Audio Recording Location Identification Using Acoustic Environment Signature

Hong Zhao and Hafiz Malik, *Member, IEEE*

***Abstract*—An audio recording is subject to a number of possible distortions and artifacts. Consider, for example, artifacts due to acoustic reverberation and background noise. The acoustic reverberation depends on the shape and the composition of the room, and it causes temporal and spectral smearing of the recorded sound. The background noise, on the other hand, depends on the secondary audio source activities present in the evidentiary recording. Extraction of acoustic cues from an audio recording is an important but challenging task. Temporal changes in the estimated reverberation and background noise can be used for dynamic *acoustic environment identification* (AEI), audio forensics, and ballistic settings. We describe a statistical technique based on spectral subtraction to estimate the amount of reverberation and nonlinear filtering based on particle filtering to estimate the background noise. The effectiveness of the proposed method is tested using a data set consisting of speech recordings of two human speakers (one male and one female) made in eight acoustic environments using four commercial grade microphones. Performance of the proposed method is evaluated for various experimental settings such as microphone independent, semi- and full-blind AEI, and robustness to MP3 compression. Performance of the proposed framework is also evaluated using Temporal Derivative-based Spectrum and Mel-Cepstrum (TDSM)-based features. Experimental results show that the proposed method improves AEI performance compared with the direct method (i.e., feature vector is extracted from the audio recording directly). In addition, experimental results also show that the proposed scheme is robust to MP3 compression attack.**

***Index Terms*—Acoustic environment identification, acoustic reverberation, audio forensics, background noise, particle filtering.**

I. Introduction

THE use of digital media (audio, video, and images) as evidence in litigation and criminal justice is increasingly common. For digital media to be admitted as evidence in a court of law, its authenticity and integrity must be verified. This requirement is a complex and challenging task, especially without *helping data*, such as *digital watermarks* or *fingerprints*, and if the media is only available in a compressed format. The availability of powerful, sophisticated, and easy-to-use digital media manipulation tools has made authenticating the integrity of digital media even more difficult. In this context, digital media forensics aims to determine the underlying facts about an evidentiary recording and to provide authoritative answers (in the absence of helping data) to various questions, such as:

- 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 captured using acquisition device X at location L, as claimed?
- Is the auditory scene in the evidentiary recording original or was it digitally altered to deceive the listener?

It is therefore critical to authenticate the integrity of digital evidence. Digital audio forensic techniques have been developed to detect traces of forgeries and tempering by exploiting:

- inconsistencies in the *electric network frequency* (ENF) [1]–[6],
- acquisition device nonlinearities [7]–[15],
- artifacts due to acoustic reverberation [10]–[12], [16]–[20],
- inconsistencies due to spectral distance and phase shift [21] gunshot characterization[22], [23], and
- inconsistencies due to lossy compression [24]–[27] and '*butt-splicing*' [28].

The acoustic environment identification (AEI) has a wide range of applications ranging from audio recording integrity authentication to real-time acoustic space localization/identification. For instance, consider a scenario where an audio recording presented in the court as evidence claiming that the recording was made in the claimed environment, e.g., office, hallway, outdoors, etc. For the admissibility of the evidence in a court of law, integrity authentication of the evidentiary recording is required. Temporal consistency of estimated acoustic signatures (e.g., reverberation and background noise) from the evidentiary recording can be used for integrity authentication. Similarly, 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 and reverberations in the test audio recording can be used to

Manuscript received June 09, 2013; revised August 03, 2013; accepted August 03, 2013. Date of publication August 16, 2013; date of current version September 26, 2013. This work was supported in part by the National Natural Science Foundation of China under Grant 61170226, in part by the Fundamental Research Funds for the Central Universities under Grants SWJTU11CX047 and SWJTU12ZT02, in part by the Young Innovative Research Team of Sichuan Province under Grant 2011JTD0007, and in part by Chengdu Science and Technology program under Grant 12DXYB214JH-002. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jiwu Huang.

H. Zhao is with the School of Science and Technology, Southwest Jiaotong University, Jiaotong, 610031, China, and also with the Department of Electronic and Electrical Engineering, South University of Science and Technology, Shenzhen, 518055, China (e-mail: zh1985444@gmail.com).

H. Malik is with the Department of Electrical and Computer Engineering, The University of Michigan, Dearborn, MI 48128 USA (e-mail: hafiz@umich.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2013.2278843

determine the acoustic space (i.e., car, street, neighborhood, living room, bath room, bed room, kitchen, etc.) of the crime scene.

This paper presents a new approach using a detailed analysis of audio data to provide evidence in terms of features characterizing the location where the recording was made. Motivation behind considering acoustic artifacts for audio forensics and AEI is that existing audio forensic analysis methods, e.g., ENF-based methods [3], [21], [29], [30] and recording device identification based methods [7]–[10] cannot withstand *lossy compress attack*, e.g., MP3 compression. In our recent work [20] we have shown that acoustic reverberations can survive the lossy compression attack. Location specific acoustic features therefore can be used for AEI and digital audio forensic applications.

The main goal of this paper is to develop a statistical framework for automatic AEI and its applications to digital audio forensics. Here we exploited specific artifacts introduced at the time of recording for the AEI and for audio recording integrity authentication. Both the acoustic reverberation and the background (ambient) noise are considered to achieve this objective. Audio reverberation is caused by the persistence of sound after the source has terminated which 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 amount of literature on modeling and estimating audio reverberation (see, for example, [31]). Inverse filtering using spectral subtraction based methods is considered to estimate acoustic reverberation. The blind reverberation estimation approach method used here is a variant of that described in [32]. The background noise is modeled using a dynamical system and estimated using nonlinear filtering based on particle filtering. An 128-dimensional feature vector is extracted from the estimated background noise and acoustic reverberation components. A multiclass Support Vector Machine (SVM) classifier is trained using the data set recorded in eight acoustic environments and tested for the AEI. The performance of the proposed scheme is tested using a data set consisting of 2240 audio recordings made in eight acoustic environments using four microphones. Experimental results show that the proposed system can successfully identify a recording environment for both the compressed and uncompressed audio recordings.

The rest of the paper is organized as follows: a brief overview of the existing state of art audio forensics is provided in Section II; details of background noise modeling and estimation are given in Section III-A. Reverberation acoustic environment artifacts modeling is outlined in Section III. Block-based inverse filtering for blind dereverberation is provided in Section III-B. Details of feature extraction for the estimated reverberation and background noise components are provided in Section III-C. Experimental setup, results, and performance analysis are provided in Section IV. Finally, the concluding remarks along with future research directions are discussed in Section V.

II. Audio Forensics: State of the Art

Forensic experts have been analyzing audio recordings since the 1960s. For example, the U.S. Federal Bureau of Investigation has conducted an examination of audio recordings for speech intelligibility enhancement and authentication [33]. In the United States, US vs. McKeever (169 F.Supp. 426, 430, S.D.N.Y. 1958) established a set of requirements for the admissibility of audio recordings in a court of law.

Audio forensics has traditionally focused on analog magnetic tape recordings by relying on analog recorder fingerprints, such as *head switching transients*, *mechanical splices*, and *overdubbing signatures*, to determine the integrity of the recording [34]–[38]. The question of authenticity becomes more complicated and challenging for digital recordings because digital recorders do not leave such traces in the recording. Therefore, linking a recorder to the recording, detecting copies, and determining the chronology of recorded events is difficult to determine for digital recordings.

Over the last few decades, several efforts have been initiated to fill the rapidly growing gap between digital media manipulation technologies and digital media authentication tools. For example, recent efforts have focused on residual signals (i.e., *electric network frequency (ENF)*, which has power-line frequency 60/50 Hz) may be present in frequency to the digital recording system [1]–[6] to authenticate the digital recording. The ENF-based methods use the random fluctuations in the power-line frequency caused by mismatches between the electrical system load and generation for authentication purposes. The ENF-based approaches may not always be applicable if well-designed audio equipment (e.g., precessional microphones) or battery-operated devices (e.g., smartphones) are used to capture the recordings.

Statistical pattern recognition-based techniques have been proposed for identifying recording locations [11]–[13], [39]–[44] and acquisition devices [7]–[10], [45]–[47]. However, these methods are limited by their low accuracy and the inability to link a recording to an acquisition device in a unique manner. Additionally, these techniques work only in the raw digital domain. We have also developed model-driven approaches to estimate acoustic reverberation signatures for automatic acoustic environment identification and forgery detection [14]–[20].

Techniques based on time-domain analysis [25]–[27] have been proposed to determine the authenticity of MP3 audio files against editing and double compression attacks. Similarly, a framework based on frequency-domain statistical analysis has also been proposed by Grigoras [24] to detect traces of audio (re)compression and to discriminate among different audio compression algorithms. Similarly, Liu *et al.* [48] have also proposed a statistical learning based method to detect traces of double compression. The performance of Liu *et al.*'s method deteriorates for low to high bit rate transcoding. Qiao *et al.* [49] addresses this issue by considering nonzero dequantized Modified Discrete Cosine Transform (MDCT) coefficients for their statistical machine learning method. Brixen in [50] has proposed a time-domain method based on acoustic reverberation estimated from digital audio recordings of mobile phone calls for crime scene identification. A method based on higher-order time-differences and correlation analysis has been proposed by Cooper [28] to detect traces of "*butt-splicing*" in digital recordings. Recently, Pan *et al.* [51] have also proposed a time-domain method based on higher-order statistics to detect traces of splicing. The proposed method uses differences in

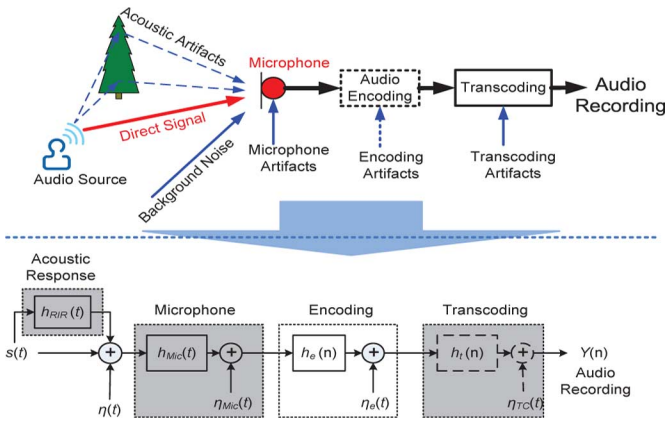

Fig. 1. Simplified digital audio recording diagram.

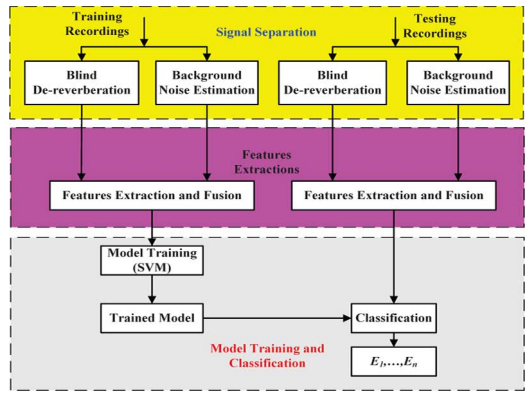

Fig. 2. Conceptual information flow of the proposed acoustic environment identification system.

the local noise levels in an audio signal for splice detection. Similarly, Zhou *et al.* in [52] proposed to use the noise patterns of different codecs to identify the speech codec used and tested the effectiveness of their method on commonly-used audio codecs such as GSM-HR, GSM-EFR, GSM-AMR, and SILK.

III. Methodology

Consider a digital audio recording signal $y(t)$, which is a combination of several components such as, direct speech signal $s(t)$, acoustic environment distortion (consisting of *reverberant signal* $r(t)$ and *background noise* $\eta(t)$), microphone distortion $\eta_{Mic}(t)$, encoding distortion $\eta_e(t)$, and transcoding distortion $\eta_{TC}(t)$. A simplified model for digital audio recording is shown in Fig. 1.

The combined effect of the direct and reflected signals and the background (or ambient) noise at the input of the microphone can be expressed as,

$$y_1(t) = s(t) + h_{RIR}(t) * s(t) + \eta(t) \tag{1}$$

where h_{RIR} denotes the room impulse response (RIR). If h_{Mic} denotes the microphone impulse response, then the digital audio recording, $y(t)$, can be expressed as,

$$y(t) = h_{final}(t)*s(t)+h_{Mic}(t)*\eta(t)+\eta_{Mic}(t)+\eta_{TC}(t) \tag{2}$$

where $h_{final}(t)$ can be expressed as,

$$h_{final}(t) = h_{RIR}(t) * h_{Mic}(t). \tag{3}$$

Extraction of the dry signal, $s(t)$, from the observation, $y(t)$, without the knowledge of $h_{final}(t)$, $\eta(t)$, $\eta_{Mic}(t)$, and $\eta_{TC}(t)$, is a challenging task. To simplify the complicated model, we make the following necessary assumptions:

- The impulse response of microphone $h_{Mic}(t)$ is time-invariant or at least short-time-invariant. Under this assumption, the estimation of $h_{RIR}(t)$ is possible, without any prior knowledge of $h_{Mic}(t)$.
- The recording system introduces negligible microphone distortion, i.e., $\eta_{Mic}(t) \approx 0$, and transcoding distortion $\eta_{TC}(t) \approx 0$. These are reasonable assumptions especially when a noise-canceling microphone is used for audio recording and when the recorded audio is saved in raw format.
- The background noise is not very strong. This is a reasonable assumption for non-Gaussian background noise. Removing strong non-Gaussian noise from the audio signal is a challenging task.
- Audio recording is captured in a stationary acoustic environment. It is important to note that an acoustic space does not have a stationary impulse response when either the sound source or the microphone is moving (even by a small distance). Theoretically speaking, for any given room there are effectively an infinite number of possible impulse responses, as there are effectively an infinite number of possible combinations of locations of sound source and microphone pairs. Estimating the reverberant component for nonstationary acoustic environments is a challenging task. It is therefore reasonable to assume that the recordings are made with a fixed set of microphone and sound source locations in each acoustic environment.

With the above assumptions, the observation can be expressed as,

$$y(t) \approx x(t) + \eta(t)$$

where

$$x(t) = h_{final}(t) * s(t) + s(t). \tag{4}$$

In the following section, we will describe how to estimate the reverberation signal $r(t) = h_{RIR}(t) * s(t)$ and the background noise $\eta(t)$ without exploiting any prior knowledge of the acoustic environment.

The proposed system can be divided into three subsystems:

1) The background noise estimation subsystem,
2) The blind dereverberation subsystem, and
3) The feature extraction, fusion, and classification subsystem.

Shown in Fig. 2 is the conceptual information flow of the proposed acoustic environment identification system. Details of each processing block are provided in the following sections.

A. Background Noise Estimation

The background (or ambient) noise provides very useful acoustic cues that can be used for acoustic environment identification. Various noise estimation methods have been proposed [53], [54] in the past. These methods work well for stationary

and synthetic noise. Most of the real-world background noise, however, is dynamic (nonstationary) in nature. Therefore, we modeled background noise using a dynamical system. To estimate the noise as accurately as possible, a particle filter-based sequential estimation method has attracted much attention. During past two decades, various *particle filter* (PF)-based approaches have been proposed to track nonstationary additive distortions on speech features in the log-power frequency domain [55]–[60]. It has been demonstrated that the PF based speech feature enhancement technique will improve the recognition accuracy. The PF based noise estimation can be formulated as a tracking problem where the noise feature have to be estimated for each frame, given the current observation and its history of the noisy features. In this section, particle filter based noise estimation will be introduced.

1) Dynamical System for Noise Estimation: The short-time discrete-time representation of the observation signal model described in (4) can be expressed as,

$$\begin{aligned} y[n,k] &= x[n,k] + \eta[n,k] \\ &= s[n,k] * h_{final}[n] + \eta[n,k] \end{aligned} \tag{5}$$

where positive integer k denotes the index of the time-frame.

The corresponding observation signal in the Mel-frequency domain can be expressed as follow. Let vectors $\mathbf{Y}_k$, $\mathbf{X}_k$ and Υ_k denote the logarithmic output energy of the Mel-filter bank [60] of the k^{th} time-frame of observation $y[n,k]$, clean speech $s[n,k]$ and noise $\eta[n,k]$, respectively. The dynamical system for noise sequence can be represented as [57], [60],

$$\mathbf{Y}_k = \mathbf{X}_k + log(\mathbf{I} + exp(\boldsymbol{\Upsilon}_k - \mathbf{X}_k)) + \mathbf{V}_k \tag{6}$$

$$= \mathfrak{F}(\mathbf{X}_k, \boldsymbol{\Upsilon}_k) + \mathbf{V}_k \tag{7}$$

where,

$$\mathbf{V}_k \sim \mathcal{N}(\mathbf{0}, \Sigma_v) \tag{8}$$

represents the process noise in the Mel-frequency domain, Σ_v denotes the diagonal covariance matrix of the Gaussian distribution, and $\mathbf{I}$ represents the identity matrix with the same dimension as $\mathbf{V}_k$.

The Gaussian Mixture Model (GMM) [61] is used to model the clean speech feature vector $\mathbf{X}$ in the Mel-frequency domain, that is, the clean speech features $\mathbf{X}_k$ at time index k is generated by a Gaussian distribution trained using clean speech. In addition to the observation equation, the state transition equation of noise is the most important factor for the state-space model based noise signal estimation. Several state transition models, e.g., random walk, random walk by Polyak averaging, and feedback, predicted walk by static AR processes, etc., have been proposed to model nonstationary background noise [59]. Random walk, the simplest next state predictor, is commonly used to predict the next state of a dynamical system. It simply takes the previous state as the estimate of the current state and adds a random variable W_k, which is considered to be independent and an identically distributed (iid) Gaussian random variable with zero mean and covariance Σ_w, i.e., $\mathbf{W}_k \sim \mathcal{N}(\mathbf{0}, \Sigma_w)$. The random walk-based prediction of the next state is give as,

$$\boldsymbol{\Upsilon}_k = \boldsymbol{\Upsilon}_{k-1} + \mathbf{W}_k. \tag{9}$$

The random walk process is the simplest model used for tracking the state transition of a dynamical system. Therefore, it is unable to accurately model the state transition of nonstationary real-world noise processes. To this end, random walk by *Polyak averaging* and feedback based state transition model is used to model the nonstationarity in the background noise. The motivation behind Polyak averaging is to limit the range of the predicted noise hypothesis within a fixed interval of the preceding frames. The weighted average of noise sample for the next prediction is calculated over the particles is,

$$\widetilde{\boldsymbol{\Upsilon}}_{k-1} = \sum_{j=1}^{J} w_k^j \boldsymbol{\Upsilon}_{k-1}^j. \tag{10}$$

The polyak averaging based state transition can be expressed as,

$$\boldsymbol{\Upsilon}_k^j = (1-\alpha)\boldsymbol{\Upsilon}_{k-1}^j + \alpha\widetilde{\boldsymbol{\Upsilon}}_{k-1} + \alpha\beta(\mu_k^j - \boldsymbol{\Upsilon}_{k-1}^j) + \mathbf{W}_k^j \tag{11}$$

where $\alpha \in \mathcal{R}$ represents forgetting factor, $\beta \in \mathcal{R}$ represents scaling factor of feedback, and μ_k^j represents the Polyak average of preceding T frames given as,

$$\mu_k^j = \frac{1}{T} \sum_{i=k-T+1}^{k} \boldsymbol{\Upsilon}_i^j. \tag{12}$$

It is important to mention that the first and the second terms in (11) shift noise sample $\widetilde{\boldsymbol{\Upsilon}}_{k-1}$ close to the weighted average using a forgetting factor which helps to reduce the scattering of samples and removes outliers in the predicted state. The third term, Polyak averaging and feedback, controls the compensation range for the predicted state parameters.

The Polyak averaging and feedback based state-transition equation can estimate the noise samples more accurately than the random walk based state-transition equation because it predicts the next frame parameters depending on the previous frames.

2) Sequential Importance Sampling (SIS) for PF: Given the dynamical system characterized by (7) and (11), the joint a *posteriori* probability density function (pdf) for the noise sequence, Υ_k, can be represented by a first-order Markov chain, i.e.,

$$p(\boldsymbol{\Upsilon}_{0:k}|\mathbf{Y}_{0:k}) = p(\boldsymbol{\Upsilon}_0|\mathbf{Y}_0) \prod_{j=1}^{k} p(\boldsymbol{\Upsilon}_j|\boldsymbol{\Upsilon}_{j-1}) p(\mathbf{Y}_j|\boldsymbol{\Upsilon}_j) \tag{13}$$

where $\boldsymbol{\Upsilon}_{0:k} = \{\boldsymbol{\Upsilon}_0, \boldsymbol{\Upsilon}_1, \ldots, \boldsymbol{\Upsilon}_k\}$ and $\mathbf{Y}_{0:k} = \{\mathbf{Y}_0, \mathbf{Y}_1, \ldots, \mathbf{Y}_k\}$.

The $p(\boldsymbol{\Upsilon}_{0:k}|\mathbf{Y}_{0:k})$ in (13) can be obtained recursively from $p(\boldsymbol{\Upsilon}_{0:k-1}|\mathbf{Y}_{0:k-1})$, i.e.,

$$p(\boldsymbol{\Upsilon}_{0:k}|\mathbf{Y}_{0:k}) = \frac{p(Y_k|\boldsymbol{\Upsilon}_k) p(\boldsymbol{\Upsilon}_k|\boldsymbol{\Upsilon}_{k-1})}{p(Y_k|\mathbf{Y}_{0:k-1})} p(\mathbf{Y}_{0:k-1}|\boldsymbol{\Upsilon}_{0:k-1}). \tag{14}$$

Since the derivation from $p(\boldsymbol{\Upsilon}_{0:k-1}|\mathbf{Y}_{0:k-1})$ to $p(\boldsymbol{\Upsilon}_{0:k}|\mathbf{Y}_{0:k})$ is analytically intractable. To get around this problem, Monte Carlo sampling is used to approximate the joint a *posteriori*. The $p(\boldsymbol{\Upsilon}_k|\mathbf{Y}_k)$ approximated using Monte Carlo sampling is given as [60], [62],

$$p(\boldsymbol{\Upsilon}_k|\mathbf{Y}_k) \approx \sum_{j=1}^{J} \widetilde{w}_k^j \delta(\boldsymbol{\Upsilon}_k - \boldsymbol{\Upsilon}_k^j) \tag{15}$$

where $\widetilde{w}_k^j$ is the normalized importance weight and $\delta(\cdot)$ denotes a Dirac delta function.

It is important to mention that drawing samples directly from the posterior density is often infeasible, therefore, a suboptimal *importance density* is generally used [63]. When samples are drawn from the importance density $q(\mathbf{\Upsilon}_k|\mathbf{Y}_k)$ then the importance weights $\widetilde{w}_k^j$ can be defined as,

$$\widetilde{w}_k^j \propto \frac{p(\mathbf{\Upsilon}_k|\mathbf{Y}_k)}{q(\mathbf{\Upsilon}_k|\mathbf{Y}_k)}. \tag{16}$$

Now, let us approximate $p(\mathbf{\Upsilon}_k|\mathbf{Y}_k)$ as,

$$p(\mathbf{\Upsilon}_k, \mathbf{Y}_k|\mathbf{Y}_{1:k-1}) \approx \sum_{j=1}^{J} p(\mathbf{\Upsilon}_k^j|\mathbf{\Upsilon}_{k-1}^j) p(\mathbf{Y}_k|\mathbf{\Upsilon}_k^j) \tag{17}$$

and

$$p(\mathbf{Y}_k|\mathbf{Y}_{1:k-1}) \approx \sum_{j=1}^{J} p(\mathbf{Y}_k|\mathbf{\Upsilon}_k^j). \tag{18}$$

According to (15)–(18), the weights can be expressed as,

$$\widetilde{w}_k^j \propto p(\mathbf{Y}_k|\mathbf{\Upsilon}_k^j). \tag{19}$$

The weights are represented by the corresponding likelihood for each sample j out of J samples. Those samples are known as *particles* and the filter process is called particle filtering. It is also called a *sequential important sampling (SIS)* particle filter [64].

In practice, sampling step many results in *degeneracy condition*, that is, it might select a reasonable number of samples with insignificant weights. To address this issue, residual resampling [64] is used which discard samples with insignificant weights and maintain a constant number of samples. Accordingly, weights after the resampling step are also proportionally redistributed.

After resampling, the samples are distributed approximately according to (15). However, the discrete nature of the approximation can skew the importance weight distribution. To get around this problem, Metropolis-Hastings sampling is used. The Metropolis-Hastings sampling method [65] samples a candidate according to a proposed importance distribution given the current state.

3) Weights Calculation: The likelihood-based method [62] is used to calculate the weight of each sample $\mathbf{\Upsilon}_k^j$ given in (19), that is,

$$p(\mathbf{Y}_k|\mathbf{\Upsilon}_k^j) = \frac{p_{speech}(\mathbf{Y}_k) + log(\mathbf{I} - exp(\mathbf{\Upsilon}_k^j - \mathbf{Y}_k))}{\prod_{b=1}^{B} |\mathbf{I} - e^{\mathbf{\Upsilon}_k^j - \mathbf{Y}_k}|} \tag{20}$$

where $p_{speech}(\cdot)$ denotes the prior speech density represented by a GMM, which has been trained on clean speech. The model is represented in a B dimensional space, where each dimension b represents a frequency bin.

The weight of each sample is then normalized as,

$$\widetilde{w}_k^j = \frac{p(\mathbf{Y}_k|\mathbf{\Upsilon}_k^j)}{\sum_{j=1}^{J} p(\mathbf{Y}_k|\mathbf{\Upsilon}_k^j)}. \tag{21}$$

It can be observed from (21) that the normalized weight can be evaluated if the estimated noise is less than the observation in all spectral bins, that is,

$$\mathbf{\Upsilon}_k^j < \mathbf{Y}_k, \forall b \in B. \tag{22}$$

If this condition is not satisfied then the $p(\mathbf{Y}_k|\mathbf{\Upsilon}_k^j)$ cannot be evaluated. One simple solution to such situations is to set weights of those samples to zero. However, it may result in a complete annihilation especially when all the weights are set to zeros. To handle this problem, a *fast acceptance test* (FAT) [62] is used to prevent aggressive dropouts in the number of particles from the candidate list. The FAT accepts a drawn sample only if the likelihood can be evaluated. Otherwise, a new sample is drawn by propagating a randomly chosen particle.

4) Particles Initialization: The first step for particle filtering is to initialize the samples from a joint *prior noise density*, $p(\mathbf{\Upsilon}_{0:k}|\mathbf{Y}_{0:k})$. To achieve this goal, the $p(\mathbf{\Upsilon}_0)$ can be initialized as,

$$p(\mathbf{\Upsilon}_0) = \mathcal{N}(\mu_\Upsilon, \Sigma_\Upsilon). \tag{23}$$

Here is the variance terms Σ_Υ, Σ_v and Σ_w are initialized by setting each of them to a small real value. It is important to note that particle filtering performance does not depend on the initial value of Σ_Υ, Σ_v, and Σ_w but it does depend on the initial value of the mean term, e.g., μ_Υ. Therefore, the μ_Υ is initialized by setting it equal to the estimated mean of the nonvoiced portions of the audio recordings, which can be obtained through voice activity detection.

B. Blind Dereverberation

Dereverberation is the next stage in the process of acoustic environment identification (AEI). Blind dereverberation is the separation of the reverberant and dry signal from a single channel audio recording without exploiting knowledge of the room impulse response, $h_{RIR}(t)$. Dereverberation is a widely used method with application ranging from speech enhancement to distant speech recognition, acoustic environment identification, etc. It is important to note that speech recognition applications rely on the direct signal component, $s(t)$, whereas, the AEI applications rely on the reverberant component, $r(t)$; the reverberant component $r(t)$ embodies the characterization of the acoustic environment. The goal of dereverberation here is to estimate $r(t)$ from the enhanced (denoised) recording, $\tilde{y}(t)$ and use it for the AEI. The output of the denoising stage can be expressed as $\tilde{y}(t) \approx s(t) + r(t)$, where $s(t)$ represents direct sound component, also referred as the dry signal, and $r(t)$ which represents the reverberant component. Furthermore, the reverberant signal $r(t)$ can be expressed as,

$$r(t) = s(t) * h_{RIR}(t) \tag{24}$$

where $*$ represents the convolution operation and $h_{RIR}(t)$ represents acoustic environment (or room) impulse response.

Under the assumption of a stationary acoustic environment, the $h_{RIR}(t)$ can be modeled by a finite impulse response (FIR) filter, provided that filter is of sufficient length [32]. Shown in Fig. 3 is an example of a typical room impulse response. An

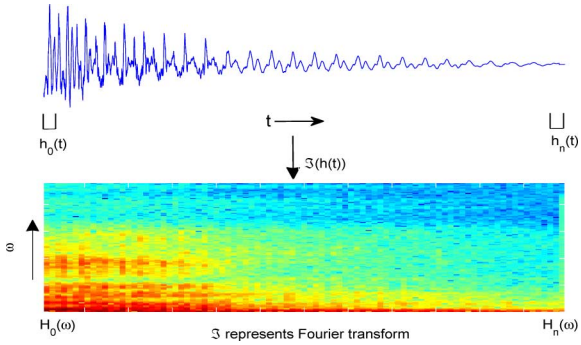

Fig. 3. Block-based representation of room impulse response (RIR).

impulse response can also be expressed in the frequency domain $H(\omega)$, as shown in Fig. 3. The Fourier representation of the impulse response provides us with both the magnitude response and the phase response. Changing the location of a sound source or the microphone does not have much effect on the magnitude response, whereas, it does have a pronounced effect on the phase response. Fortunately, the human auditory system is relatively insensitive to phase over a short period of time. So, instead of estimating the phase information, the phase of the reverberant input signal will be used to approximate the phase of the original dry signal. To this end, the magnitude of the reverberant signal, $|R(\omega)|$, is estimated using block-based processing of $\tilde{y}(t)$. The following section provides the details of estimating the $|R(\omega)|$ from $\tilde{y}(t)$.

It can be observed from Fig. 3 that the room impulse response, $h_{RIR}(t)$, can be divided into $K+1$ blocks consisting of $h_{RIR,\,i}(t) : i = 0, 1, \ldots, K$ (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 $\tilde{y}(t)$ can be expressed using block-based FIR filtering as,

$$\tilde{y}(t) = \sum_{i=0}^{K} s(t) * \delta(t - iL) * h_{RIR,\,i}(t). \tag{25}$$

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

$$r(t) = \sum_{i=1}^{K} s(t) * \delta(t - iL) * h_{RIR,\,i}(t). \tag{26}$$

Similarly, the observation signal $\tilde{y}(t)$ and reverberant component can be expressed in frequency domain as,

$$\widetilde{Y}(\omega) = S(\omega)H_{RIR,0}(\omega) + \sum_{i=1}^{K} S(\omega)Z^{-iL}H_{RIR,i}(\omega) \tag{27}$$

$$R(\omega) = \sum_{i=1}^{K} S(\omega)Z^{-iL}H_{RIR,i}(\omega). \tag{28}$$

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 $\tilde{y}(t)$ using such a filter. For example, if the FIR filter response ($h_{RIR}(t)$ or $H_{RIR}(w)$) is known then the reverberant component $r(t)$ can be estimated using (28). Under blind dereverberation however estimation of FIR filter response from a monophonic audio recording, $\tilde{y}(t)$ is an ill-posed problem. To overcome this problem, a *perceptually relevant* estimation of FIR filter block is used. The perceptually relevant estimates, $\widetilde{H}_{RIR,\,i}(\omega) : i = 0, 1, \ldots, K$, are estimated from magnitude response of the FIR filter blocks, i.e.,

$$\widetilde{H}_{RIR,\,i}(\omega) \approx \mid H_{RIR,\,i}(\omega) \mid^2 . \tag{29}$$

Details of the reverberant component estimation are provided next.

1) Estimating the Room Impulse Response (RIR): The room impulse response is estimated from the audio recording in the frequency domain, $\widetilde{Y}(\omega)$. More specifically, a block-based framework based on spectral subtraction is used for blind RIR estimation from $\widetilde{Y}(\omega)$. Consider the audio frame containing the direct signal $S_0(\omega)$ only. This is generally the first frame of the voiced part of the audio recording, that is, $\widetilde{Y}_0(\omega) = S_0(\omega)$. The impulse response for the first block $\widetilde{H}_{RIR,0}(\omega)$ is estimated by calculating ratio of the magnitude of the second block, $|\widetilde{Y}_1(\omega)|^2$, to that of a previous block $|\widetilde{Y}_0(\omega)|^2$. Similarly, impulse response for the i^{th} block $\widetilde{H}_{RIR,i}(\omega)$ is estimated by calculating the ratio of the magnitude of the current block, $|\widetilde{Y}_i(\omega)|^2$, to that of the previous block $|\widetilde{Y}_{i-1}(\omega)|^2$, that is,

$$|C_i(\omega)|^2 = \begin{cases} \frac{|\widetilde{Y}_i(\omega)|^2}{|\widetilde{Y}_{i-1}(\omega)|^2}; & \frac{|\widetilde{Y}_i(\omega)|^2}{|\widetilde{Y}_{i-1}(\omega)|^2} < |\widetilde{H}_i(\omega)|^2 \\ |\widetilde{H}_i(\omega)|^2 \times Bias_i(\omega) + \varepsilon; & otherwise \end{cases} \tag{30}$$

where $i = 1, \ldots, L$, and $Bias_i(\omega) > 1$ and ε is a small positive value.

The minimum of the above ratio corresponds to the fastest rate of decay which is used to estimate $\widetilde{H}_{RIR,i}(\omega)$. It is important to mention that the unbounded rate of decay may lead to nonreal acoustic spaces. The following constraints are used to ensure a bounded impulse response estimate $\widetilde{H}_{RIR,i}(\omega)$, i.e.,

$$|C_i(\omega)|^2 = \begin{cases} MaxValue_i(\omega); & |C_i(\omega)|^2 > MaxValue_i(\omega) \\ |C_i(\omega)|^2; & otherwise \end{cases} \tag{31}$$

where the frequency-dependent $MaxValue_i(\omega)$ reflects the type of decay expected in real reverberant systems.

Finally, frequency-dependent temporal smoothing is applied to obtain a stable estimate of the room impulse response as,

$$|\widetilde{H}_{RIR,\tau}(\omega)|^2 = \alpha_i(\omega)|\widetilde{H}_{RIR,\tau-1}(\omega)|^2 + (1 - \alpha_i(\omega))|C_i(\omega)|^2 \tag{32}$$

where τ indicates the current time frame, and $\alpha_i(\omega)$ is a frequency-dependent parameter that controls the amount of smoothing $0 \leq \alpha_i < 1$.

The current estimates of RIR $\widetilde{H}_{RIR,i}(\omega), i = 1, \ldots, L$ are used for reverberant component estimation. For the following frames, the $\widetilde{H}_{RIR,i}(\omega)$ is updated recursively to obtain a stable RIR estimate.

2) Reverberant Component Estimation: With the perceptually relevant estimates, $\widetilde{H}_{RIR,\,i}(\omega)$, (27) can be expressed as,

$$\widetilde{S}(\omega)\widetilde{H}_{RIR,0}(\omega) = \widetilde{Y}(\omega) - \sum_{i=1}^{K} S(\omega)Z^{-iL}\widetilde{H}_{RIR,i}(\omega). \tag{33}$$

To better understand this operation, consider the process for a given block of the input signal $\widetilde{Y}_0(\omega)$, which consists of the current block of the dry signal convolved with $H_{RIR,0}(\omega)$, plus the previous block of the dry signal convolved with $H_{RIR,1}(\omega)$, and so on. The operation of the block-based IIR structure can be described mathematically as,

$$\widetilde{S}_0(\omega)\widetilde{H}_{RIR,0}(\omega) = \widetilde{Y}_0(\omega) - \sum_{i=1}^{K} \widetilde{S}_i(\omega)\widetilde{H}_{RIR,i}(\omega) \quad (34)$$

where $\widetilde{S}_i(\omega)$ is an estimate of the true value of $S_i(\omega)$. If the block size L is chosen to be small enough, setting $\widetilde{H}_{RIR,0(\omega)}$ to 1 is reasonable, so (34) now can be expressed as,

$$\widetilde{S}_0(\omega) = \widetilde{Y}_0(\omega) - \sum_{i=1}^{K} \widetilde{S}_i(\omega)\widetilde{H}_{RIR,i}(\omega). \quad (35)$$

Here estimate of the reverberant signal component (the second term in the above (35)) can be expressed as,

$$\widetilde{R}_0(\omega) = \sum_{i=1}^{K} \widetilde{S}_i(\omega)\widetilde{H}_{RIR,i}(\omega). \quad (36)$$

The block-based impulse response is approximated by the magnitude of the frequency domain representations of the $L+1$ blocks. Lack of phase information leads us to make the following perceptually motivated approximations,

$$|\widetilde{S}_0(\omega)|^2 = |\widetilde{Y}_0(\omega) - \sum_{i=1}^{K} \widetilde{S}_i(\omega)\widetilde{H}_{RIR,i}(\omega)|^2 \quad (37)$$

$$\approx |\widetilde{Y}_0(\omega)|^2 - \sum_{i=1}^{K} |\widetilde{S}_i(\omega)|^2|\widetilde{H}_{RIR,i}(\omega)|^2 \quad (38)$$

and

$$|\widetilde{R}_0(\omega)|^2 = |\sum_{i=1}^{K} \widetilde{S}_i(\omega)\widetilde{H}_{RIR,i}(\omega)|^2 \quad (39)$$

$$\approx \sum_{i=1}^{K} |\widetilde{S}_i(\omega)|^2|\widetilde{H}_{RIR,i}(\omega)|^2 \quad (40)$$

$$|\widetilde{R}_0(\omega)|^2 \approx \sum_{i=1}^{K} |\widetilde{S}_i(\omega)|^2|\widetilde{H}_{RIR,i}(\omega)|^2. \quad (41)$$

With these approximations, the necessary gains required to estimate $\widetilde{S}_0(\omega)$ and $\widetilde{R}_0(\omega)$, is defined as,

$$G_s(\omega) = \frac{|\widetilde{S}_0|^2}{|\widetilde{Y}_0|^2}. \quad (42)$$

Substituting (38) into (42), the gain can be expressed as,

$$G_s(\omega) = 1 - \frac{\sum_{i=1}^{L} |\widetilde{S}_i|^2|\widetilde{H}_{RIR,i}(\omega)|^2}{|\widetilde{Y}_0|^2}. \quad (43)$$

To avoid the gain being negative, it should be limited to an appropriate range. In order to guarantee the reverberant component extraction process is stable, the $G_S(\omega)$ should not exceed 1, i.e.,

$$G_S(\omega) = \begin{cases} MinGain(\omega); & G_S(\omega) < MinGain(\omega) \\ G_s(\omega); & otherwise \end{cases}. \quad (44)$$

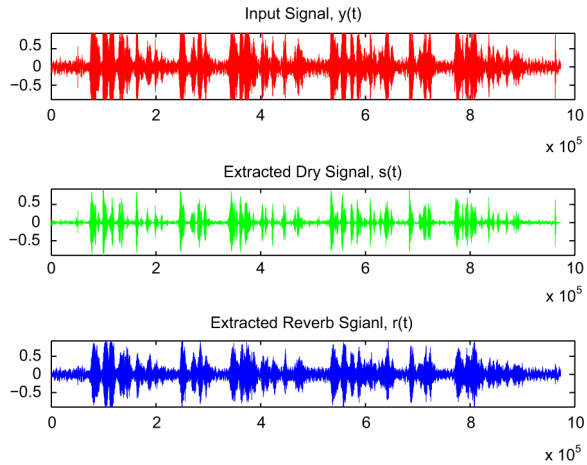

Fig. 4. Dereverberation experimental result for a weakly reverberant environment in the presence of weak background noise: shown in the top panel is the plot of the test recording $\tilde{y}(t)$, in the middle panel is the estimated dry speech $s(t)$, and in the bottom panel is the estimated reverberant speech $r(t)$.

The gain for the reverberant component estimation $G_R(\omega)$ can be expressed as,

$$G_R(\omega) = 1.0 - G_S(\omega). \quad (45)$$

Smoothing is used to mitigate any abrupt changes in the estimated gain,

$$G'_{R,\tau}(\omega) = (1 - \gamma(\omega))G'_{R,\tau-1} + \gamma(\omega)G_{R,\tau}(\omega) \quad (46)$$

where τ indicates the current time frame of the process and $\gamma(\omega)$ ranges between 0 and 1. It determines the amount of smoothing that is applied to gain vectors at each frequency over time.

The estimated reverberant component in the frequency-domain can be expressed as,

$$\widetilde{R}_0(\omega) = G'_R(\omega)\widetilde{Y}_0(\omega). \quad (47)$$

The above block-based blind reverberant component estimation process is repeated for each frame of the input signal.

The effectiveness of the block-based blind dereverberation method discussed in this Section is tested for a speech recording captured in a small office. Shown in Fig. 4 are the temporal plots of the test recording captured in a reverberant environment (top), estimated dry signal (middle), and estimated reverberant signal (bottom). It can be observed from Fig. 4 that the blind dereverberation scheme is effective in separating the dry and reverberant signal from the test speech recording.

Effectiveness of this method has also been evaluated through a number of experiments on speech recordings of two speakers reading different texts and made in four different environments with different levels of reverberation and ambient noise. Due to space limitation, dereverberation results for a highly reverberant acoustic environment with strong background noise are included here. Shown in Fig. 5 are the temporal plots of the test recording made in a highly reverberant environment with strong background noise (top), estimated dry signal (middle), and estimated reverberant signal (bottom). It can be observed from Fig. 5 that even for highly reverberant environment with the presence of strong background noise, this method still works at an acceptable level.

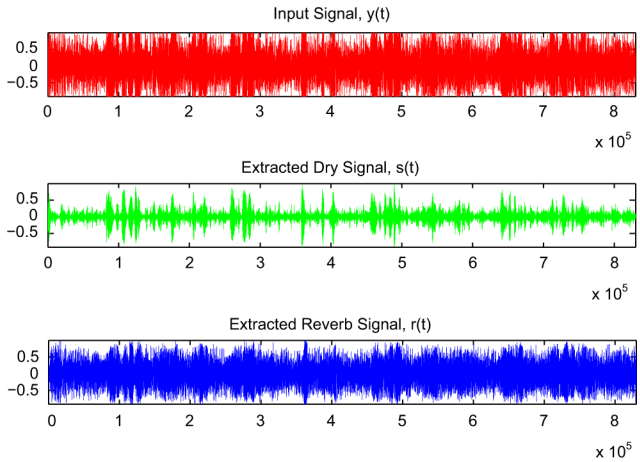

Fig. 5. Dereverberation experimental result for a highly reverberant environment in the presence of strong background noise: shown in the top panel is the plot of the test recording $\tilde{y}(t)$, in the middle panel is the estimated dry speech $s(t)$, and in the bottom panel is the estimated reverberant speech $r(t)$.

C. Acoustic Feature Fusion

The background noise and the reverberant components estimated from the test-recording are fused to obtain a feature vector characterizing the acoustic environment. A feature vector consisting of feature vectors estimated from background noise $\eta(t)$ and the reverberant component $r(t)$ is obtained. Estimated background noise and reverberant component can be modeled as the output of RIR filter $h_{RIR}(t)$. Cepestral analysis is therefore performed to capture traces of acoustic environment. We have shown in [19], [20] that *Mel-frequency Cepstral Coefficients* (MFCC) and *Logarithmic Mel-spectral Coefficients* (LMSC) perform significantly better than other features such as DFT (discrete Fourier transform), DCT (discrete cosine transform), etc.

The motivation behind selecting MFCC for acoustic environment identification can be justified base on the fact that the MFCC has been successfully used for speaker recognition applications. In speaker recognition systems, the speech signal $s(t)$ is generally modeled as $s(t) = h_{VT}(t) * u(t)$, where $h_{VT}(t)$ is the vocal track impulse response, $u(t)$ represents the excitation source and $*$ represents the convolution operator. The *vocal track impulse response* $h_{VT}(t)$ is used to fully characterize each speaker. In MFCC-based speaker recognition systems, the MFCC relies on the $h_{VT}(t)$ to discriminate between the speakers. There exists a strong analogy between the acoustic environment identification process and the speaker recognition process. As, an acoustic environment is also characterized by the *acoustic impulse response*, $h_{RIR}(t)$. The recording made in an acoustic environment with impulse response $h_{RIR}(t)$ can be modeled as, $y(t) = h_{RIR}(t) * d(t)$, where, $y(t)$ is the recording and $d(t)$ is the direct signal. It is therefore expected that the MFCC/LMSE will perform well for the acoustic environment identification task.

To extract feature vector, estimated components $\eta(t)$ and $r(t)$ are transformed into the cepstral domain. More specifically, MFCC and LMSC are computed from both $\eta(t)$ and $r(t)$. To this end, a 60-D feature vector (consisting of 30-D MFCC and 30-D LMSC vectors) is obtained from $r(t)$. Similarly, a 60-D feature vector is also obtained from $\eta(t)$. It is important to mention that the PF-based noise estimation method works in the logarithmic spectral domain, e.g., $\boldsymbol{\Upsilon}_k$ is the LMSC. The 30-D feature vector consisting of MFCC is therefore estimated by calculating forward DCT of $\boldsymbol{\Upsilon}_k$. In addition, for each estimated component, four higher-order statistics, i.e., mean, variance, skewness, and kurtosis are also used to capture the characteristic of environments which resulted in a 128-D feature vector.

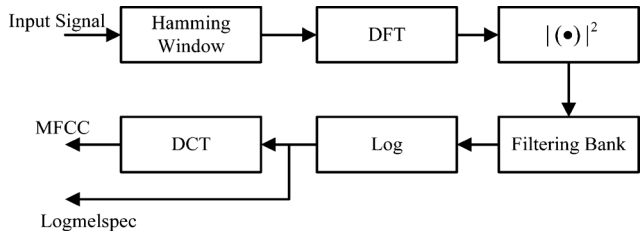

Fig. 6. Flowchart of feature extraction stage.

The proposed framework for joint background noise and reverberation estimation is summarized in the Fig. 7 and details of each processing block is provided in the following Algorithm.

Algorithm: Outline of each step of our proposed framework

The following processing steps are used to estimate noise $\boldsymbol{\Upsilon}_k$ and reverberation components from the input audio.

1) *Blind Dereverberation*
   The reverberant component $r(t)$ is estimated according to (25)–(32).
2) *Spectral Estimation*
   The spectral estimation is achieved by segmenting the input audio into 25 ms overlap blocks with 50% overlapping followed by temporal smoothing using 2048-point Hamming window, a frequency domain transformation using the magnitude square of a 2048-point DFT, filtering using Mel-filtering bank and rescaling using the logarithmic operator. The flowchart of the spectral estimation process is shown in Fig. 6.
3) *Prior Noise Density Estimation*
   The prior noise density $p(\boldsymbol{\Upsilon}_0)$, required to initialize the particle filter, is estimated via (23). The parameters are trained from the noise-only signal, which is obtained by the voice activity detection process.
4) *Particle Evolution*
   All particles, $\boldsymbol{\Upsilon}_k^j, j = 1 : J$ are propagated by the state transition function described in (11).
5) *Prediction Model Updating*
   The prediction model given in (11) is updated for each iteration and model parameters are updated using (10) and (12).
6) *Noise Evaluation*
   The noise samples $\boldsymbol{\Upsilon}_k^j$ are evaluated according to (20) and (21).
7) *Importance Resampling*
   The weights are resampled to avoid the degeneracy problem.
8) Steps: 2 to 7 are repeated until all frames are processed.

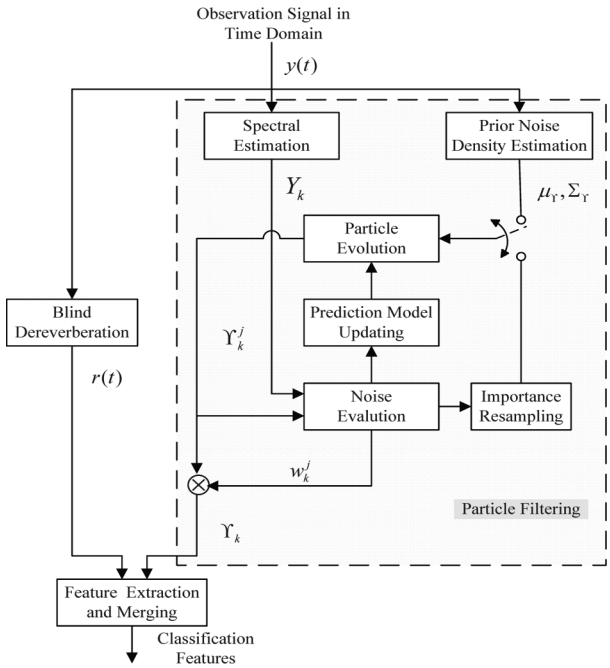

Fig. 7. Proposed framework for generating feature vector used for AEI. Solid arrow represents the flow of the signal. The part highlighted by dotted rectangles represents the particle filtering-based noise estimation.

TABLE I
DESCRIPTION OF ACOUSTIC ENVIRONMENTS CONSIDERED

Environments	Types	Descriptions
E_1 & E_2	Small Office 1 & 2	11' × 12.5' × 9' predominantly carpet & drywall
E_3 & E_4	Restroom 1 & 2	17' × 14' × 9' predominantly ceramic tiles
E_5 & E_6	Hallway 1 & 2	32' × 8' × 9' predominantly concrete floor and drywalls
E_7 & E_8	Outdoors 1 & 2	Outdoors

TABLE II
MICROPHONES USED

Microphone	Description
M_1	Shure SM58-LC Cardioid Vocal Mic.
M_2	Dynamic Omnidirectional Mic. (Electro-Voice 635A/B)
M_3	Cond. Mic. (BEHRINGER ECM8000)
M_4	Right Built-in Mic. of ZOOM R16 Recorder

IV. PERFORMANCE EVALUATION

The effectiveness of the proposed framework is evaluated using a data set of human speech. Details for the data set used, experimental setup and experimental results are provided next.

A. Data Set

The performance of the proposed scheme is evaluated using a data set consisting of more than 2240 speech recordings. Speech recordings of two speakers, a male and a female, were recorded with four different types of microphones in eight different acoustic environments. A description of each acoustic environment is listed in Table I. Audio was recorded using four commercial-grade microphones with 44.1 kHz sampling rate and 16 bits/sample resolution. Details of makes and models of the microphones used are provided in Table II.

TABLE III
CLASSIFICATION PERFORMANCE OF MICROPHONE INDEPENDENT TEST

Mic. Type	Orig. Audio	Rever. Comp.
M_2	87.50%	100.0%
M_3	97.22%	100.0%
M_4	90.28%	98.85%

B. Experimental Setup

Given the input audio, $y(t)$, the background noise, $\eta(t)$, is estimated according to the Algorithm described above. To extract the reverberation component from the input audio recording, the reverberation estimation parameters L, $Bias_i(\omega)$, ε and $\gamma(\omega)$ are set to 30, 1.1, 0.0001 and 0.5 respectively.

Before applying feature extraction from the reverberation component, it is pre-emphasized according to $r(t) = r(t) - p \times r(t-1)$ with $p = 0.97$. The pre-emphasized signal is decomposed into overlapping frames with 25 ms duration and shift of 10 ms. For each frame, a 60-D feature vector is extracted. Similar procedure is applied to the observation $y(t)$, which is used for noise estimation. Noise is estimated in the Mel-Cepstral domain so that input audio is pre-emphasized according to $y(t) = y(t) - p \times y(t-1)$ before applying the particle filtering.

For classification, a multiclass Support Vector Machine (SVM)[66] with a *radial basis kernel function (RBF)* is used. The kernel selection of SVM plays a role on the classification accuracy. Motivation behind selecting SVM with RBF kernel is that it outperforms the linear and the polynomial kernels [67], [68]. In addition, the SVM with RBF also performs significantly better than other classifiers such as Fisher Linear Discernment [69], Logistic regression [70]. For each experiment, the optimal parameters for the classifier are determined using a grid-search technique with five-fold cross-validation on the training data. The kernel parameter C (for all the kernels) and γ (for RBF and polynomial kernels) were carried out on the multiplicative grid $\mathcal{G}_C \times \mathcal{G}_\gamma$, $\mathcal{G}_C = \{2^a\}, a \in \{-10, 10\}$, $\mathcal{G}_\gamma = \{2^b\}, b \in \{-10, 10\}$. Each experiment is repeated 10 times and classification accuracy averaged over all runs is used for performance evaluation.

C. Experimental Results

1) Microphone Independent Test: In our first experiment, we investigated the *microphone independent acoustic environment identification* performance. The motivation behind using a microphone independent analysis is that for AEI we do not have any knowledge of the microphone used for the recording. So, acoustic environment identification algorithm should be robust (not sensitive) to the type of microphones used for recording. To this end, we considered a two-class identification scenario, that is, the identification of two acoustic environments, say E_1 and E_3. The SVM classifier was trained using feature vectors extracted from the recordings made in acoustic environments E_1 and E_3 with microphone M_1. The trained classifier was tested using recordings made in acoustic environments E_1 and E_3 with microphones M_2, M_3 and M_4.

Shown in Table III is the classification accuracy for the *microphone independent test*. Shown in the second column

TABLE IV
CLASSIFICATION PERFORMANCE OF SEMIBLIND ENVIRONMENT IDENTIFICATION

Mic. Type	Orig. Audio	Rever. Comp.	Rever. Comp. + Noise
M_1	90.41%	**99.32%**	98.63%
M_2	95.42%	98.24%	**98.57%**
M_3	82.99%	91.32%	**93.75%**
M_4	92.71%	**96.88%**	94.79%

TABLE V
CLASSIFICATION PERFORMANCE OF BLIND ENVIRONMENT IDENTIFICATION USING REVERBERATION AND NOISE FEATURES

Mic. Type	Orig. Audio	Rever. Comp. + Noise
M_1	80.56%	96.53%
M_3	79.17%	94.10%
M_4	74.65%	97.46%

of Table III are the classification performance for the original audio recordings without dereverberation or denoising, that is, the feature vector is extracted from the original audio directly. Classification performance for the proposed framework is shown in the third column. It can be observed from Table III that the AEI performance for the proposed method is independent of the microphone type used. It is important to note that the *original audio case* classification performance is not stable, since, it varies from 87.5% to 97.22%. However, the performance of the proposed method is stable and on average higher than the *original audio case*. One of the reasons for improved AEI performance for the proposed method is that the training and testing recordings come from the same environments so the source mismatch problem does not exist. It is reasonable to claim that the proposed reverberation-based method not only improves the AEI performance but is also independent of the microphone used.

2) Semi-blind Environment Identification: In our second experiment, we considered a four-class blind classification scenario, that is, the SVM classifier is trained using recordings made in environments E_1, E_3, E_5 and E_7 and tested on the recordings made in environments E_2, E_4, E_6 and E_8. For this experiment recording made with M_1, M_2, and M_3 are used. Shown in Table IV is the AEI performance of the original audio (second column), proposed method using reverberation component only (third column), and proposed method using reverberation + background noise components (fourth column). It can be observed from Table IV that dereverberation provided on average 6.7% performance improvement averaged over all microphones. Merging dereverberation and noise estimation provided addition 1% performance improvement. It can also be observed from Table IV (fourth column) that AEI using joint reverberation and noise components on average does improve the classification performance.

3) Blind Environment Identification: In the third experiment, performance of the proposed framework is tested for the blind AEI case, that is, knowledge of the microphone (used for making the recording) and the recording environment are not exploited during classification process. For this test, the SVM classifier was trained using feature vectors based on joint reverberation and noise components extracted from recordings made in environments E_1, E_3, E_5 and E_7 using microphone M_2, and tested on the recordings captured in environments E_2, E_4, E_6 and E_8 using microphones M_1, M_3 and M_4. The classification performance for the four-class blind environments identification are shown in Table V.

TABLE VI
CONFUSION MATRIX OF BLIND ENVIRONMENT IDENTIFICATION USING REVERBERATION AND NOISE FEATURES FOR M_1

Trained Environments	Predicted Environments			
	E_2	E_4	E_6	E_8
E_1	100.0%	0	0	0
E_3	0	100.0%	0	0
E_5	4.17%	1.39%	90.28%	4.17%
E_7	0	2.78%	1.39%	95.83%

TABLE VII
CONFUSION MATRIX OF BLIND ENVIRONMENT IDENTIFICATION USING REVERBERATION AND NOISE FEATURES FOR M_3

Trained Environments	Predicted Environments			
	E_2	E_4	E_6	E_8
E_1	97.22%	0	2.78%	0
E_3	5.56%	87.50%	6.94%	0
E_5	0	8.33%	91.67%	0
E_7	0	0	0	100.0%

TABLE VIII
CONFUSION MATRIX OF BLIND ENVIRONMENT IDENTIFICATION USING REVERBERATION AND NOISE FEATURES FOR M_4

Trained Environments	Predicted Environments			
	E_2	E_4	E_6	E_8
E_1	100.0%	0	0	0
E_3	0	98.55%	1.45%	0
E_5	0	4.35%	91.30%	4.35%
E_7	0	0	0	100.0%

It can be observed from Table V that classification performance for the original recordings deteriorates significantly (roughly 18% on average) compared with the semi-blind case (see Table IV). However, performance for the proposed system deteriorates marginally, that is, it decreases only about 1% on average compared with the semi-blind case (see Table IV). Similar observations were made when we trained the classifier on recordings collected using microphone M_1, M_3 or M_4 and tested on the recordings made using the rest of the microphones.

Blind identification performance of the proposed system is also evaluated using confusion matrix based measure. To this end, confusion matrices are computed for all three microphones. Shown in Tables VI–VIII are the confusion matrices for M_1, M_3 and M_4, respectively. It can be observed from Tables VI–VII that the error distribution is slightly dependent on a given microphone types. Moreover, it can also be observed that for all microphones classification errors for E_4 and E_6 are slightly higher than E_2 and E_8, and the false rates between E_4 and E_6 are also higher. The higher error for E_4 and E_6 can be attributed to the similarity in the acoustic environment structure, that is, both acoustic environments, e.g., E_4 and E_6 have concrete floors.

It is worth mentioning that classification accuracy of SVM classifier depends on the underlying kernel used. In AEI results shown in Table V are obtained for RBF kernel. Superior classification accuracy is the major motivation behind selection RBF kernel [67], [68]. To validate this claim, we compared classification performance of SVM with RBF kernel with SVM with

TABLE IX
BLIND ENVIRONMENT IDENTIFICATION PERFORMANCE COMPARISON OF SVM WITH RBF, LINEAR, AND POLYNOMIAL KERNELS

Mic. Type	Rever. Comp. + Noise		
	RBF	Linear	Polynomial
M_1	**96.53**%	89.93%	92.01%
M_3	**94.10**%	66.31%	67.36%
M_4	**97.46**%	60.50%	69.92%

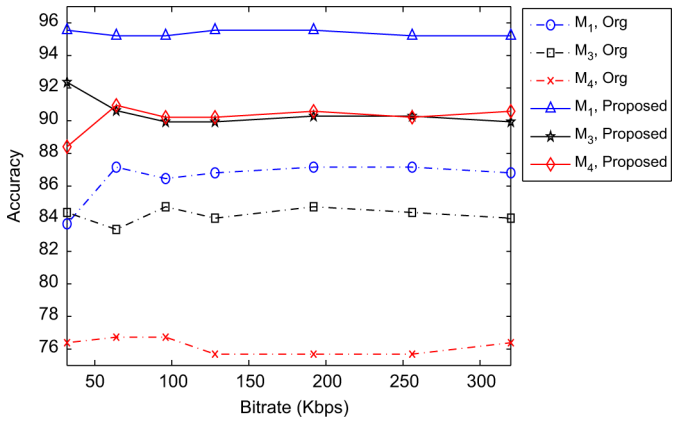

Fig. 8. Classification performance of blind environment identification under MP3 compression with various compression biterates.

linear and SVM with polynomial kernels. The this end, the blind environment identification experiment was repeated for both the linear and the polynomial kernels with same configurations (that is, training on M_2 and testing on the rest of the microphones) and similar parameter selection as used for RBF kernel. The average classification accuracies for RBF, linear, and polynomial kernels are shown in Table IX.

It can be observed from Table IX that on average RBF kernel performs significantly better that both the linear and the polynomial kernels. It can also be observed that the polynomial kernel performs relatively better than the linear kernel, which is not a surprising observation.

4) Robustness Against MP3 Compression: In our fourth experiment, we tested the robustness of our scheme against MP3 compression attacks. To this end, audio recordings were compressed into MP3 format using ffmpeg [71] with bitrates R, $R \in \{32, 64, 96, 128, 192, 256, 320\}$ kbps. For feature extraction, the compressed recordings were converted back to wave format. The SVM classifier was trained using feature vectors based on joint reverberation and noise components extracted from decompressed recordings made in environments E_1, E_3, E_5 and E_7 using microphone M_2, and tested on the recordings captured in environments E_2, E_4, E_6 and E_8 using microphones M_1, M_3 and M_4.

The average classification performance, averaged over 100 runs, of the proposed scheme for blind identification of the data set compressed using biterates ranging from 32 to 320 kbps is shown in Fig. 8. Here, the dashed-lines with circles, squares and cross represent the classification results for compressed recordings (without dereverberation) made with microphones M_1, M_3 and M_4, respectively; and the solid lines triangle, stars and diamonds represent the blind identification performance of the proposed scheme.

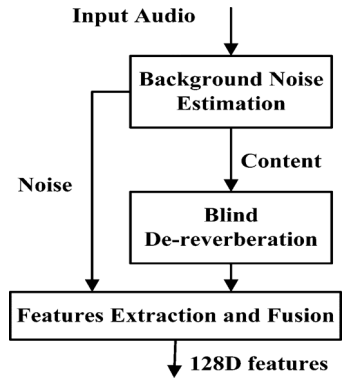

Fig. 9. Cascaded feature extraction framework.

TABLE X
CLASSIFICATION PERFORMANCE OF CASCADED FEATURE EXTRACTION FRAMEWORK

Mic. Type	Reverb. Comp. + Noise **(Parallel)**	Reverb. Comp.+ Noise **(Cascaded)**
M_1	96.53%	97.91%
M_3	94.10%	67.01%
M_4	97.46%	78.12%

It can be observed from Fig. 8 that the performance of the proposed scheme is robust to MP3 compression attack. It can also be observed from both the Table V and Fig. 8 that microphone M_3 and M_4 are more sensitive to MP3 compression, since, compression attack resulted in an average performance drop of 4.05% and 6.91%, respectively. Whereas, microphones M_1 is experienced a negligible performance degradations of 1.12%. Whatever the compression bitrate, our proposed scheme has better performance than that of using original audio.

Another interesting observation can be made from Fig. 8 that classification performance is relatively more unstable performance for bitrate ≤ 96 kbps. Performance fluctuation in low bitrate can be attributed to relatively unstable particle filtering performance in the presence of strong distortion due to aggressive MP3 compression. In addition, distortion due to MP3 compression might have also resulted in dereverberation performance degradation.

5) Performance Comparison-Parallel Vs Cascaded Framework: In our fifth experiment, performance of the proposed framework for parallel reverberation and noise estimation is compared with the cascaded architecture, that is, noise estimation followed by reverberation estimation. The feature extraction from the cascaded architecture is shown in Fig. 9. Here, the particle filtering (PF) stage estimates background noise from the input audio and the output of PF stage is applied at the input of the blind dereverberation stage for reverberant component estimation. For each architecture, the SVM classifier was trained on the recordings made in environments E_1, E_3, E_5 and E_7 by microphone M_2, and tested on the recordings made in environments E_2, E_4, E_6 and E_8 using microphones M_1, and M_4,.

Classification performance for the cascaded framework is shown in Table X. It can be observed from Table X that the cascaded framework does provide a marginal performance improvement (1% to be exact) for M_1; whereas, there is a significant performance degradation for M_3 and M_4. The performance degradation for cascade framework can be attributed to the fact

TABLE XI
CLASSIFICATION ACCURACY USING TDSM FEATURES

Mic. Type	Feature	
	TDSM	MFCC+LMCC
M_1	78.82%	96.53%
M_3	56.25%	94.10%
M_4	38.19%	97.16%

that denoising removes some of the late reverberations. The distortion introduced by denoising stage propagates to the dereverberation stage which also results in performance degradation. This experiment also indicates that the performance of the cascaded framework is strongly dependent on the microphone sensitivity. The performance degradation for a low quality microphone M_4 might be attributed to the following: (1) Denoising removes the late reverberation which is the primary contributor of the AEI, and (2) Distortion introduced by denoising decreases the dereverberation accuracy. The parallel architecture of the proposed framework, on the other hand, decouples both signal estimation stages, that is, the noise and reverberation estimation stages. In case of parallel architecture, the artifacts in one estimation stage does not deteriorate estimation performance of the other stage. The parallel architecture is therefore expected perform better than the cascade architecture. The classification performance shown in Tables V and X confirms this claim. It is worth mentioning that parallel architecture is not an optimal solution to this interdependent component estimation problem. We have observed through extensive experimentation that both the background noise and the reverberation components are tightly coupled. The optimal solution therefore would require joint estimation of both the noise and reverberation. In our future work we will investigate this issue.

6) Performance Comparison-Using TDSM-Based Features: In our final experiment, we compare the performance of proposed framework using existing state-of-the-art features used in audio steganalysis [72]–[74]. Liu *et al.* [72]–[74] proposed *Temporal Derivative-based Spectrum and Mel-Cepstrum (TDSM)* features for audio steganalysis. Liu *et al.* in [72]–[74], have experimentally confirmed that the TDSM-based features are effective for audio steganalysis, and exhibit superior performance than existing state-of-the-art [75].

To investigate impact of the TDSM-based features on the performance of the proposed AEI framework, we perform blind AEI using temporal derivative-based spectrum and Mel-Cepstrum features instead of MFCC and LMCC. Shown in Table XI are the detection performance for TDSM-based feature vector. The results shown in Table XI are obtained using the same experimental setting as used in Section IV-C3.

It can be noticed from Table XI that the classification accuracies decrease significantly for TDSM when compared with the used MFCC-based features. In addition, the performance for TDSM-based features is highly dependent on the microphone sensitivity. For example, for built-in microphone M_4, the TDSM-based AEI exhibits the worst accuracy (e.g., 38.19%) which is relatively better than the random guessing (e.g., 25% for four-class classification). The performance degradation due to TDSM-based features can be attributed to the following reasons: (1) As, TDSM-based feature extraction uses second-order derivative (a kind of high-pass filter) which removes the speech content as well as the reverberations (the main contributor of environment identification), and (2) The speech band filter used in Liu's feature extraction methods [72]–[74] removes the speech component along with the reverberation component.

V. CONCLUSION

In this paper, a novel method for acoustic environment identification (AEI) is proposed. Acoustic reverberations and background noise are used to characterize the acoustic environment. Background noise is modeled using a dynamical system and estimated using particle filtering. A blind dereverberation method based on spectral subtraction and inverse filtering is used to estimate the reverberation component. Both the background noise and the reverberation components are used for feature extraction. An 128-D feature vector consisting of MFCC, LMSC, and higher order statistics is used to characterize the acoustic environment. The SVM based classifier is used for the AEI. Performance of the proposed scheme is evaluated using a data set consisting 2240 speech recordings made with four different type of microphones in eight different acoustic environments. The proposed method is tested for various experimental settings such as microphone independent, semi- and full-blind AEI and robustness against MP3 compression attacks. In addition, performance of the proposed framework is also evaluated using Temporal Derivative-based Spectrum and Mel-Cepstrum (TDSM)-based features. Experimental results show that the proposed method improves AEI performance compared with the direct method (i.e., the feature vector is extracted from the audio recording directly). In addition, the proposed scheme is robust to MP3 compression.

Currently we are investigating joint reverberation and background noise estimation and extensions of the proposed method to audio forensic application.

ACKNOWLEDGMENT

The authors would like to thank Prof. Hongxia Wang for the useful discussion and funding support.

REFERENCES

[1] D. Rodriguez, J. Apolinario, and L. Biscainho, "Audio authenticity: Detecting ENF discontinuity with high precision phase analysis," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 3, pp. 534–543, Sep. 2010.

[2] C. Grigoras, "Digital audio recording analysis: The electric network frequency (ENF) criterion," *Int. J. Speech Lang. Law*, vol. 12, no. 1, pp. 1350–1771, 2005.

[3] C. Grigoras, "Application of ENF analysis method in authentication of digital audio and video recordings," *J. Audio Eng. Soc.*, vol. 57, no. 9, pp. 643–661, 2009.

[4] E. Brixen, "ENF: Quantification of the magnetic field," in *Proc. Audio Eng. Soc. 33rd Conf., Audio Forensics: Theory and Practice, 2008*, pp. 1–6.

[5] A. Cooper, "The electric network frequency (ENF) as an aid to authenticating forensic digital audio recordings-an automated approach," in *Proc. Audio Eng. Soc. 33rd Conf., Audio Forensics: Theory and Practice*, Denver, CO, USA, 2008.

[6] E. Brixen, "Techniques for the authentication of digital audio recordings," in *Proc. Audio Eng. Soc. 122nd Convention*, 2007.

[7] D. Garcia-Romero and C. Espy-Wilson, "Speech forensics: Automatic acquisition device identification," *J. Audio Eng. Soc.*, vol. 127, no. 3, pp. 2044–2044, 2010.

[8] D. Garcia-Romero and C. Espy-Wilson, "Automatic acquisition device identification from speech recordings," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing (ICASSP'10)*, Dallas, TX, USA, 2010, pp. 1806–1809.

[9] D. Garcia-Romero and C. Espy-Wilson, "Automatic acquisition device identification from speech recordings," *J. Audio Eng. Soc.*, vol. 124, no. 4, pp. 2530–2530, 2009.

[10] C. Kraetzer, M. Schott, and J. Dittmann, "Unweighted fusion in microphone forensics using a decision tree and linear logistic regression models," in *Proc. 11th ACM Multimedia and Security Workshop*, 2009, pp. 49–56.

[11] C. Kraetzer, A. Oermann, J. Dittmann, and A. Lang, "Digital audio forensics: a first practical evaluation on microphone and environment classification," in *Proc. 9th Workshop on Multimedia and Security*, Dallas, TX, USA, 2007, pp. 63–74.

[12] R. Buchholz, C. Kraetzer, and J. Dittmann, "Microphone classification using Fourier coefficients," in *Lecture Notes in Comput. Sci*. Berlin/ Heidelberg, Germany: Springer, 2010, vol. 5806/2009, pp. 235–246.

[13] A. Oermann, A. Lang, and J. Dittmann, "Verifier-tuple for audio-forensic to determine speaker environment," in *Proc. ACM Multimedia and Security Workshop*, New York, NY, USA, 2005, pp. 57–62.

[14] H. Malik, "Securing speaker verification systen against replay attack," in *Proc. AES 46th Conf. on Audio Forensics*, Denver, CO, USA, 2012.

[15] H. Malik and J. Miller, "Microphone identification using higher-order statistics," in *Proc. AES 46th Conf. on Audio Forensics*, Denver, CO, USA, 2012.

[16] H. Malik and H. Farid, "Audio forensics from acoustic reverberation," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing (ICASSP'10)*, Dallas, TX, USA, 2010, pp. 1710–1713.

[17] U. Chaudhary and H. Malik, "Automaric recording environment classification using acoustic features," in *Proc. Audio Eng. Soc. 129th Convention*, San Francisco, CA, USA, 2010.

[18] S. Ikram and H. Malik, "Digital audio forensics using background noise," in *Proc. IEEE Int. Conf. on Multimedia and Expo*, 2010, 2010, pp. 106–110.

[19] H. Malik and H. Zhao, "Recording environment identification using acoustic reverberation," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing (ICASSP'12)*, Kyoto, Japan, 2012, pp. 1833–1836.

[20] H. Zhao and H. Malik, "Audio forensics using acoustic environment traces," in *Proc. IEEE Statistical Signal Processing Workshop (SSP'12)*, Ann Arbor, MI, USA, 2012, pp. 373–376.

[21] D. Nicolalde and J. Apolinario, "Evaluating digital audio authenticity with spectral distances and ENF phase change," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, Taipei, Taiwan, 2009, pp. 1417–1420.

[22] R. Maher, "Acoustical characterization of gunshots," in *Proc. Signal Processing Applicat. for Public Security and Forensics*, Washington, DC, USA, 2007, pp. 11–13.

[23] R. Maher, "Modeling and signal processing of acoustic gunshot recordings," in *Proc. IEEE Signal Processing Soc. 12th DSP Workshop*, 2006, pp. 24–27.

[24] C. Grigoras, "Statistical tools for multimedia forensics," in *Proc. Audio Eng. Soc. 39th Conf., Audio Forensics: Practices and Challenges*, 2010, pp. 27–32.

[25] R. Yang, Y. Shi, and J. Huang, "Detecting double compression of audio signal," in *Proc. SPIE Media Forensics and Security II 2010*, , vol. 7541.

[26] R. Yang, Y. Shi, and J. Huang, "Defeating fake-quality MP3," in *Proc. of the 11th ACM Workshop on Multimedia and Security (MM Sec)'09*, 2009, 2010, pp. 117–124.

[27] R. Yang, Z. Qu, and J. Huang, "Detecting digital audio forgeries by checking frame offsets," in *Proc. 10th ACM Workshop on Multimedia and Security (MM Sec '08)*, 2008, pp. 21–26.

[28] A. Cooper, "Detecting butt-spliced edits in forensic digital audio recordings," in *Proc. Audio Eng. Soc. 39th Conf., Audio Forensics: Practices and Challenges*, Hillerod, Denmark, 2010.

[29] 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, Eur. Netw. Forensic Sci. Institutes, 2009.

[30] C. Grigoras, "Applications of ENF criterion in forensic audio, video, computer and telecommunication analysis," *Forensic Sci. Int.*, vol. 167, pp. 136–145, 2007.

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

[32] G. Soulodre, "About this dereverberation business: A method for extracting reverberation from audio signals," in *Proc. AES 129th Convention*, San Francisco, CA, USA, 2010.

[33] B. Koenig, D. Lacey, and S. Killion, "Forensic enhancement of digital audio recordings," *J. Audio Eng. Soc.*, vol. 55, no. 5, pp. 352–371, 2007.

[34] D. Boss, "Visualization of magnetic features on analogue audiotapes is still an important task," in *Proc. Audio Eng. Soc. 39th Int. Conf. Audio Forensics*, 2010, pp. 22–26.

[35] D. Begault, B. Brustad, and A. Stanle, "Tape analysis and authentication using multitrack recorders," in *Proc. Audio Eng. Soc. 26th Int. Conf.: Audio Forensics in the Digital Age*, 2005, pp. 115–121.

[36] *AES Recommended Practice for Forensic Purposes-Managing Recorded Audio Materials Intended for Examination*, AES Standard 27–1996 Std., 1996.

[37] *AES Standard for Forensic Purposes-Criteria for the Authentication of Analog Audio Tape Recordings*, AES Standard 43–2000 Std., 2000.

[38] H. David and P. Michael, Swgde best practices for forensic audio, version 1.0., Tech. Rep, 2008.

[39] H. Hollien, *The Acoustics of Crime, The New Science of Forensic Phonetics*. New York, NY, USA: Plenum, 1990.

[40] H. Hollien, *Forensic Voice Identification*. New York, NY, USA: Academic, 2001.

[41] B. Pellom and J. Hansen, "Voice analysis in adverse conditions: The centennial olympic park bombing 911 call," in *Proc. IEEE Midwest Symp. on Circuits and Syst.*, 1997, pp. 873–876.

[42] R. Malkin and A. Waibel, "Classifying user environment for mobile applications using linear autoencoding of ambient audio," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, 2005, vol. 5, pp. 509–512.

[43] S. Chu, S. Narayanan, C.-C. Kuo, and M. Mataric, "Where am i? scene recognition for mobile robots using audio features," in *Proc. IEEE Int. Conf. on Multimedia and Expo*, 2006, pp. 885–888.

[44] A. Eronen, "Audio-based context recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 321–329, Jan. 2006.

[45] Y. Panagakis and C. Kotropoulos, "Automatic telephone handset identification by spares representation of random spectral features," in *Proc. Multimedia and Security*, 2012, pp. 91–96.

[46] C. Kraetzer, K. Qian, M. Schott, and J. Dittmann, "A context model for microphone forensics and its application in evaluations," *Proc. SPIE Media Watermarking,Security, and Forensics III*, vol. 7880, pp. 78800P:1–15, 2011.

[47] C. Kraetzer, K. Qian, and J. Dittmann, "Extending a context model for microphone forensics," *Proc. SPIE Media Watermarking, Security, and Forensics III*, vol. 8303, pp. 83030S:1–12, 2012.

[48] Q. Liu, A. Sung, and M. Qiao, "Detection of double MP3 compression," *Cognitive Computation, Special Issue: Advances in Computational Intell. and Applicat.*, vol. 2, no. 4, pp. 291–296, 2010.

[49] M. Qiao, A. Sung, and Q. Liu, "Revealing real quality of double compressed MP3 audio," in *Proc. Int. Conf. Multimedia*, 2010, pp. 1011–1014.

[50] E. Brixen, "Acoustics of the crime scene as transmitted by mobile phones," in *Proc. Audio Eng. Soc. 126th Convention*, Munich, Germany, 2009.

[51] X. Pan, X. Zhang, and S. Lyu, "Detecting splicing in digital audios using local noise level estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP'12)*, Kyoto, Japan, 2012, pp. 1841–1844.

[52] J. Zhou, D. Garcia-Romero, and C. Espy-Wilson, "Automatic speech codec identification with applications to tampering detection of speech recordings," in *Proc. Interspeech*, 2011, pp. 2533–2536.

[53] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1997.

[54] K. Yao, K. K. Paliwal, and S. Nakamura, "Noise adaptive speech recognition based on sequential noise parameter estimation," *Speech Commun.*, vol. 42, no. 1, pp. 5–23, 2004.

[55] K. Yao and S. Nakamura, "Sequential noise compensation by sequential Monte Carlo methods," *Adv. Neural Inf. Process. Syst.*, vol. 14, pp. 1213–1220, 2004.

[56] R. Singh and B. Raj, "Tracking noise via dynamical systems with a continuum of staes," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing (ICASSP'03)*, 2003, vol. 1, pp. 396–399.

[57] M. F. Naakamura, "Pa0rticle filter based non-stationary noise tracking for robust speech feature enhancement," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing (ICASSP'05)*, 2005, pp. 257–260.
[58] F. Faubel and M. Wolfel, "Coupling particle filters with automatic speech recognition for speech feature enhancement," in *Proc. Interspeech*, 2006, pp. 37–40.
[59] M. Wolfel and J. W. McDonough, *Distant Speech Recognition*. Hoboken, NJ, USA: Wiley.
[60] M. Fujimoto and S. Naakamura, "Particle filtering and polyak averaging-based non-stationary noise tracking for asr in noise," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005, pp. 337–342.
[61] S. Axelrod, V. Goel, R. Gopinath, P. Olsen, and K. Visweswariash, "Subspace constrained Gaussian mixture models for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 6, pp. 1144–1160, Nov. 2005.
[62] M. Wolfel, "Enhanced speech features by single-channel joint compensation of noise and reverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 312–323, Feb. 2009.
[63] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Application*. Boston, MA, USA: Artech House, 2004.
[64] M. S. Arulampalam, N. G. S. Maskell, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.
[65] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.
[66] C. Chang and C. Lin, Libsvm: A library for support vector machines [Online]. Available: http://www.csie.ntu.edu.tw/~ cjin/libsvm
[67] S. Lyu and H. Farid, "Detecting hidden messages using higher-order statistics and support vector machines," in *Proc. 5th Int.Workshop on Information Hiding*, Noordwijkerhout, The Netherlands, 2002.
[68] T. L. S. Zhu and M. Ogihara, "Using discriminant analysis for multi-class classification: an experimental inverstigation," *Knowl. Inf. Syst.*, vol. 10, no. 4, pp. 453–472, 2006.
[69] S. Lyu and H. Farid, "Steganalysis using higher-order image statistics," *IEEE Trans. Inf. Forensics Security*, vol. 1, no. 1, pp. 111–119, Mar. 2006.
[70] I. Lubenko and A. Ker, "Steganalysis using logistic regression," *Proc. SPIE Media Watermarking,Security, and Forensics III*, vol. 7880, pp. 78800K:1–11, 2011.
[71] [Online]. Available: www.ffmpeg.org
[72] Q. Liu, A. H. Sung, and M. Qiao, "Temporal derivative-based spectrum and Mel-Cepstrum audio steganalysis," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 3, pp. 359–368, Sep. 2009.
[73] Q. Liu, A. H. Sung, and M. Qiao, "Novel stream mining for audio steganalysis," in *Proc. 17th ACM int. conf. on Multimedia*, 2009, pp. 95–104.
[74] Q. Liu, A. H. Sung, and M. Qiao, "Derivative-based audio steganalysis," *ACM Trans. Multimedia Comput., Commun., Applicat.*, vol. 7, no. 3, pp. 18:1–19, 2011.
[75] K. Christian and D. Jana, "Mel-cepstrum-based steganalysis for voip steganography," in *Proc. SPIE, Security, Steganography, and Watermarking of Multimedia Contents IX*, E. J. Delp, III and P. W. Wong, Eds., 2007, vol. 6505.

Hong Zhao received the B.S. and Ph.D. degrees in information security from the Southwest Jiaotong University, Chengdu, China, in 2007 and 2013. From 2010 to 2012, he was a visiting scholar at the University of Michigan–Dearborn. He is currently a research fellow with the Department of Electrical and Electronic Engineering, South University of Science and Technology of China. His current research interests include steganalysis, audio forensics, wireless communication security, etc.

Hafiz Malik (S'02–GS'05–M'08) is an Assistant Professor with the Electrical and Computer Engineering (ECE) Department, The University of Michigan–Dearborn. His research in multimedia forensics and security, wireless sensor networks, steganography/steganalysis, and biometric security is funded by the National Academies and other agencies. He has published more than 40 papers in leading journals, conferences, and workshops. Dr. Malik is serving as an Associate Editor for the *Springer Journal of Signal, Image, and Video Processing* (SIVP) since 2013; he is also on the Review Board of IEEE Technical Committee on Multimedia Communications (MMTC). He organized Special Track on Doctoral Dissertation in Multimedia, in the 6th IEEE International Symposium on Multimedia (ISM) 2006. He is serving on several technical program committees including the IEEE ICASSP, AVSS, ICME, ICIP, MINES, ISPA, CCNC, and ICC.