.

Conclusion

The acoustic environment identification (AEI) has a wide range of applications ranging from audio recording integrity authentication to real-time crime acoustic space localization/identification. For instance, consider a scenario where a police call center receives an emergency call from a victim being harassed or chased by an offender. Under such crime situations it is very common that the harassed persons are unable to provide any relevant information about their actual location. The acoustic signals in the audio recording can be used to determine the acoustic space (i.e. car, street, neighborhood, living room, bath room, bed room, kitchen, etc.) of the crime scene. Similarly, for gun shooting cases, the sound of the firearms in the recording can be used to obtain important information about the crime scene such as weapon type.

In this paper we proposed a statistical framework for automatic recording environment identification (AEI) using acoustic signature of an audio recording. Late reverberant tail is modeled using an exponentially damped uncorrelated noise sequence obeying Gaussian distribution, which is then used for acoustic signature estimation using maximum likelihood estimation framework. Similarity measure based on Euclidian distance is used to classify estimated reverberation parameters for AEI. Density-based clustering method DBSCAN is used for automatic AEI. Performance of the proposed method is evaluated using two datasets consisting of (i) hand-clap recordings and (ii) speech recordings. The audio recordings used for performance evaluation were collected in a diverse set of acoustic environments using commercial grade external and built-in microphones. Simulation results indicate that the proposed framework is efficient for most of the considered acoustic environments. We have also shown that accuracy and reliability of the proposed AEI depends on the microphone type (used to capture audio recording). Sensitivity of the proposed method to various frequency bands has also been evaluated. Performance comparison with Hong's statistical learning based method [18] indicates that the proposed method achieves relatively higher accuracy. We expect this approach to be a useful forensic tool when used in conjunction with other techniques that measure microphone characteristics, background noise, and compression artifacts.

Acknowledgments

This work was supported by the NPST program by the King Saud University under grant number 12-INF2634-02 and a grant from the National Science Foundation (CNS-1440929).

Author details

[1]Electrical and Computer Engineering Department, University of Michigan - Dearborn, Dearborn, MI 48128, USA. [2]Department of Electronics, Quaid-i-Azam University, Islamabad 45320, Pakistan.

Received: 2 April 2014 Accepted: 1 August 2014
Published online: 02 September 2014

References

1. Cellphone popcorn viral videos (2008). Avaliable on: http://www.wired.com/underwire/2008/06/cellphones-cant/
2. Iran 'Modifies' pictures of missile test (2008). Avaliable on: http://www.switched.com/2008/07/11/iran-photoshops-pictures-of-missile-test/
3. H Farid, A survey of image forgery detection. IEEE Signal Process. Mag. **2**(26), 16–25 (2009)
4. S Gupta, S Cho, CC Kuo, Current developments and future trends in audio authentication. IEEE Multimedia. **19**, 50–59 (2012)
5. C Grigoras, A Cooper, M Michalek, Forensic speech and audio analysis working group - best practice guidelines for ENF analysis in forensic authentication of digital evidence, in *European Network of Forensic Science Institutes*, (2009). http://www.cs.dartmouth.edu/~farid/dfd/index.php/publications/show/103
6. D Nicolalde, J Apolinario, Evaluating digital audio authenticity with spectral distances and ENF phase change, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'09)* (Taipei, Taiwan, 2009), pp. 1417–1420
7. C Grigoras, Applications of ENF criterion in forensic audio, video, computer and telecommunication analysis. Forensic Sci Int. **167**, 136–145 (2007)
8. C Grigoras, Application of ENF analysis method in authentication of digital audio and video recordings, in *Proceedings of the 123rd Convention of the Audio Engineering Society,* (New York, NY, 2007), p. 7273
9. H Hollien, *The Acoustics of Crime, The New Science of Forensic Phonetics*. (Plenum Publishing Corporation, New York, NY, ISBN-13: 978-0306434679, 1990)
10. H Hollien, *Forensic Voice Identification*. (Academic Press, Philadelphia, PA, ISBN-13: 978-0123526212, 2001)
11. D Garcia-Romero, C Espy-Wilson, Speech forensics: automatic acquisition device identification. J. Acoust. Soc. Am. **127**(3), 2044–2044 (2010)
12. D Garcia-Romero, C Espy-Wilson, Automatic acquisition device identification from speech recordings, in *Proceedings of the IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP'10)* (Dallas TX, 2010), pp. 1806–1809
13. D Garcia-Romero, C Espy-Wilson, Automatic acquisition device identification from speech recordings. J. Acoustic Soc. Am. **124**(4), 2530–2530 (2009)
14. C Kraetzer, M Schott, J Dittmann, Unweighted fusion in microphone forensics using a decision tree and linear logistic regression models, in *Proceedings of the 11th ACM Multimedia and Security Workshop* (Princeton, NJ, 2009), pp. 49–56
15. C Kraetzer, A Oermann, J Dittmann, A Lang, Digital audio forensics: a first practical evaluation on microphone and environment classification, in *Proceedings of the 9th workshop on Multimedia and Security* (Dallas TX, 2007), pp. 63–74
16. A Oermann, A Lang, J Dittmann, Verifier-tuple for audio-forensic to determine speaker environment, in *Proceedings of the ACM Multimedia and Security Workshop 2005,* (New York, NY, 2005), pp. 57–62
17. R Buchholz, C Kraetzer, J Dittmann, Microphone classification using fourier coefficients, in *proceedings of 11th International Workshop on Information Hiding, Lecture Notes in Computer Science, Springer Berlin/Heidelberg Volume 5806/2009,* (Darmstadt, Germany, 2010), pp. 235–246
18. H Zhao, H Malik, Audio forensics using acoustic environment traces, in *Proceedings of the IEEE Statistical Signal Processing Workshop (SSP'12)* (MI Ann Arbor, 2012), pp. 373–376
19. H Malik, H Farid, Audio forensics from acoustic reverberation, in *Proceedings of the IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP'10)* (Dallas, TX, 2010), pp. 1710–1713
20. S Ikram, H Malik, Digital audio forensics using background noise, in *Proceedings of IEEE Int. Conf. on Multimedia and Expo 2010* (Singapore, 2010), pp. 106–110

21. H Malik, H Zhao, Recording environment identification using acoustic reverberation, in *Proceedings of the IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP'12)* (Kyoto, Japan, 2012), pp. 1833–1836
22. U Chaudhary, H Malik, Automatic recording environment identification using acoustic features, in *Proceedings of Audio Engineering Society 129th Convention 2010* (San Francisco, CA, 2010), p. 8254
23. H Malik, J Miller, Microphone identification using higher-order statistics, in *Proceedings of AES 46th Conference on Audio Forensics 2012* (Denver, CO, 2012), pp. 5–2
24. H Farid, *Detecting digital forgeries using bispectral analysis. Tech. rep., AIM-1657*. (Massachusetts Institute of Technology, 1999)
25. R Yang, Z Qu, J Huang, Detecting digital audio forgeries by checking frame offsets, in *Proceedings of the 10th ACM Workshop on Multimedia and Security (MM & Sec)'08,* (Oxford, UK, 2008), pp. 21–26
26. R Yang, Y Shi, J Huang, Defeating fake-quality, MP3, in *Proceedings of the 11th ACM Workshop on Multimedia and Security (MM & Sec)'09,* (Princeton, NJ, 2009), pp. 117–124
27. R Yang, Y Shi, J Huang, Detecting double compression of audio signal, in *Proceedings of SPIE Media Forensics and Security II 2010, Volume 7541,* (San Jose, CA, 2010)
28. A Cooper, Detecting butt-spliced edits in forensic digital audio recordings, in *Proceedings of Audio Engineering Society 39th Conf., Audio Forensics: Practices and Challenges,* (Hillerod, Denmark, 2010), pp. 1–1
29. S Hicsonmez, H Sencar, I Avcibas, Audio codec identification through payload sampling, in *IEEE Int. Workshop on Information Forensics and Security, (WIFS'11),* (Iguacu Falls, Brazil, 2011), pp. 1–6
30. C Grigoras, Statistical tools for multimedia forensics, in *Proceedings of Audio Engineering Society 39th Conf., Audio Forensics: Practices and Challenges*, (2010), pp. 27–32
31. H Zhao, H Malik, Audio recording location identification using acoustic environment signature. IEEE Trans. Inf. Forensics Secur. **8**(11), 1746–1759 (2013)
32. H Zhao, H Malik, Acoustic environment identification and its applications to audio forensics. IEEE Trans. Inf. Forensics Secur. **8**(11), 1746–1759 (2013)
33. R Ratnam, DL Jones, BC Wheeler, WD O'Brien Jr, CR Lansing, AS Feng, Blind estimation of reverberation time. J. Acoust. Soc. Am. **5**(114), 2877–2892 (2003)
34. H Lollmann, P Vary, Estimation of the reverberation time in noisy environments, in *Proceedings of International Workshop on Acoustic Echo and Noise Control* (Seattle, WA, 2008)
35. I Tashev, D Allred, Revereberation reduction for improved speech recognition, in *HSCMA* (Piscataway, NJ, 2005)
36. H Kuttruff, *Room Acoustics, 3rd edition*. (Elsevier, New York, 1991)
37. Y Lu, PC Loizou, A geometric approach to spectral subtraction. Speech Commun. **50**(6), 453–466 (2008)
38. M Schroeder, New method for measuring reverberation time. J. Acoust. Soc. Am. **3**(37), 409–412 (1965)
39. M Ester, HP Kriegel, J Sander, X Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in *Proceedings of Second International Conference on Knowledge Discovery and Data Mining (KDD-96),* (AAAI Press Portland, OR, 1996), pp. 226-231
40. PN Tan, M Steinback, V Kumar, *Introduction to Data Mining*. (Pearson Addison Wesley, Indianapolis, IN, ISBN-13: 978-0321321367, 2006)
41. C Chang, C Lin, Libsvm: a library for support vector machines (2012). [http://www.csie.ntu.edu.tw/cjin/libsvm]
42. H Malik, Audio recording location identification using acoustic environment signature. IEEE Trans. Inf. Forensics Secur. **8**(11), 1827–1837 (2013)

doi:10.1186/s13388-014-0011-7
Cite this article as: Malik and Mahmood: **Acoustic environment identification using unsupervised learning.** *Security Informatics* 2014 **3**:11.