# Digital Recording System Identification Based on Blind Deconvolution

Michel Kulhandjian[†], Hovannes Kulhandjian[‡*], Claude D'Amours[†], Dimitris Pados[††]

[†]School of Electrical Engineering and Computer Science, University of Ottawa,
Ottawa, Ontario, K1N 6N5, Canada
E-mail: mkk6@buffalo.edu, cdamours@uottawa.ca
[‡]Department of Electrical and Computer Engineering, California State University, Fresno,
Fresno, CA 93740, U.S.A.
E-mail: hkulhandjian@csufresno.edu
[††]Computer & Electrical Engineering & Computer Science & I-SENSE Center, Florida Atlantic University,
Florida, FL 33434, U.S.A.
E-mail: dpados@fau.edu

***Abstract*—In this work, we develop a theoretical framework for reliable digital recording system identification from digital audio files alone, for forensic purposes. A digital recording system consists of a microphone and a digital sound processing card. We view the cascade as a system of unknown impulse response. We expect the same manufacturer and model microphone-sound card combinations to have very similar/near identical impulse responses, bar any unique manufacturing defect. Input voice (or other) signals are modeled as non-stationary processes. The technical problem under consideration becomes blind deconvolution with non-stationary inputs, as it manifests itself in the specific application of digital audio recording equipment classification. We propose a conditionally maximum-likelihood (CML) algorithm to estimate underlying systems impulse response together with a novel nearest neighborhood algorithm for recording system identification. Experimental results demonstrate over $99.2\%$ accuracy in identification of the recording devices.**

***Index Terms*—Audio fingerprinting, blind deconvolution, system identification.**

## I. Introduction

Digital recording system identification allows us to distinguish between different recording systems regardless of the audio format. Each manufacturer or brand of the recording system has unique characteristics in the audio recordings they produce. Recording system identification technology is able to extract recording system characteristics (e.g., impulse response of a recording system) and make comparisons. If similar, they belong to one type of recording system, otherwise they belong to different recording systems. The technology should be robust even if the audio signal has undergone a certain degree of modification. Such modifications may include, as an example, linear disruption such as level changes or bandwidth limitation encountered in the case of radio broadcasting. Other modifications include non-linear disturbances, for example, encoding as MP3 format, or recording with a distorted/high noise recording system. Many of these techniques have applications in criminology and forensics, where determining whether a certain recording is from an original device and thus determining its validity. Moreover, this aspect can be used to distinguish between the normal and live versions of an audio recording, which can improve monitoring of illegal copying of original music for piracy issues. Although there have been significant advances in image forensics, audio forensics is still in its infancy. Over the past several years, recording device identification has gained more attention. For example, Kotropoulos and Samaras [1] studied mobile phone identification from recorded speech signals alone, based on Mel frequency cepstral coefficients (MFCCs), which are extracted from the recorded audio to train a Gaussian Mixture Model (GMM) with diagonal covariance matrices. Garcia-Romero and Espy-Wilson [2] have studied on automatic identification of acquisition devices from speech recordings alone, by using a support vector machine (SVM) classifier to perform closed-set identification experiments in which they focused on two classes of acquisition devices. Panagakis and Kotropoulos [3] studied acquisition device identification using random spectral features (RSFs) and the labeled spectral features (LSFs), which are extracted by applying unsupervised and supervised feature selection to the mean spectrogram of each speech signal. Kraetzer, *et al.,* [4] proposed to identify four microphones in which short-term features and MFCCs are combined together to form the feature vectors, then Naive Bayes classifiers are applied to classify these four microphones. Other results that are based on MFCCs were reported by Hanilçi, *et al.,* [5], Zou, *et al.,* [6], Qin *et al.,* [7] and Qi, *et al.,* [8], where latter uses noise as the intrinsic fingerprint traces of an audio recording device.

Unlike the present methods used in literature, we propose a blind system identification technique without any prior

*Corresponding author.

978-1-5386-8380-4/19//$31.00 

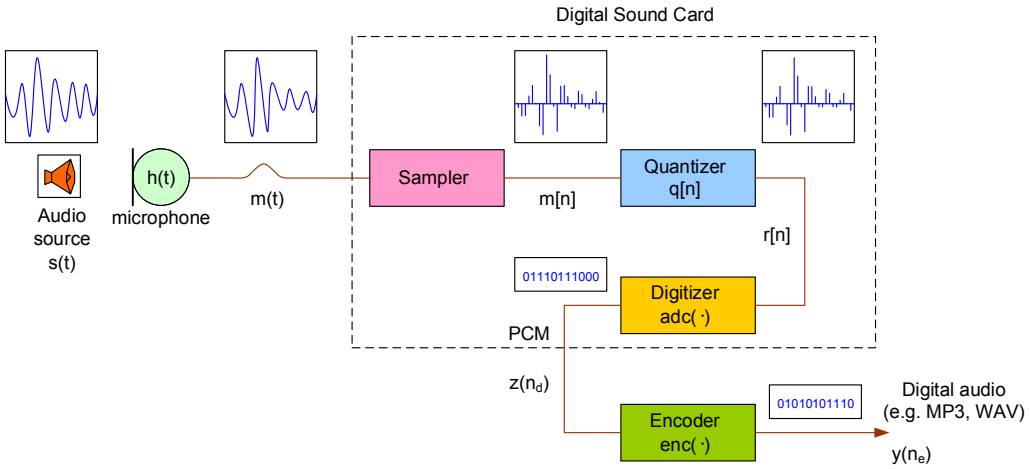


Fig. 1: Digital audio recording system model.

information. Due to memory and low complexity constrictions, we consider blind deconvolution, which does not use any reference signal. Identifying a sound recording system can be very challenging, as we are considering a digitally compressed recorded file (e.g., MP3) without any information of the source or any reference audio signals. The basic mechanism of audio recording can be simply explained as follows: sound source goes through an impulse response of a system, specifically, in our system the impulse response includes a microphone and a sound processing card. However, the profile of a system imposed on the sound source during recording process is quite impossible to be identified by means of human ears when it is played back. Since, the sound heard by a person is not simply the original sound, but instead it is processed through a recording system.

A considerable number of linear signal processing problems reduce to the fundamental task of deconvolution. Those problems are blind, in the sense that neither the source signals nor the impulse response of the system are known, which substantially increases the difficulty of the problem. Conventional techniques of blind deconvolution cannot be applied on this system, since the human voice is a non-stationary process. In our solution, we search for the reliable estimates of the system profile by identifying the location of the poles. Given those poles we develop a near neighbouring algorithm that can classify recorded audio files according to their impulse response of microphone/recording system. Experiments are conducted to evaluate the performance of the proposed classifier, which demonstrate over $99.2\%$ accuracy in identifying the recording devices.

The rest of the paper is organized as follows. In Section II, we present the system and signal model of digital recording system and formulate the recording system identification problem. In Section III, we propose a blind deconvolution solution and the recording system identification performances is presented in Section IV, followed by a few concluding remarks, which are drawn in Section V.

The following notations are used in this paper. All boldface lower case letters indicate column vectors and upper case letters indicate matrices, $()^T$ denotes transpose operation, $|.|$ denotes cardinality of the set, $*$ is the convolution operator, $\|\cdot\|_F^2$ is the Frobenius norm 2 and $\lfloor . \rfloor$ is the flooring function.

## II. Digital Recording System and Problem Formulation

In this section, we first present a digital recording system model with block diagrams. Then we formulate a mathematical model for the underlying problem.

### A. Signal Model and Notation

A simple single channel digital audio recording system consists of a microphone and a digital sound processing card, as shown in Fig. 1. The audio signal $s(t) \in \mathbb{R}, t \in \mathcal{T}$, $\mathcal{T} = \{1, ..., T\} \subset \mathbb{Z}$, is picked up by a microphone, which acts as a transducer converting the air pressure caused by the audio source to an electrical continuous time signal $m(t)$. It is reasonable to assume that the channel between the audio signal $s(t)$ and the microphone is distortionless. Therefore, within the short time frame $T$, the channel from audio signal $s(t)$ to the output of microphone can be considered as a linear time-invariant (LTI) system and is represented by continuous-time impulse response $h(t)$. The output of microphone signal $m(t)$ can thus be formulated as

$$m(t) = h(t) * s(t). \tag{1}$$

The continuous waveform, $m(t)$ is processed by a digital sound card. Most of the sound cards perform pulse code modulation (PCM) on continuous audio waveforms, which involves sampling, quantizing and digitizing by analog-to-digital converter (ADC). The audio signal first passes through a sampler that samples the source $m(t)$ at the *sampling rate* $f_s$, as follows $m[n] = m(n\tau_s) \in \mathbb{R}$ with $n \in \mathcal{N}$, $\mathcal{N} = \{1, ..., N\} \subset \mathbb{Z}$, where $N = \lfloor T/\tau_s \rfloor$ and $\tau_s = 1/f_s$ is the sampling interval. Then the continuous value sampled

audio $m[n]$ is converted to discrete value sampled audio $r[n]$ through a uniform quantizer $Q$, with the impulse response $q[n]$.

After the quantization step, the discrete real-valued signal $r[n]$ are converted into binary format by $\mathsf{adc}(\cdot)$ function, which is the ADC operation. The output of the sound card $z(n_d)$ is a PCM signal, which is an uncompressed digital audio signal. Most of the sound cards' codecs do not do any further processing like compression, which is usually performed in the software. There are two types of compression techniques; lossless (e.g., PCM, WAVE, etc.) and lossy (e.g., MP3, etc.). The raw uncompressed PCM audio signals $z(n_d)$ can be encoded using any encoder $\mathsf{enc}(\cdot)$ to produce audio stream $y(n_e)$.

### B. Cascaded transform system model

Our main objective of recording system identification is given $L$ streams of encoded audio signal $y_l(n_e)$, $1 \leq l \leq L$ identify whether $y_l(n_e)$ are recorded using the same microphone/sound card system or a different one. More precisely, we would like to classify reliably the encoded audio streams $y_l(n_e)$ according to the similarities of the microphone/sound card system utilized for each $1 \leq l \leq L$. In order to identify different recording system the algorithm should be able to extract unique characteristics of the recording systems. One characteristics that is able to uniquely identify a recording system from a model or manufacturer is the impulse response of microphone/sound card system. If $y_l(n_e)$ are recorded using only one microphone/sound card system then the impulse response of microphone/sound card system should be the same for all the $1 \leq l \leq L$. If they are not the same, then we should see differences in the impulse responses. The main potential difference of a specific manufacturer or model will come from the microphone and possibly the manufacturing design of the sound card, as well as, the type of the quantizer used. In order to construct a realizable system for the digital recording system identification purpose, we need to make reasonable assumptions:

1) For all the recordings $1 \leq l \leq L$ the encoding and decoding as well as analog-to-digital and digital-to-analog converting will be the same. Therefore, it is reasonable to assume that we have perfect knowledge of $enc(\cdot)$ and $adc(\cdot)$ and their corresponding $dec(\cdot)$ and $dac(\cdot)$ deterministic functions for $1 \leq l \leq L$.
2) Also, we assume that process of $adc(\cdot)$, $enc(\cdot)$, $dec(\cdot)$ and $dac(\cdot)$ introduce insignificant distortion to discrete real-valued $r[n]$ signal, such that the impulse response of microphone/sound card system can be properly estimated from $x[n]$, shown in Fig. 2 for identification purposes. Therefore, we can assume $x[n] = c \times r[n]$ for some constant $c$.
3) Ignoring any manufacturing defects, we can assume that each device from the same manufacturer or model should have the same fixed characteristics profile. Therefore, the distinct characteristics of the model or manufacturer can be captured in the impulse response of $\mathsf{model}[n]$.
4) Assume the impulse response of the uniform quantizer $Q$ is invertable. We know that $Q$ is non-invertible LTI and non-linear system if non-uniform quantizer is involved.
5) Assume that the channel between the audio signal $s(t)$ and microphone is negligible and does not affect the overall distortion operator.
6) Assume that the microphone impulse response $h(t)$ is an LTI system.
7) Since $x[n]$ in Fig. 2 is in discrete time domain, without loss of generality, the microphone's continuous impulse response $h(t)$ and audio signal $s(t)$ can be equivalently viewed in discrete time domain as $h[n] = h(n\tau_s)$ and $s[n] = s(n\tau_s)$, respectively. As a result, the $m[n] \simeq h[n] * s[n]$ becomes a valid argument, which allows us to work in discrete time domain, instead of continuous.
8) Suppose that overall system distortion operator, $\mathcal{D}$, is an LTI system.
9) Ignore any errors due to other non-linear effects of the mapping of input source $s(t)$ to the output value $y(n_e)$ in addition to sampler, quantization effects, $adc(\cdot)$, $enc(\cdot)$, $dac(\cdot)$ and $dec(\cdot)$.

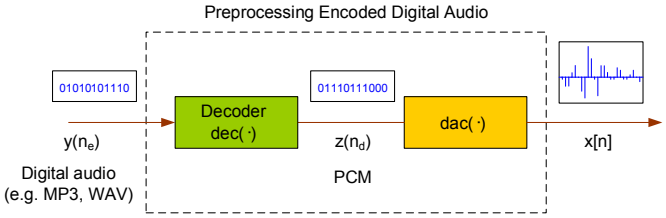


Fig. 2: Preprocessing with known $dec(\cdot)$ and $dac(\cdot)$.

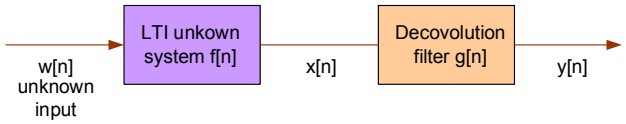


Fig. 3: Blind deconvolution of an unknown LTI system.

We can therefore model the microphone's discrete time impulse response $h[n]$ together with the impulse response of the quantizer $q[n]$ and impulse response of the manufacturer or model sound card design $\mathsf{model}[n]$ as a cascade series connection of impulse responses $f[n] = h[n] * q[n] * \mathsf{model}[n]$ in discrete time domain. Identifying $f[n]$ through digital audio recording identification algorithm is the same as identifying different underlying recording systems. The resultant impulse response of the system $f[n]$ is also LTI, hence, the problem reduces to blind deconvolution [9]. Blind deconvolution is a technique that permits recovery of the source signal that has been convolved (distorted) by the impulse response $f[n]$. Although source signal is unknown, we assume it is white, stationary, independent and identically distributed (i.i.d.) process applied to an LTI system. Since, it is an LTI system then filter operation becomes convolution and we need to

find deconvolution (inverse) impulse response $g[n]$ such that $c[n] = \sum_k f[k]g[n-k] = \alpha\delta[n-n_0]$, where $\alpha$ is an arbitrary non-zero scalar and $n_0$ is a time shift. Fig. 3 represents the blind deconvolution of an unknown LTI system. In literature, deconvolution is used to estimate the source signal $s[n]$. The approach is to first estimate the distortion parameters $\mathcal{D}$ by treating $s[n]$ as a nuisance parameter, and then deconvolve $x[n]$ with $\mathcal{D}$ to recover $s[n]$. In our case, we only need to estimate distortion parameters $\mathcal{D}$ to identify the underlying recording system.

### C. Problem Formulation

In all the mentioned previous techniques they assume that $w[n]$ is an i.i.d. stationary process [10]. However, in our recording system the audio source signal is highly correlated and possesses non-stationary statistical characteristics. Hence, the prior work proposed techniques mentioned above will not work. Therefore, we look into highly correlated and non-stationary methods to solve the blind deconvolution problem. The most common approach to model non-stationary processes is to represent the signal in the form of a stationary model, autoregressive (AR) model [11], [12] with time-varying parameters. The audio source signal can be modeled as time-varying autoregressive (TVAR) model and the impulse response $f[n]$ as an all-pole (AP) infinite impulse response (IIR) model.

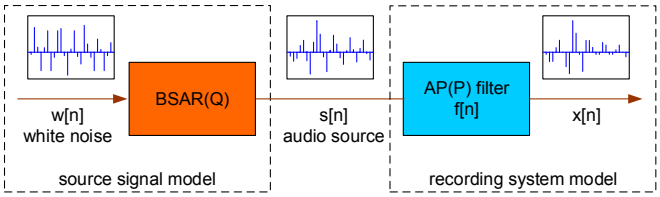


Fig. 4: System model: the output, $x[n]$, is a BSAR process, $s[n]$ is filtered by AP model with parameter $\mathbf{a}$.

Our assumption of the model is a stochastic process that is *globally* non-stationary, yet approximately *locally* stationary; these can be represented by a *quasi-stationary* model. The audio source signal, $s[n]$, is modeled by a block stationary AR (BSAR), which is given by (2). Here, $s[n]$, is partitioned into $M$ contiguous disjoint blocks, block $i \in \mathcal{M}$, $\mathcal{M} = \{1, ..., M\}$ with the length of $N_i = n_{i+1} - n_i$. Within the block $i$, beginning at sample $n_i$, $s[n]$ is assumed to be given by stationary AR model of order $Q_i$. The distortion impulse response $f[n]$, is modeled by an LTI (IIR) filter of order $P$ such that the $x[n]$ is given in terms of $s[n]$. Consider, in block $i$ the preprocessed $x[n]$ of a BSAR process $s[n]$, which is filtered by an all-pole model, as shown in Fig. 4. The equations governing the above system model are expressed as

$$\left.\begin{aligned} s_i[n] &= -\sum_{q=1}^{Q_i} b_i[q]s_i[n-q] + w_i[n] \\ x_i[n] &= -\sum_{p=1}^{P} a[p]x_i[n-p] + s_i[n] \end{aligned}\right\} n \in \mathcal{N}_i, \quad (2)$$

where $w_i[n] \sim \mathcal{N}(0, \sigma_i^2)$, $\sigma_i^2 \in \mathbb{R}^+$, $\mathcal{N}_i = \{n_i, n_i + 1, ..., n_{i+1} - 1\}$ such that $s[n] = s_i[n]$, $x[n] = x_i[n]$, $\forall n \in \mathcal{N}_i \subset \mathcal{N}$. Define vector $\mathbf{b}_i = \{b_i[q], q \in \mathcal{Q}_i = \{1, ..., Q_i\}\}$ and $\mathbf{a} = \{a[p], p \in \mathcal{P} = \{1, ..., P\}\}$, which are the model parameters with $Q_i$, $P$ number of poles, respectively. Therefore, the excitation samples in block $i \in \mathcal{M}$ can be written as

$$\mathbf{w}_i = \mathbf{s}_i + \mathbf{S}_i\mathbf{b}_i, \quad (3)$$

where $\mathbf{s}_i = [s[n_i], s[n_i + 1], \ldots, s[n_i + N_i]]^T$, $\mathbf{b}_i = [b_i[1], b_i[2], \ldots, b_i[Q_i]]^T$ and the data matrix $\mathbf{S}_i = [\bar{\mathbf{s}}_1, \bar{\mathbf{s}}_2, ..., \bar{\mathbf{s}}_{Q_i}]$, where $\bar{\mathbf{s}}_j = [s[n_i - j], s[n_i + 1 - j], ..., s[n_i + N_i - j]]^T$ for $1 \le j \le Q_i$. The probability distribution for the excitation in block $i$ is therefore given by

$$\begin{aligned} \mathbf{w}_i \sim Pr_{\mathbf{w}_i}(\mathbf{w}_i \mid \sigma_i^2) &= \mathcal{N}(\mathbf{w}_i \mid \mathbf{0}_{N_i}, \sigma_i^2\mathbf{I}_{N_i}) \\ &= \frac{1}{(\sqrt{2\pi\sigma_i^2})^{N_i}} \exp\left\{-\frac{\mathbf{s}_i^T\mathbf{s}_i}{2\sigma_i^2}\right\}. \end{aligned} \quad (4)$$

The probability chain rule is given by

$$Pr(\mathbf{s}_1, ..., \mathbf{s}_M) = Pr(\mathbf{s}_1)\prod_{i=2}^{M} Pr(\mathbf{s}_i \mid \mathbf{s}_{i-1}, ..., \mathbf{s}_1). \quad (5)$$

Since the BSAR process depends only on previous $Q_i$ outputs such that if $Q_i \le N_i$, then $Pr(\mathbf{s}_i \mid \mathbf{s}_{i-1}, ..., \mathbf{s}_1) = Pr(\mathbf{s}_i \mid \mathbf{s}_{i-1})$. The probability chain rule (5) reduces to

$$Pr(\mathbf{s}_1, ..., \mathbf{s}_M) = Pr(\mathbf{s}_1)\prod_{i=2}^{M} Pr(\mathbf{s}_i \mid \mathbf{s}_{i-1}). \quad (6)$$

The likelihood function $Pr(\mathbf{s}_i \mid \mathbf{s}_{i-1})$ can not be obtained without any prior information. Instead, we look into $Pr(\mathbf{s}_i \mid \mathbf{s}_{i-1}, \sigma_i^2, \mathbf{b}_i)$, where we assume $\{\sigma_i^2, \mathbf{b}_i\}$ is given. Since the distribution of $\mathbf{w}_i$ is independent of BSAR output and filter parameter $\{\mathbf{s}_{i-1}, \mathbf{b}_i\}$ then $Pr_{\mathbf{w}_i}(\mathbf{w}_i \mid \mathbf{s}_{i-1}, \sigma_i^2, \mathbf{b}_i) = Pr_{\mathbf{w}_i}(\mathbf{w}_i \mid \sigma_i^2)$. Using the transformation of multivariate random variables we can write

$$Pr_{\mathbf{s}_i}(\mathbf{s}_i \mid \mathbf{s}_{i-1}, \sigma_i^2, \mathbf{b}_i) = \frac{Pr_{\mathbf{w}_i}(\mathbf{w}_i' \mid \mathbf{s}_{i-1}, \sigma_i^2, \mathbf{b}_i)}{|J(\mathbf{s}_i, \mathbf{w}_i \mid \mathbf{s}_{i-1}, \sigma_i^2, \mathbf{b}_i)|}, \quad (7)$$

where $\mathbf{w}_i' = \mathbf{s}_i + \mathbf{S}_i\mathbf{b}_i$. Given the fact that the filter operation is linear with known $\{\mathbf{s}_{i-1}, \sigma_i^2, \mathbf{b}_i\}$ parameters the filter operation in (3) can be written as $\mathbf{s}_i = e_i(\mathbf{w}_i \mid \mathbf{s}_{i-1}, \sigma_i^2, \mathbf{b}_i)$ then the inverse filtering is expressed as follows $\mathbf{w}_i = e_i^{-1}(\mathbf{s}_i \mid \mathbf{s}_{i-1}, \sigma_i^2, \mathbf{b}_i)$. The Jacobian* becomes unity, i.e., $J(\mathbf{s}_i, \mathbf{w}_i \mid \mathbf{s}_{i-1}, \sigma_i^2, \mathbf{b}_i) = 1$. Therefore, the likelihood function for the audio source signal, $\mathbf{s}_i$, in block $i \in \mathcal{M}_{\{-1\}}$ is expressed as

$$\begin{aligned} Pr(\mathbf{s}_i \mid \mathbf{s}_{i-1}, \sigma_i^2, \mathbf{b}_i) &= \frac{1}{(\sqrt{2\pi\sigma_i^2})^{N_i}} \\ &\times \exp\left\{-\frac{(\mathbf{s}_i + \mathbf{S}_i\mathbf{b}_i)^T(\mathbf{s}_i + \mathbf{S}_i\mathbf{b}_i)}{2\sigma_i^2}\right\}, \end{aligned} \quad (8)$$

*The Jacobian for the transformation $\mathbf{y} = f(\mathbf{x})$ is $J(\mathbf{y}, \mathbf{x}) = \left|\frac{\delta f^T}{\delta \mathbf{x}}\right|$.

where $\mathcal{M}_{\{-1\}}$ denotes the set $\mathcal{M}$ not including the element "1". Denote $\boldsymbol{\sigma} = \{\sigma_i^2; i \in \mathcal{M}\}$ and $\boldsymbol{\beta} = \{\mathbf{b}_i; i \in \mathcal{M}\}$. Assuming that the $\{\sigma_i^2, \mathbf{b}_i\}$ are independent between blocks, such that $\mathbf{s}_i$ depends only on $\{\mathbf{s}_{i-1}, \sigma_i^2, \mathbf{b}_i\}$ and not on $\{\sigma_j^2, \mathbf{b}_j\}$ for all $j \neq i$, then (6) can be written as

$$Pr(\mathbf{s}_1, ..., \mathbf{s}_M | \boldsymbol{\sigma}, \boldsymbol{\beta}) = Pr(\mathbf{s}_1 | \sigma_1^2, \mathbf{b}_1) \prod_{i=2}^{M} Pr(\mathbf{s}_i | \mathbf{s}_{i-1}, \sigma_i^2, \mathbf{b}_i), \tag{9}$$

and if $N_1 \gg Q_1$, which is often the case with audio signals, it is common practice to approximate $Pr(\mathbf{s}_1 \mid \sigma_1^2, \mathbf{b}_1)$ with $Pr(\mathbf{s}_1 \mid \mathbf{s}_0, \sigma_1^2, \mathbf{b}_1)$, where $\mathbf{s}_0$ is the initial values of audio signal. Then we can write (9) as

$$Pr(\mathbf{s}_1, ..., \mathbf{s}_M \mid \boldsymbol{\sigma}, \boldsymbol{\beta}) = \prod_{i=1}^{M} Pr(\mathbf{s}_i \mid \mathbf{s}_{i-1}, \sigma_i^2, \mathbf{b}_i). \tag{10}$$

Similarly, the preprocessed signal $x[n]$ in block $i \in \mathcal{M}$ can be written as

$$\mathbf{s}_i = \mathbf{x}_i + \mathbf{X}_i \mathbf{a}, \tag{11}$$

where $\mathbf{x}_i = [x[n_i], x[n_i + 1], ..., x[n_i + N_i]]^T$, $\mathbf{a} = [a[1], a[2], ..., a[P]]^T$ and $\mathbf{X}_i = [\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, ..., \bar{\mathbf{x}}_P]$ is the preprocessed matrix, where $\bar{\mathbf{x}}_j = [x[n_i - j], x[n_i + 1 - j], ..., x[n_i + N_i - j]]^T$ for $1 \leq j \leq P$. Using the transformation of multivariate random variables we can write

$$Pr_{\mathbf{x}_i}(\mathbf{x}_i \mid \mathbf{s}_{i-1}, \sigma_i^2, \mathbf{b}_i) = \frac{Pr_{\mathbf{s}_i}(\mathbf{s}_i' \mid \mathbf{s}_{i-1}, \sigma_i^2, \mathbf{b}_i)}{|J(\mathbf{x}_i, \mathbf{s}_i \mid \mathbf{s}_{i-1}, \sigma_i^2, \mathbf{b}_i)|}, \tag{12}$$

where $\mathbf{s}_i' = \mathbf{x}_i + \mathbf{X}_i \mathbf{a}$. According to (11), $\mathbf{s}_{i-1}$ depends on $\{\mathbf{x}_{i-1}, \mathbf{x}_{i-2}, \mathbf{a}\}$ then equivalently we can rewrite (12) as follows

$$Pr_{\mathbf{x}_i}(\mathbf{x}_i \mid \mathbf{x}_{i-1}, \mathbf{x}_{i-2}, \sigma_i^2, \mathbf{b}_i, \mathbf{a}) = \frac{Pr_{\mathbf{s}_i}(\mathbf{s}_i' \mid \mathbf{s}_{i-1}, \sigma_i^2, \mathbf{b}_i)}{|J(\mathbf{x}_i, \mathbf{s}_i \mid \mathbf{x}_{i-1}, \mathbf{x}_{i-2}, \sigma_i^2, \mathbf{b}_i, \mathbf{a})|}. \tag{13}$$

Since the transformation is linear and with the given $\mathbf{x}_{i-1}, \mathbf{x}_{i-2}, \sigma_i^2, \mathbf{b}_i, \mathbf{a}$ parameters, is can be easily shown that $J(\mathbf{x}_i, \mathbf{s}_i \mid \mathbf{x}_{i-1}, \mathbf{x}_{i-2}, \sigma_i^2, \mathbf{b}_i, \mathbf{a}) = 1$. Therefore, the likelihood function for the preprocessed signal, $\mathbf{x}_i$, in block $i \in \mathcal{M}_{\{-1,-2\}}$ in terms of $\{\mathbf{s}_i, \mathbf{s}_{i-1}, \sigma_i^2, \mathbf{b}_i\}$ can be expressed as

$$Pr(\mathbf{x}_i \mid \mathbf{x}_{i-1}, \mathbf{x}_{i-2}, \sigma_i^2, \mathbf{b}_i, \mathbf{a}) = \frac{1}{(\sqrt{2\pi\sigma_i^2})^{N_i}} \times \exp\left\{ -\frac{(\mathbf{s}_i + \mathbf{S}_i \mathbf{b}_i)^T (\mathbf{s}_i + \mathbf{S}_i \mathbf{b}_i)}{2\sigma_i^2} \right\}, \tag{14}$$

where $\mathcal{M}_{\{-1,-2\}}$ denotes the set $\mathcal{M}$ not including the element "1" and "2". Since $\mathbf{x}_i$ depends only on $\{\mathbf{x}_{i-1}, \mathbf{x}_{i-2}, \sigma_i^2, \mathbf{b}_i, \mathbf{a}\}$ then $Pr(\mathbf{x}_i \mid \mathbf{x}_{i-1}, \mathbf{x}_{i-2}, ..., \mathbf{x}_1, \sigma_i^2, \mathbf{b}_i, \mathbf{a}) = Pr(\mathbf{x}_i \mid \mathbf{x}_{i-1}, \mathbf{x}_{i-2}, \sigma_i^2, \mathbf{b}_i, \mathbf{a})$ and $\{\sigma_i^2, \mathbf{b}_i\}$ are independent between blocks, i.e., $\mathbf{x}_i$ does not depend on $\{\sigma_j^2, \mathbf{b}_j\}$ for all $j \neq i$, then (14) can be written as

$$\begin{aligned} Pr(\mathbf{x}_1, ..., \mathbf{x}_M \mid \boldsymbol{\sigma}, \boldsymbol{\beta}, \mathbf{a}) &= Pr(\mathbf{x}_1 \mid \sigma_1^2, \mathbf{b}_1, \mathbf{a}) \\ &\times Pr(\mathbf{x}_2 \mid \mathbf{x}_1, \sigma_1^2, \mathbf{b}_1, \mathbf{a}) \\ &\times \prod_{i=3}^{M} Pr(\mathbf{x}_i \mid \mathbf{x}_{i-1}, \mathbf{x}_{i-2}, \sigma_i^2, \mathbf{b}_i, \mathbf{a}), \end{aligned} \tag{15}$$

and if $N_1, N_2 \gg P$, we can approximate $Pr(\mathbf{x}_1 \mid \sigma_1^2, \mathbf{b}_1, \mathbf{a})$ and $Pr(\mathbf{x}_2 \mid \mathbf{x}_1, \sigma_1^2, \mathbf{b}_1, \mathbf{a})$ with $Pr(\mathbf{x}_1 \mid \mathbf{x}_0, \mathbf{x}_{-1}, \sigma_1^2, \mathbf{b}_1, \mathbf{a})$ and $Pr(\mathbf{x}_2 \mid \mathbf{x}_1, \mathbf{x}_0, \sigma_1^2, \mathbf{b}_1, \mathbf{a})$, respectively, where $\mathbf{x}_0$ and $\mathbf{x}_{-1}$ are the initial values of the preprocessed signal.

Then we can write (15) as

$$\begin{aligned} Pr(\mathbf{x}_1, ..., \mathbf{x}_M \mid \boldsymbol{\sigma}, \boldsymbol{\beta}, \mathbf{a}) &= \prod_{i=1}^{M} Pr(\mathbf{x}_i \mid \mathbf{x}_{i-1}, \mathbf{x}_{i-2}, \sigma_i^2, \mathbf{b}_i, \mathbf{a}) \\ &= \prod_{i=1}^{M} Pr_{\mathbf{s}_i}(\mathbf{s}_i' \mid \mathbf{s}_{i-1}, \sigma_i^2, \mathbf{b}_i) \\ &= \prod_{i=1}^{M} \mathcal{N}(\mathbf{w}_i \mid \mathbf{0}_{N_i}, \sigma_i^2 \mathbf{I}_{N_i}), \end{aligned} \tag{16}$$

where $\mathbf{0}_N$ and $\mathbf{I}_N$ are all-zero column vector with dimension of $N$ by 1 and identity matrix with dimension of $N$ by $N$, respectively.

If we let $\boldsymbol{\theta} = \{\boldsymbol{\sigma}, \boldsymbol{\beta}, \mathbf{a}\}$, i.e., the parameters to be estimated be fixed and unknowns, we can apply maximum-likelihood (ML) estimation, which aims to maximize $Pr(\mathbf{x}_1, ..., \mathbf{x}_M \mid \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. The ML approach for parameter estimation is presented next.

## III. Digital recording system Identification

### A. Audio Recording Identification Algorithm

Our approach to solve the problem of estimating the distortion impulse response $f[n]$ of the unknown system, which in our model is parameterized in $\mathbf{a}$, is to maximize (16) in terms of $\boldsymbol{\theta} = \{\boldsymbol{\sigma}, \boldsymbol{\beta}, \mathbf{a}\}$. We let the parameters $\boldsymbol{\theta} = \{\boldsymbol{\sigma}, \boldsymbol{\beta}, \mathbf{a}\}$ of (16) be unknown, but fixed.

Let $y[n]$ be the output signal of the AR filter, $\mathbf{b}_i$, in block $i$, for the input signal $x[n]$, which can be expressed as

$$y_i[n] = \sum_{q=1}^{Q_i} b_i[q] x_i[n - q] + x_i[n]. \tag{17}$$

We can apply the AP filter, $\mathbf{a}$, to obtain the excitation signal $w_i[n]$, expressed as

$$w_i[n] = \sum_{p=1}^{P} a[p] y_i[n - p] + y_i[n]. \tag{18}$$

Therefore, the expression in (3) is equivalent to

$$\mathbf{w}_i = \mathbf{y}_i + \mathbf{Y}_i \mathbf{a}, \tag{19}$$

where $\mathbf{y}_i = [y[n_i],\ y[n_i + 1],\ \ldots\ , y[n_i + N_i]]^T$ and the data matrix $\mathbf{Y}_i = [\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2, ..., \bar{\mathbf{y}}_P]$, where $\bar{\mathbf{y}}_j = [y[n_i - j], y[n_i + 1 - j], ..., y[n_i + N_i - j]]^T$ for $1 \leq j \leq P$.

Using the likelihood function (16) the ML expression is given by

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} Pr(\mathbf{x}_1, ..., \mathbf{x}_M \mid \boldsymbol{\theta}). \tag{20}$$

We consider the partial derivative of (20) with respect to $\{\sigma_i^2, \mathbf{b}_i\}$ for $1 \leq i \leq M$ and parameter $\mathbf{a}$. Since $\{\sigma_i^2, \mathbf{b}_i\}$ are independent among the blocks, the ML solution is given as follows,

$$\hat{\mathbf{b}}_i = -(\mathbf{S}_i^T \mathbf{S}_i)^{-1} \mathbf{S}_i^T \mathbf{s}_i, \tag{21}$$

$$\hat{\sigma}_i^2 = \frac{\mathbf{w}_i^T \mathbf{w}_i}{N_i}, \tag{22}$$

$$\hat{\mathbf{a}} = -\left( \sum_{i=1}^{M} \frac{\mathbf{Y}_i^T \mathbf{Y}_i}{\sigma_i^2} \right)^{-1} \sum_{i=1}^{M} \frac{\mathbf{Y}_i^T \mathbf{y}_i}{\sigma_i^2}, \tag{23}$$

where $\mathbf{y}_i = \mathbf{x}_i + \mathbf{X}_i \mathbf{b}_i$. The results are in terms of preprocessed signals $\{\mathbf{x}_1, ..., \mathbf{x}_M\}$ and the unknown parameters $\{\boldsymbol{\sigma}, \boldsymbol{\beta}, \mathbf{a}\}$, which is considered as conditionally optimal ML. The $\{\boldsymbol{\sigma}, \boldsymbol{\beta}\}$ parameters are not needed for our purpose, which can be considered as nuisance.

The ML algorithm can be performed using the following steps. Initialize $\hat{\mathbf{a}}$ arbitrarily (or by an educated guess, if side information is available), compute $\hat{\mathbf{b}}_i$, $\hat{\sigma}_i^2$ for $1 \leq i \leq M$ and $\hat{\mathbf{a}}$ iteratively to obtain at each step the conditionally ML optimal estimates of the parameters given by other estimated parameters. Stop when convergence is observed. The conditional ML (CML) algorithm is summarized below. Superscripts denote the iteration index.

**Conditionally ML algorithm (CML)**

**Input:** Initialize: $k \leftarrow 0$; generate $\hat{\mathbf{a}}$ arbitrarily.

1: $k \leftarrow k + 1$;
2: **for** $i \leftarrow 1$ to $M$
3: $\quad \hat{\mathbf{b}}_i^{(k)} \leftarrow -((\mathbf{S}_i^{(k-1)})^T \mathbf{S}_i^{(k-1)})^{-1} (\mathbf{S}_i^{(k-1)})^T \mathbf{s}_i^{(k-1)}$
4: $\quad (\hat{\sigma}_i^{(k)})^2 \leftarrow \frac{(\mathbf{w}_i^{(k-1)})^T \mathbf{w}_i^{(k-1)}}{N_i}$
5: $\quad \hat{\mathbf{a}}^{(k)} \leftarrow -\left( \sum_{i=1}^{M} \frac{(\mathbf{Y}_i^{(k)})^T \mathbf{Y}_i^{(k)}}{(\sigma_i^{(k)})^2} \right)^{-1} \sum_{i=1}^{M} \frac{(\mathbf{Y}_i^{(k)})^T \mathbf{y}_i^{(k)}}{(\sigma_i^{(k)})^2}$
6: repeat until $\|\hat{\mathbf{a}}^{(k)} - \hat{\mathbf{a}}^{(k-1)}\|_F^2 < \Delta$

**Output:** $\hat{\mathbf{a}}^{(k)}$

where $\Delta$ denotes a small scalar value, $\mathbf{s}_i^{(k-1)} = \mathbf{x}_i + \mathbf{X}_i \mathbf{a}^{(k-1)}$, $\mathbf{w}_i^{(k-1)} = \mathbf{s}_i^{(k-1)} + \mathbf{S}_i^{(k-1)} \mathbf{b}_i^{(k)}$ and $\mathbf{y}_i^{(k)} = \mathbf{x}_i + \mathbf{X}_i \mathbf{b}_i^{(k)}$. The convergence of the CML in general is not guaranteed. Convergence of the iterative procedure depends on a proper initialization point. Certainly, availability of prior information can facilitate the choice of a good initialization point, which in turn can increase the possibility for global (and fast) convergence.

Applying Baye's rule to (16) we can obtain the posterior probability density function for unknown parameters $\{\boldsymbol{\sigma}, \boldsymbol{\beta}, \mathbf{a}\}$. Since we are interested in estimating underlying recording system, parameter $\mathbf{a}$, the nuisance parameters $\boldsymbol{\sigma}$ and $\boldsymbol{\beta}$ are marginalized. A marginal maximum a posteriori (MMAP) estimate for the parameter $\mathbf{a}$ can be calculated by

$$\widehat{\mathbf{a}} = \arg\max_{\mathbf{a}} Pr(\mathbf{a} \mid \mathbf{x}_1, ..., \mathbf{x}_M, \boldsymbol{\sigma}, \boldsymbol{\beta}). \tag{24}$$

In principle, an MMAP for the unknown channel parameters, $\mathbf{a}$, can be formed by solving (24). The optimization can be performed using *deterministic* or *stochastic* optimization methods. Since sampling from the distribution in (24) is difficult, estimates of the parameter, $\mathbf{a}$, are obtained using Markov chain Monte Carlo (MCMC) routines, such as the Gibbs sampler [13]. The Monte Carlo method can the used to marginalize the nuisance parameters $\boldsymbol{\sigma}$ and $\boldsymbol{\beta}$. Similarly, we can solve (24) using the extensions of expectation-maximization (EM), stochastic approximation EM (SAEM) [14] or Monte Carlo EM (MCEM) [15] methods. All of the mentioned methods have polynomial complexity in the number of estimated parameters.

For system identification purposes, we consider the roots $\hat{\mathbf{r}}$, which correspond to the estimated coefficients, $\hat{\mathbf{a}}$, obtained by the proposed CML algorithm. The estimated roots of the stationary poles of a specific impulse response would be concentrated in particular fixed regions in a complex plain. Estimates of the poles of different distortion filters would correspond to different fixed regions and therefore can be identified. Our problem now becomes more accurate compared to those fixed regions that correspond to distortion filters and we can differentiate them according to the location of the fixed regions. If the estimated poles of two different stationary distortion filters have their fixed regions very close to each other then they correspond to the same distortion filter, otherwise different. The sets of roots of the poles of similar and different filters as an example are shown in Figs. 5 and 6. For the case of similar microphones, as can be seen from Fig. 5, most of the poles are almost overlapping with each other indicating same microphone is detected while for the case of different microphones, as can be seen from Fig. 6, the poles are further away from each other, which indicates that different microphones have been used to perform the recordings.

### *B. Nearest Neighboring Based Clustering*

The idea is to cluster poles of each distortion impulse response according to their fixed regions. We develop a simple neighboring rule that takes two sets of poles and tries to find the fixed regions of each set. In particular, we first construct the set of estimated roots, $\mathcal{R}_l \in \{\hat{\mathbf{r}}_l(u); u \in \mathcal{U} = \{1, ..., U\}\}$ for $1 \leq l \leq L$, where $L$ is the number of distortion impulse response being estimated and $U$ is the number of estimates of parameter $\hat{\mathbf{a}}$ for a given $l$, which is returned by the CML algorithm. Before we describe the algorithm let us define a histogram in the complex plane as

$$\mathbf{H}_l(y, x) = |\{\hat{\mathbf{r}}_l(p, u) \in x + jy \; \forall p \in \mathcal{P}, u \in \mathcal{U}\}|, \tag{25}$$

where $\mathbf{H}_l \in \mathbb{N}^{Y \times X}$, $1 \leq x \leq X$, $1 \leq y \leq Y$, $j$ is imaginary number, and $X, Y \in \mathbb{N}$. The higher the value of a point in the histogram means the more likely that it corresponds

to a certain fixed regions of an impulse response. It can be concluded from the observation of poles in complex plane that $\mathbf{H}_l$ contains many zeros. Now, let us define a sum over a small region of $\mathbf{H}_l$ histogram as follows

$$s_l(y,x) = \sum_{y',x' \in \mathcal{L}_{y,x}} \mathbf{H}_l(y',x'), \tag{26}$$

where $\mathcal{L}_{y,x} = \{y', x' | y - r \leq y' \leq y + r, x - r \leq x' \leq x + r\}$, $r \in \mathbb{N}$, which defines a square around the center point $(y,x)$ in $\mathbf{H}_l$. This sum represents the number of points lying in a complex plane around the center point $(y,x)$. The value of $r$ can effect the decision of the fixed region and hence, can be properly chosen to serve our purpose. The histogram of the poles corresponding to a particular distortion impulse response have most of its *reliable* estimated poles located at small region and *unreliable* estimated poles are in a larger region. The convergence estimates that are concentrated in a fixed region are close to the true poles. Our task now is to locate those *reliable* estimated poles in a histogram that belong to the fixed regions of the poles as close as possible.

The basic idea behind our approach is to search those *reliable* estimated poles among all the poles returned by the proposed CML algorithm. However, in audio recording system identification problem, the locating of such reliable estimates is not an easy task, due to the lack of knowledge of true parameters and priori information. The *reliable* estimated poles can be identified by examining the $s_l(y,x)$ defined by (26). The $s_l(y,x)$ not only takes into consideration the $\mathbf{H}_l(y,x)$ value but around that point, which is a reasonable measure. In order to decide whether $\mathbf{H}_l(y,x)$ belongs to a fixed region we compute $s_l(y,x)$ and it compare to an upper threshold $h_l$, which is a design parameter. If $\mathbf{H}_l(y,x) > h_l$ then we can consider $\mathbf{H}_l(y,x)$ belongs to the fixed region of the distortion impulse response.

The comparison of roots in complex plane in terms of fixed regions of impulse responses can be performed in pairwise fashion. For example, let us take $\mathbf{H}_1$ and $\mathbf{H}_2$ to make the comparison. For each point in $\mathbf{H}_1$ we compute $s_1(y,x)$ and the corresponding $s_2(y,x)$, if both $s_1(y,x) > h_1$ and $s_2(y,x) > h_2$ are satisfied for each $(y,x)$ then the stationary poles belong to the same fixed region, hence, the impulse response is the same, otherwise, it is different. A reasonable measure is utilized for a reliable identification, which is summarized in *Criterion 1* below.

*Criterion 1: Classifying histogram of estimated poles*

If both $s_1(y,x) > h_1$, $s_2(y,x) > h_2$, $\forall \{y,x\} \in \mathcal{L}_{y,x}$ and $\mathbf{H}_1(y,x) < \bar{h}_1$, $\mathbf{H}_2(y,x) < \bar{h}_2$ $\forall \{y,x\} \notin \mathcal{L}_{y,x}$ conditions are satisfied then declare them similar or "1". Here $\bar{h}_1$ and $\bar{h}_2$, lower thresholds are design parameters. Otherwise, if the condition are not satisfied that means the two histograms $\mathbf{H}_1$ and $\mathbf{H}_2$ do not have common reliable estimated poles and hence, they have different distortion impulse response functions, declare them not similar or "2".

To address our goal of identifying the recording systems, we now present a complete description of the algorithm. Assume we are given $L$ recorded audio source files from different/similar microphone recording systems. We take each audio files (e.g., MP3, WAV) and preprocess to obtain $\mathbf{x}_l$ for $1 \leq l \leq L$. For each $\mathbf{x}_l$, we run the proposed CML algorithm for $U$ times to produce stationary poles $\mathbf{r}_l(u)$, $1 \leq u \leq U$. We create $\mathbf{H}_l$ from $\mathbf{r}_l$, $1 \leq l \leq L$ then utilizing the *Criterion 1*, we can identify the *reliable* poles and classify those estimated poles according to their difference/similarity. Eventually, this algorithm classifies each recorded audio files according to their distortion impulse response of microphone/recording system.

The ML-nearest neighbor based audio recording identification algorithm (ML-NNA) that returns expected classes $\hat{C}$ of $L$ audio recordings is presented below.

**ML-Nearest Neighbor Algorithm (ML-NNA)**

**Input:** Initialize: $\hat{C} \leftarrow \{1, ..., L\}$ classes
1: **for** $l \leftarrow 1$ to $L-1$
2: **for** $m \leftarrow l+1$ to $L$
3: run CML to obtain $\{\hat{\mathbf{a}}_l(u); u \in \mathcal{U}\}$
4: compute roots $\hat{\mathbf{r}}_l(u)$ for $1 \leq u \leq U$
5: run CML to obtain $\{\hat{\mathbf{a}}_m(u); u \in \mathcal{U}\}$
6: compute roots $\hat{\mathbf{r}}_m(u)$ for $1 \leq u \leq U$
7: compute $\mathbf{H}_l$ and $\mathbf{H}_m$, from $\hat{\mathbf{r}}_l$ and $\hat{\mathbf{r}}_m$ by (25)
8: **if** *Criterion 1* is satisfied; then $\hat{C}(m) \leftarrow \hat{C}(l)$
**Output:** $\hat{C}$

Integers in $\hat{C}$ represents the classes; if all the integers are the same that means the audio recording files are recorded with similar recording system and if the integers are not the same that means some or all the audio files are recorded using different recording systems.

## IV. Experimental Results

In order to assess the performance of the proposed algorithm for identification of device similarity, experiments were conducted in a laboratory environment. Speech recordings from 15 different speakers using two different microphones, $L$-Logitech Stereo Headset H111 and $P$-Philips PH62080 Unidirectional Microphone are used. Audio is recorded as MP3 mono audio files in the same laboratory using 44.1 kHz sampling rate and 16 bit quantization, and we assume that the acoustical environment does not change.

The estimate $\hat{\mathbf{a}}$ obtained from the ML algorithm is based on the assumption that we have perfect knowledge of the exact model and parameters. However, the nature of identifying the underlying recording system problem prohibits us to have any knowledge of the model parameters (e.g., $P, Q_i, N_i, 1 \leq i \leq M$). Therefore, we look into a heuristic methods to choose proper model parameters for identification purposes. As an example, $AR(80)$ BSAR process can be used to model speech signal and $AP(68)$ process to model distortion filter. To model quasi-stationary signal $M = 500$ blocks of length $N_i = 350$ can be used, note that natural speech is locally stationary in every $25ms$ [16].

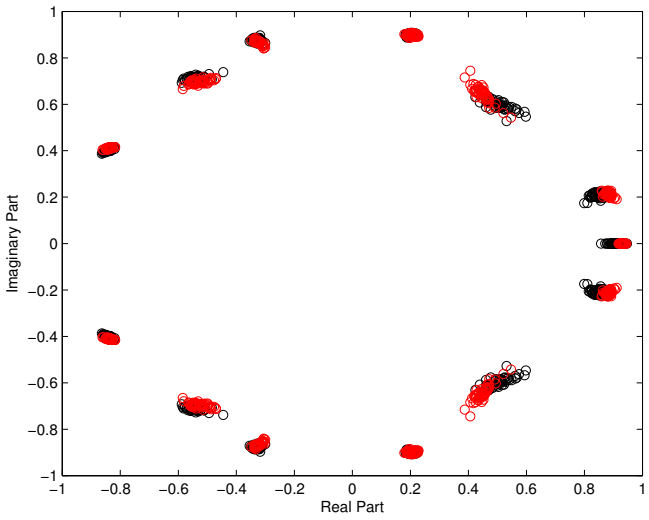


Fig. 5: Philips microphone - proposed CML.

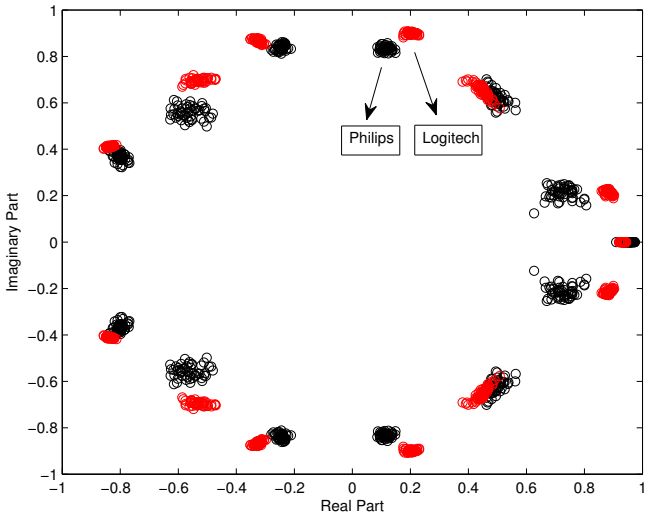


Fig. 6: Logitech and Philips microphones - proposed CML.

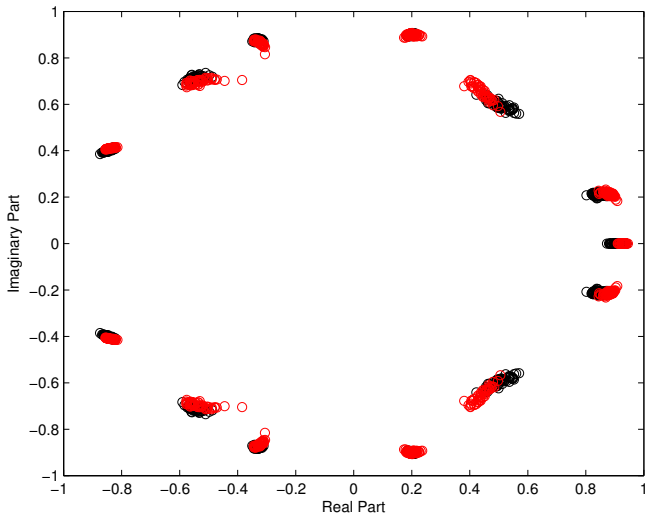


Fig. 7: Philips microphone - SAEM.

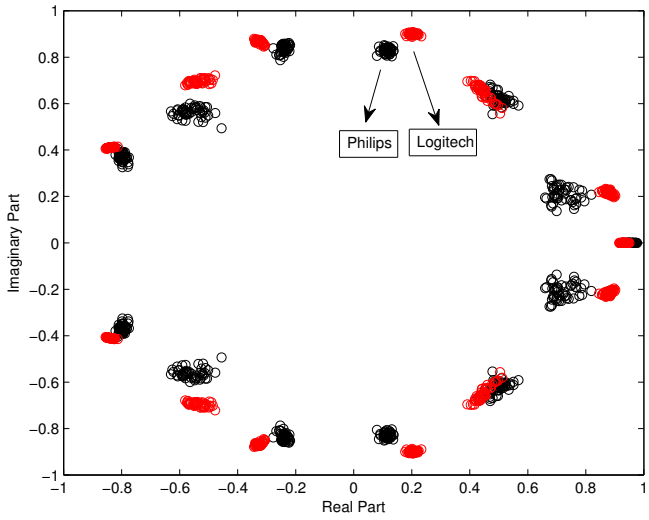


Fig. 8: Logitech and Philips microphones - SAEM.

Through extensive experimentation, we discovered that using the $AR(2)$ BSAR to model the speech signal and $AP(13)$ to model the distortion filter serves well for our identification purposes. In our experiments, we selected $M = 250$ blocks of length $N_i = 940$.

We recorded $400$ MP3 recordings using $15$ different speakers for each microphone. In this manuscript, we illustrate the performance of the proposed CML algorithm that is compared to Gibbs sampler, SAEM, and MCEM algorithms for classifying the same microphone ($P$ and $P$) and different microphones ($P$ and $L$).

The proposed CML algorithm in Figs. 5 and 6, demonstrates the classifications of two different speakers recorded with the same Philips microphone $P$ and one speaker using Logitech $L$ and Philips $P$ microphones, based on MP3 files alone. We can observe from Fig. 5 using similar filters the sets of roots are closer to each other compared to the one using two microphones, as shown in Fig. 6. Similar performances can be seen for SAEM, MCEM and Gibbs Sample, shown in Figs. 7 and 8, Figs. 9 and 10, Figs. 11 and 12, respectively.

In order to have better judgment on the performance of each estimators, we measure the accuracy of the classification algorithm using total of $1600$ MP3 files about $30$s long each that are recorded using the two microphones. In Table I, we compared the accuracy of the proposed CML classification algorithm with SAEM, MCEM, and Gibbs estimators. The accuracy of classifying the same microphone $P$, and detecting difference between the two microphones $P$ and $L$ using the proposed CML estimator is slightly higher than those using the SAEM, MCEM and Gibbs estimators.

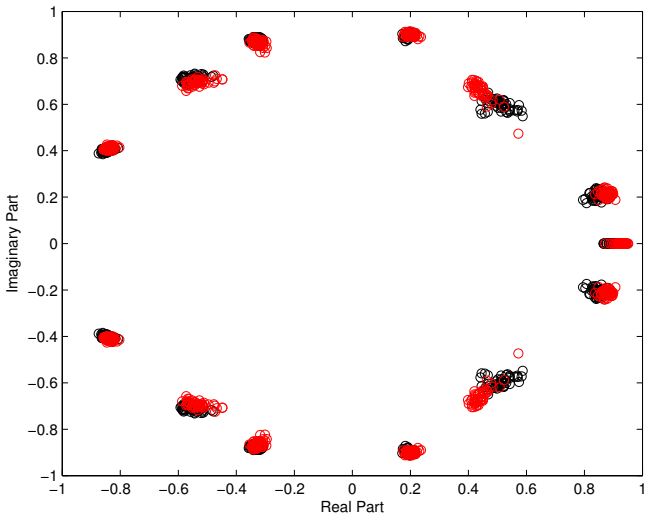


Fig. 9: Philips microphone - MCEM.

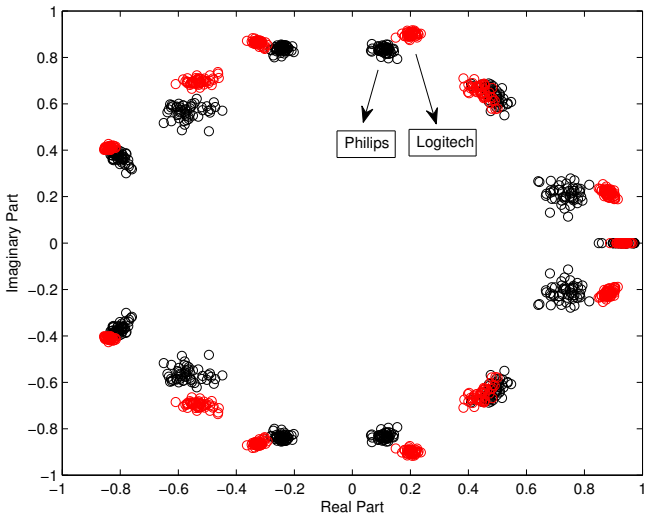


Fig. 10: Logitech and Philips microphones - MCEM.

TABLE I: Accuracy of different estimators

| Estimators | Rec. Systems | Accuracy (%) |
|---|---|---|
| **Proposed CML** | $P$ & $P$ | **99.28** |
| SAEM | $P$ & $P$ | 97.46 |
| MCEM | $P$ & $P$ | 96.23 |
| Gibbs | $P$ & $P$ | 93.46 |
| **Proposed CML** | $L$ & $P$ | **99.51** |
| SAEM | $L$ & $P$ | 97.16 |
| MCEM | $L$ & $P$ | 95.89 |
| Gibbs | $L$ & $P$ | 94.24 |

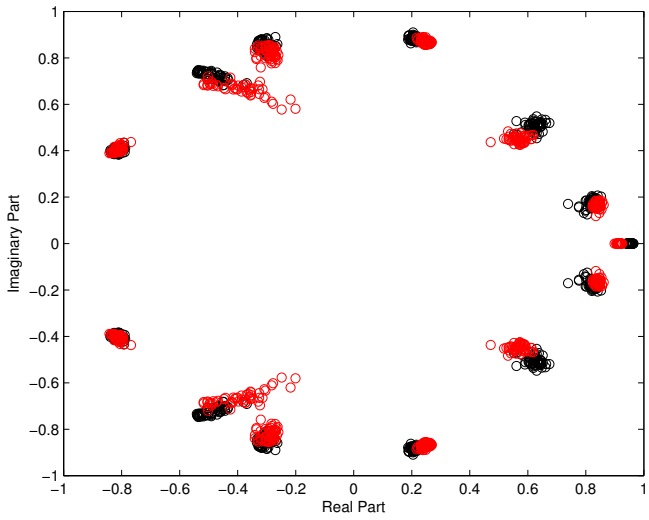


Fig. 11: Philips microphone - Gibbs samppler.

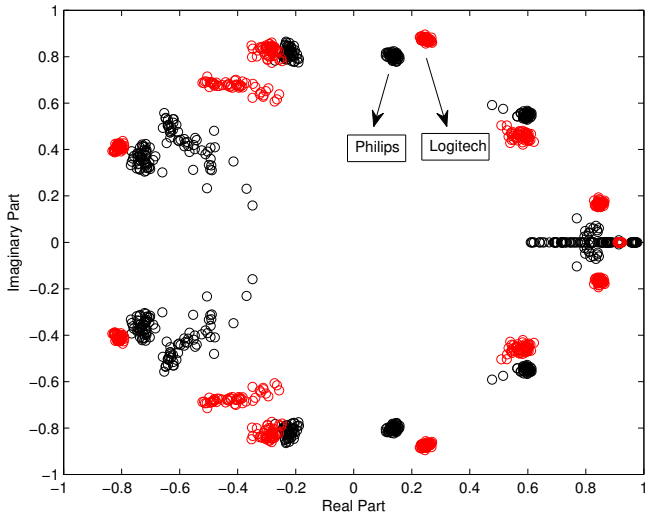


Fig. 12: Logitech and Philips microphones - Gibbs sampler.

## V. Conclusion

We presented a novel digital recording system identification system where no prior information is provided and it is based on audio files alone. The non-stationary audio source is modeled as a block stationary AR (BSAR) and cascaded impulse response of the recording system, which accordingly is modeled as an all-pole (AP) infinite impulse response (IIR) filter. We then proposed a conditionally maximum-likelihood (CML) algorithm to estimate the coefficients of the unknown AP distortion filter. To determine the difference or similarity of the recording system being analyzed, namely the distortion filter from CML, we developed a novel nearest neighborhood algorithm to cluster poles of the AP filter. Experimental results demonstrate high accuracy, over $99.2\%$, in identification of the recording devices using the proposed scheme.

## REFERENCES


[1] C. Kotropoulos, and S. Samaras, "Mobile phone identification using recorded speech signals," in *Proc. IEEE 19th International Conference on Digital Signal Processing (DSP)*, Hong Kong, pp. 586-591, Aug. 2014.

[2] D. Garcia-Romero, and C. Y. Espy-Wilson, "Automatic acquisition device identification from speech recordings," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, Texas, USA, pp. 1806-1809, March 2010.

[3] Y. Panagakis, and C. Kotropoulos, "Telephone handset identification by feature selection and sparse representations," in *Proc. IEEE International Workshop on Information Forensics and Security (WIFS)*, Tenerife, Spain, pp. 73-78, Dec. 2012.

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

[5] C. Hanilçi, F. Ertaş, T. Ertaş, and Ö. Eskidere, "Recognition of brand and models of cell-phones from recorded speech signals," *IEEE Trans. on Information Forensics and Security*, vol. 7, no. 2, pp. 625-634, April 2012.

[6] L. Zou, Q. He, J. Yang, and Y. Li, "Source cell phone matching from speech recordings by sparse representation and KISS metric," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shangai, China, pp. 2079-2083, March 2016.

[7] T. Qin, R. Wang, D. Yan, L. Lin, "Source Cell-Phone Identification in the Presence of Additive Noise from CQT Domain," *Information*, vol. 9, no. 8, pp. 205, Aug. 2018.

[8] S. Qi, Z. Huang, Y. Li, and S. Shi, "Audio recording device identification based on deep learning," in *Proc. IEEE International Conference on Signal and Image Processing (ICSIP)*, Beijing, China, pp. 426-431, Aug. 2016.

[9] Simon Haykin, *Unsupervised Adaptive Filtering Volume 2: Blind Deconvolution*, Canada: John Wiley and Sons Inc., 2000.

[10] O. Shalvi, and E. Weinstein, "System identification using nonstationary signals," *IEEE Trans. Signal Processing*, vol. 44, no. 8, 2055-2063, Aug. 1996.

[11] J. R. Hopgood, and Peter J. W. Rayner, "Blind single channel deconvolution using nonstationary signal processing," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, no. 5, pp. 476-488, Sept. 2003.

[12] Mark Kahrs, and Karlheinz Brandenburg, *Applications of Digital Signal Processing to Audio and Acoustics*, Boston, MA: Kluwer Academic Publishers, 2002.

[13] Alan E. Gelfand, "Gibbs Sampling," in *Journal of the American Statistical Association*, Vol. 95, No. 452, pp. 1300-1304, Dec 2000.

[14] B. Delyon, M. Lavlielle, and E. Moulines, "Convergence of a stochastic approximation version of the EM algorithm," in *Annals of Statics*, Vol. 27, No. 1, pp. 94-128, 1999.

[15] G. Wei and M. Tanner, "A Monte-Carlo implementation of the EM algorithm and the Poor's Man's data augmentation algorithm," in *Journal of the American Statistical Association*, Vol. 85, pp. 699-704, 1990.

[16] W.K. Ma, T.H. Hsieh, C.Y. Chi, "DOA estimation of quasi-stationary signals via Khatri-Rao subspace," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, pp. 2165-2168, April 2009.