ABSTRACT

Title of dissertation: INTRINSICALLY EMBEDDED SIGNATURES FOR MULTIMEDIA FORENSICS

Adi Al Hajj Ahmad, Doctor of Philosophy, 2016

Dissertation directed by: Professor Min Wu
Department of Electrical and Computer Engineering

This dissertation examines the use of signatures that are intrinsically embedded in media recordings for studies and applications in multimedia forensics. These near-invisible signatures are fingerprints that are captured *unintentionally* in a recording due to influences from the environment in which it was made and the recording device that was used to make it. We focus on two types of such signatures: the Electric Network Frequency (ENF) signal and the flicker signal.

The ENF is the frequency of power distribution networks and has a nominal value of 50Hz or 60Hz. The ENF fluctuates around its nominal value due to load changes in the grid. It is particularly relevant to multimedia forensics because ENF variations captured intrinsically in a media recording reflect the time and location related properties of the respective area in which it was made. This has led to a number of applications in information forensics and security, such as time-of-recording authentication/estimation and ENF-based detection of tampering in a recording.

The first part of this dissertation considers the extraction and detection of

the ENF signal. We discuss our proposed spectrum combining approach for ENF estimation that exploits the presence of ENF traces at several harmonics within the same recording to produce more accurate and robust ENF signal estimates. We also explore possible factors that can promote or hinder the capture of ENF traces in recordings, which is important for a better understanding of the real-world applicability of ENF signals.

Next, we discuss novel real-world ENF-based applications proposed through this dissertation research. We discuss using the embedded ENF signal to identify the region-of-recording of a media signal through a pattern analysis and learning framework that distinguishes between ENF signals coming from different power grids. We also discuss the use of the ENF traces embedded in a video to characterize the video camera that had originally produced the video, an application that was inspired by our work on flicker forensics.

The last part of the dissertation considers the flicker signal and its use in forensics. We address problems in the entertainment industry pertaining to movie piracy related investigations, where a pirated movie is formed by camcording media content shown on an LCD screen. The flicker signature can be inherently created in such a scenario due to the interplay between the back-light of an LCD screen and the recording mechanism of the video camera. We build an analytic model of the flicker, relating it to inner parameters of the video camera and the screen producing the video. We then demonstrate that solely analyzing such a pirated video can lead to the identification of the video camera and the screen that produced the video, which can be used as corroborating evidence in piracy investigations.

INTRINSICALLY EMBEDDED SIGNATURES FOR MULTIMEDIA FORENSICS

by

Adi Al Hajj Ahmad

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2016

Advisory Committee:
Professor Min Wu, Chair/Advisor
Professor Gang Qu
Professor Behtash Babadi
Dr. Gwenaël Doërr
Professor Miao Yu

I dedicate this work to my father, for always having my back, to my mother, for all the warmth she brings to my world, and to Sanaa, for her knack of telling me things I need to hear.

Acknowledgments

I want to start by thanking my advisor and mentor Prof. Min Wu who has been my guide and teacher throughout this whole journey, and without whom this dissertation would not have been possible. I learned from her to be adaptable and to always be willing to re-invent myself and forge forward. I learned from her that there is no limit to where the work can go; there are always new possibilities to explore, interesting questions that can be addressed, and more than one way to go from good to better. I want to thank her for her support and genuine interest in my growth as a researcher and for all the opportunities that she has helped make possible for me during my time at UMD, and for all the ones to come.

I want to thank Dr. Gwenaël Doërr who has been a second advisor to me since being my mentor at my internship at Technicolor in Summer 2014, one of my most fulfilling working experiences to date. I am very happy that we were able to continue collaborating over the remainder of my time as a PhD student. He has definitely allowed me to broaden the research work that I did and I'm very grateful for his support and guidance, and for being such an instrumental part of my PhD journey.

I want to thank Prof. Gang Qu, Prof. Behtash Babadi and Prof. Miao Yu for serving on my dissertation committee and for their thoughtful comments and discussions during the defense. I want to thank the professors who I have had the privilege to be a teaching assistant for and who have helped shape my ideas around teaching and education: Prof. Anthony Ephremides, Prof. Prakash Narayan, Prof. Andre Tits, Prof. Eyad Abed, and Prof. Steve Marcus. I also want to thank

the stellar staff in the ECE department at UMD who I think deserve plenty of appreciation.

During my PhD training, I had the opportunity to work on two internships that helped me to broaden my horizon and research training. From Technicolor R&D France, I want to thank, in addition to Dr. Doërr, Dr. Séverine Baudry, Mr. Bertrand Chupeau, and Mr. Mario de Vito. From AT&T Research Labs, I want to thank Dr. Behzad Shahraray and Dr. Zhu Liu.

I want to thank my research teammates over the past five years. I have benefited plenty from the experiences of those more senior than me: Wei-Hong Chuang, Ravi Garg, and Hui Su; and it's been great being able to share the journey and work together with Chau-Wai Wong and Abbas Kazemipour. I also want to thank my friend and fellow Terp Andrew Berkovich whom I was able to collaborate on a research project with.

At this point, I want to thank the National Science Foundation and the Northrop Grumman Foundation for grants that supported my research.

I want to thank the friends who have enriched my world and changed my life in small ways and big ways over the past few years. As much as I would enjoy listing these people by the order in which I like them, I don't want to break any hearts by my ordering or by forgetting somebody, so this is my cop-out: If you're reading this to see if I mentioned you, then thank you for caring, and thank you for being there.

I want to end by thanking my incredible family for being who they are, and for all the transoceanic conversations: Mostafa and Meyye; Hassan, Joujou, and Sanaa; the legendary Teta Em Hassoun; Jana and Jad; and my parents. No words will really be enough so I will just stop at this: B7ebkon.

Table of Contents

List of Tables viii

List of Figures ix

1 Introduction 1
1.1 ENF Forensics . . . 2
1.2 Flicker Forensics . . . 6
1.3 Dissertation Organization . . . 8

2 Extraction of ENF Signals 10
2.1 Chapter Introduction . . . 10
2.2 Reference power recordings . . . 11
2.3 Overview of ENF Estimation Approaches . . . 14
2.3.1 Time-domain approach . . . 15
2.3.2 Non-Parametric Frequency-Domain Approaches . . . 16
2.3.3 Parametric Frequency-domain Approaches . . . 19
2.4 Comparison of Estimation Approaches . . . 23
2.4.1 Experiments on Synthetic Signals . . . 24
2.4.2 Experiments on Power Grid Data Set . . . 27
2.4.3 Matching ENF Signals from Audio and Power . . . 28
2.5 Proposed Spectrum Combining Estimation Approach . . . 30
2.5.1 Model and Proposed Approach . . . 31
2.5.1.1 Problem Formulation . . . 31
2.5.1.2 Determining Spectral Combining Weights . . . 34
2.5.1.3 Instantaneous ENF Estimation . . . 34
2.5.2 Experiments and Results . . . 35
2.5.2.1 Experimental Set-up . . . 35
2.5.2.2 Comparison Framework . . . 37
2.5.2.3 Results and Discussions . . . 38
2.6 Chapter Summary . . . 40

3 Factors Affecting Capture of ENF Traces in Recordings 42
3.1 Chapter Introduction . . . 42
3.2 Overview of previous work . . . 44
3.3 Explorations on environment-related factors . . . 46
3.3.1 Effect of wave interference . . . 47
3.3.2 Effect of recorder movement . . . 51
3.3.3 ENF traces across different locations . . . 53
3.4 Explorations on device-related factors . . . 57
3.4.1 Computation of local SNR estimate . . . 58
3.4.2 Experiments and Results . . . 60
3.4.2.1 Difference in ENF strength . . . 60
3.4.2.2 Difference in ENF harmonics captured . . . 62
3.4.2.3 Discussion . . . 63
3.5 Chapter Summary . . . 64

4 ENF-Based Region-of-Recording Identification for Media Signals 66
4.1 Chapter Introduction . . . 66
4.2 Location-Dependent ENF Database . . . 68
4.2.1 Database Description . . . 69
4.2.2 Comparison of ENF Signals from Different Grids . . . 70
4.3 Multiclass Region-of-Recording Classification . . . 73
4.3.1 Feature Extraction and Analysis . . . 74
4.3.2 SVM Classifier . . . 78
4.3.3 Results and Discussions . . . 81
4.4 Noise Adaptation using Multi-Conditional Learning . . . 87
4.4.1 System Model . . . 88
4.4.2 Experimental Setup . . . 90
4.4.3 Results and Discussions . . . 91
4.5 Further Discussions . . . 93
4.5.1 Dimensionality Reduction . . . 93
4.5.2 Grids with Varying Profiles . . . 95
4.6 Chapter Summary . . . 98

5 Exploiting Power Signatures for Camera Forensics 101
5.1 Chapter Introduction . . . 101
5.2 Model and Proposed Approach . . . 103
5.2.1 Modeling the Captured Signal . . . 103
5.2.2 Proposed Approach . . . 106
5.2.2.1 Vertical Phase Method . . . 106
5.2.2.2 Adapting to a Practical Setting . . . 108
5.3 Experiment and Results . . . 110
5.3.1 Experimental Set-up . . . 110
5.3.2 Results and Discussions . . . 112
5.4 Chapter Summary . . . 113

6 Flicker Forensics for Camcorder Piracy 114
6.1 Chapter Introduction . . . 114
6.2 Modeling the Flicker Signal . . . 117
6.2.1 LCD Screens and Back-light . . . 118
6.2.2 Light Integration with a Camcorder . . . 119
6.2.3 Sampling with a Rolling Shutter . . . 120
6.2.4 Discussion . . . 122
6.3 Estimation of the Vertical Radial Frequency . . . 124
6.3.1 Flicker Phase Method . . . 125
6.3.2 Exploiting the Harmonics of the Flicker . . . 128
6.3.3 Content Cancellation Method . . . 132
6.4 Pirate Devices Identification . . . 134
6.4.1 Extracting Internal Parameter Values from Devices . . . 135
6.4.1.1 LCD Screen Back-light Frequency . . . 136
6.4.1.2 Camcorder Read-out Time . . . 137
6.4.2 Experimental Results . . . 138
6.5 Extension to Flicker Profile Recovery . . . 140
6.5.1 Recovering the Flicker Profile . . . 141
6.5.2 Revealing LCD Back-light Technology . . . 144
6.6 Chapter Summary . . . 149

7 Conclusions and Future Perspectives 151

Bibliography 155

List of Tables

2.1 Correlation coefficient values between power and audio ENF signals . 29
2.2 Sample values of spectral combining weights w_k's 37

3.1 Sample of factors affecting capture of ENF traces in audio recordings 45
3.2 Wavelengths of sound waves at different ENF harmonics 47
3.3 SNR estimates in recordings made by Olympus and B&K 62
3.4 SNR estimates in recordings made by B&K and Max4466 63

4.1 Proposed feature components 75
4.2 Description of the trained SVM systems 80
4.3 Accuracies for different systems averaged over 20 rounds. 81
4.4 Confusion matrix for power ENF testing data on System I 83
4.5 Confusion matrix for audio ENF testing data on System II 83
4.6 Confusion matrix for power and audio ENF testing data on System III 84
4.7 Description of trained multi-conditional systems 91
4.8 Results of testing on multi-conditional systems for seven grids 92
4.9 Accuracies for classification systems with dimensionality reducion . . 94
4.10 Confusion matrix for testing Lebanon data in the new setting 96

5.1 Cameras used in our experiments 112
5.2 Estimated T_{ro} values of considered videos using our approach 112

6.1 Parameters used in the modeling of the flicker signal 117
6.2 Benefit of using harmonics with a toy example 131
6.3 LCD screens used in our experiments 137
6.4 Camcorders used in our experiments 138
6.5 Relative estimation error with and without using harmonics 139
6.6 Rank of the correct screen–camcorder pair using proposed approaches 140

List of Figures

1.1 Spectrograms of power & audio signals, and the extracted ENF signal 4
1.2 Flicker artifact when camcording a gray frame on an LCD screen . . . 7

2.1 Framework of the FNET system 12
2.2 Reference recording measurement set-up and circuit schematic 13
2.3 Synthetic Signals: Correlation Coefficient vs. SNR for very short frame sizes 25
2.4 Power Grid Data Set: Correlation Coefficient vs. SNR for 1-sec frames . . 28
2.5 Spectrogram strips about the ENF harmonics for two recording sets. . 31
2.6 ROC curves for matching audio ENF to reference power ENF 39

3.1 Diagram showing result of interference of two identical sound waves. . 48
3.2 Set-up and results of wave interference verification experiment. 50
3.3 Spectrogram strips and ENF for signal made by a moving recorder . . 52
3.4 Trajectory of 1-hour recording made. 54
3.5 Spectrogram strips for 1-hour recording made. 55
3.6 Correlation coefficient values obtained for all location cases 56
3.7 Spectrogram strips about ENF harmonics: Olympus and B&K case . 61
3.8 Spectrogram strips about ENF harmonics: B&K and Max4466 case . 63

4.1 Sample ENF signals extracted from power recordings: 60Hz case. . . 70
4.2 Sample ENF signals extracted from power recordings: 50Hz case . . . 71
4.3 The number of available examples (ENF signal segments) per grid. . . 72
4.4 Sample feature values for training data instances: 60Hz case. 77
4.5 Sample feature values for training data instances: 50Hz case 77

5.1 Timing of rolling shutter sampling 102
5.2 Sample results of applying *vertical phase method* on a video 107

6.1 Flicker artifact when camcording a gray frame on an LCD screen . . . 116
6.2 Simplified model of the back-light signal 118

6.3 Rolling shutter illustrated 121
6.4 DFT magnitude of one row average for several camcorded samples . . 126
6.5 Samples of the evolution of the estimated flicker phase 127
6.6 Sample spectrum of signal through content cancellation method . . . 134
6.7 Custom-made light sensing probe 136
6.8 Sample screenshots of camcorder video sequences 139
6.9 Sample temporal back-light signals of LED and CCFL screens 145
6.10 Illustration of row luminance signature in a toy example 145
6.11 Illustration of flicker profiles extracted from camcorded videos 146
6.12 Examples of polynomial fitting for row lumincance signatures 148
6.13 Histogram of feature values computed for several pirated copies . . . 148

Chapter 1

Introduction

Multimedia forensics addresses a series of questions about a piece of multimedia recording of interest. Is it authentic? Has it been tampered with? When was it recorded? Where was it recorded? What devices were used for recording it? The answers to such questions are valuable in shedding light on the recording's integrity, history, and origins. They also provide important evidence and assurance in crime solving, journalism, infrastructure monitoring, smart grid management and other informational operations.

We have carried out research on signatures that are embedded in multimedia recordings in an intrinsic way due to the environment in which the recordings were made. We have focused on developing approaches to extract such signatures accurately and on examining real-world applications in forensics and security that they can be used towards. The two main signatures addressed in this dissertation are the Electric Network Frequency (ENF) signal and the flicker signal.

1.1 ENF Forensics

The Electric Network Frequency (ENF) signal has emerged in recent years as an important tool for multimedia forensics. The ENF is the frequency of power distribution networks. It has a nominal value of 60Hz in the United States, and 50Hz in most other parts of the world. The ENF is not typically constant at this nominal value, but rather fluctuates around it due to load changes across the power grid. These variation trends are almost identical in all locations of the same grid at a given time due to the interconnected nature of the grid [1]. We define the *ENF signal* as the changing instantaneous value of the ENF over time.

An important property of the ENF that makes it particularly relevant to multimedia forensics is that ENF fluctuations are often intrinsically embedded in audio or video recordings made in areas where there is electrical activity. It has been shown that ENF traces can be captured by recorders connected to the power mains or by battery-powered recorders. In audio recordings, this has been attributed to electromagnetic influences, the ambient acoustic mains hum and background noise emitted by mains-powered equipment in the vicinity of the recorder being used [2, 3]. In video recordings, the captured ENF traces have been attributed to the near-invisible flickering of electric lighting [4].

The process of working with the ENF signal can be seen as a two-stage operation. The first stage requires the extraction of the ENF signal from media recordings. The second stage involves using the extracted ENF signal to address a real-world problem. Government agencies and research institutes of many countries have con-

ducted ENF-related research and development work, with a large emphasis on the extraction and detection of the ENF signal, and examining and improving applications that it can be used towards, in forensics and security, and in other fields as well [5–8]. These agencies and institutes include academia and government in Romania [2,9–12], Poland [13], Denmark [14–16], the United Kingdom [17–19], the United States [4–8,20–38], the Netherlands [39], Brazil [40–43], China [44,45], Egypt [46,47], Israel [48], Germany [3], and Singapore [49].

When ENF traces are found in a signal, a typical approach to extract the changing values of the ENF over time is to divide the signal into consecutive frames and use a frequency estimation approach on each frame to determine its dominant frequency close to the nominal 50/60Hz ENF value. Concatenating these frequency estimates together gives an estimate of the captured ENF signal.

ENF traces captured in different recordings may be of varying quality. The ENF traces captured in an audio/video recording will typically exhibit a low signal-to-noise ratio (SNR), especially when compared to the ENF traces found in recordings that are recorded directly from the power mains, which we define as *reference power recordings*. These reference recordings contain strong ENF traces and are very valuable in ENF studies. The variations in the ENF signal extracted from two simultaneously recorded signals should be very similar over the same time duration. Figure 1.1 shows an example of the spectrograms of two simultaneously recorded signals showing similar ENF variations, one a reference power signal and one a regular audio signal. The figure also shows the extracted ENF signal from the audio recording. It is worth noting that the ENF traces may appear not just around the

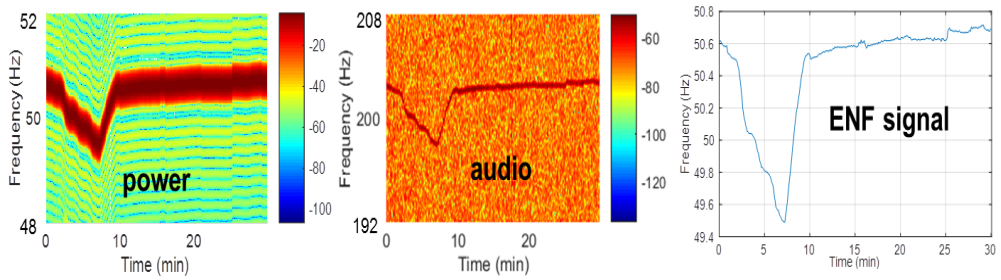


Figure 1.1: (a) and (b) show the spectrograms of simultaneous recorded power and audio signals from Lebanon, where the nominal ENF value is 50Hz. In this case, the ENF traces were captured in the audio recording strongly at around 200Hz, a harmonic of the nominal ENF value. (c) shows the ENF signal extracted from the audio recording.

nominal ENF value but around harmonics of the nominal ENF value as well, which is the case of the audio recording in this example.

The similarity observed in the variations between ENF signals extracted from simultaneously recorded signals motivated one of the early proposed ENF-based forensics applications: using the ENF to authenticate or identify the time-of-recording of a signal [2]. A database of reference ENF signals, with known times-of-recording, can be used a guide against which the ENF signal extracted from a media recording is compared to find the time duration where a match between the reference and media ENF patterns exists, thus identifying its time-of-recording. Other proposed ENF-based applications include detecting tampering/modification in media files [41–43, 50], multimedia synchronization [7, 8], characterizing the video camera producing an ENF-containing video [51], and determining the location-of-recording among different power grids [34, 37] and potentially within a grid [27, 29].

As estimating ENF signals is typically the first step in most ENF-based applications, the validity of the final results of this application will strongly depend on the accuracy of the extracted ENF signals. To this end, in the early part of

this dissertation, we focus on studying approaches to extract the ENF signals from media recordings. We also propose a novel approach, the spectrum combining approach, which exploits the presence of the ENF traces at different ENF harmonics in a media recording to obtain a more robust ENF signal estimate. Afterwards, we present our study on examining factors that can promote or hinder the capture of ENF traces in media recordings. A better understanding of these factors would lead to a stronger understanding of the real-world applicability of the ENF signal and can influence the way we use it.

We later discuss two novel ENF-based applications that we have proposed through the work done for this dissertation. The first application exploits the observed statistical differences in ENF variations between different grids to use the ENF as a signature for the grid-of-origin of a media recording. We investigate features based on these differences and use them in a multiclass machine learning framework to identify the grid-of-origin of an ENF-containing signal. The second application we discuss uses the ENF captured in a video to characterize the video camera that originally produced the video. This application was actually inspired by our work on the flicker signal, the second intrinsically embedded signature we focus on in this dissertation.

1.2 Flicker Forensics

The flicker signal is a signature that is created in videos made by camcording content shown on a Liquid Crystal Display (LCD) screen, resulting from the interplay between the camcorder and the screen. Work based on this signature addresses forensic problems in the entertainment industry for movie piracy related investigations and deterrence.

Movie piracy is still a major concern for the entertainment industry today, with the illegal releases of pirated copies before theatrical or Blu-ray release of movies holding the potential to significantly harm revenues. To address this risk, content owners have been relying on cryptography-based content to prevent consumers from easily accessing the media content [52]. Nevertheless, this type of protection needs to be lifted eventually, which leaves the content vulnerable for a pirate to place a camera in front of a screen and record a pirated copy.

A second line of defense behind cryptography-based approaches is embedding forensic watermarks within the media content, which would be able to survive digital-to-analog conversion [53]. Following this, if a pirated copy appears on unauthorized distribution platforms, it would be possible to recover the underlying watermark identifier and trace it back to the user or device that produced it [54].

A common way to produce a pirated video is to record a movie shown on on an LCD screen. It is therefore relevant to evaluate the kind of distortion that may appear when such a display is being recorded. Early works have focused on the ability to recognize this type of piracy through the use of discriminating features

and artifacts, such as combing artifacts [55], video jitter [56], ghosting artifacts [57], and the luminance flicker [58]. The latter artifact is the focus of our work here.

An example of how the flicker signal is presented in a pirated video is shown in Figure 1.2. Flicker, a result of the interplay between the camcorder and the LCD screen, is typically incarnated by more or less visible dark and bright strips scrolling up or down the recaptured video. By design, an LCD screen makes use of a source of light called a *back-light* that illuminates the liquid crystal cells of the screen from behind. This back-light is a periodic signal whose frequency lies within 120Hz-1kHz. Upon being captured by a camcorder, with a typical frame rate in the 30-60fps range, the light is aliased to a lower frequency, making it visible to humans. Many camcorders in use today employ a *rolling shutter*, which means that they acquire rows of a video frame sequentially, each at a different time, rather than all together at the same time. This rolling shutter effect results visually in the movement of the dark and bright strips up/down the recaptured video.

Figure 1.2: Flicker artifact when recording an LCD screen displaying a uniformly gray frame with a camcorder.

In our work on flicker forensics, we go beyond the focus of previous work in the literature that addressed using the flicker to identify the type of piracy – *camcorder piracy* – that creates flicker. Working with flicker-containing videos, we discuss various estimation techniques to extract parameters related to the flicker signature and examine applications that this signature can be used towards. We demonstrate how the flicker can be used to identify and characterize the camcorder and LCD screen that were used to produce the flicker-containing video, which would allow it to be used as an aid for movie piracy investigations.

1.3 Dissertation Organization

The rest of the dissertation is organized as follows.

In Chapter 2, we discuss approaches to collect reference power recordings and examine ENF signal estimation methods. This also includes a discussion on our proposed spectrum combining approach for more robust ENF signal estimates.

In Chapter 3, we carry out a study examining the factors that can affect the capture of ENF traces in recordings. Our focus is on audio recordings made by battery-powered recorders, and the factors discussed are related to the environment in which the recording was made and the recording device used to make it.

In Chapter 4, we present our proposed ENF-based application that seeks to identify the grid-of-origin of a media recording using a learning framework that exploits the statistical differences between ENF signals from different grids.

In Chapter 5, we show how analyzing the ENF traces found in a video can be used to characterize the video camera that originally produced this video, an application that was inspired by our work on the flicker signal.

In Chapter 6, we focus on flicker forensics. We build an analytic model that relates the parameters of the flicker captured in a pirated video to inner parameters of the camcorder and LCD screen used to produce it. We then use this model to build a flicker-based approach to identify and characterize the devices producing a pirated video.

In Chapter 7, we conclude this dissertation and outline research avenues for future work.

Chapter 2

Extraction of ENF Signals

2.1 Chapter Introduction

As discussed in Chapter 1, the ENF signal can have a number of useful real-world applications, in fields related to security and forensics as well as in other fields. A major first step to any of these applications is to accurately extract and estimate the ENF signal from an ENF-containing media signal. Following that, the validity of the results of these applications will strongly depend on how accurate and robust our estimates are. The focus of this chapter is on the approaches by which we can extract this ENF signal.

The rest of this chapter is organized as follows. In Section 2.2, we discuss approaches by which we can collect reference power signals, which will contain ENF traces at a high signal-to-noise ratio (SNR) and are valuable for ENF studies. In Section 2.3, we describe various ENF estimation approaches that have been proposed

in the literature. In Section 2.4, we carry out a comparison on a group of the ENF signal estimation approaches described in Section 2.3. We examine the effect of varying the frame size and the signal's SNR levels on the reliability of ENF signal estimation. In Section 2.5, we present our proposed approach for ENF extraction, which exploits the presence of ENF traces at different harmonics to produce more robust estimates. We conclude the chapter in Section 2.6.

2.2 Reference power recordings

Reference power recordings can be very useful to ENF analysis. The ENF traces found in these reference recordings are typically much stronger, and exhibit a higher SNR than the ENF traces found in audio or video recordings. They are useful for studying the properties of ENF signals and they can be used as a reference and a guide for ENF signals extracted from media recordings. As mentioned in Chapter 1, the variations in ENF signals from recordings that were simultaneously made in the same power grid should be very similar. Given an audio recording and a simultaneously recorded reference power signal, the latter can provide confident a ENF signal estimate that can be used to verify the integrity of the ENF-containing audio signal and authenticate its time-of-location. In this section, we describe different methods that can be used to acquire reference power recordings, including ones proposed in the literature and ones that we adopted in our work at the University of Maryland.

The Power Information Technology Laboratory at the University of Tennessee, Knoxville operates the North American Power Grid Frequency Monitoring Network

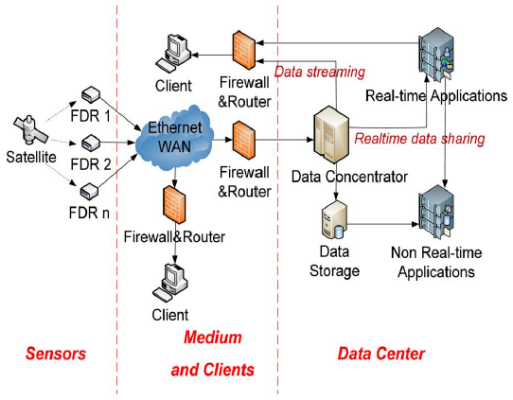
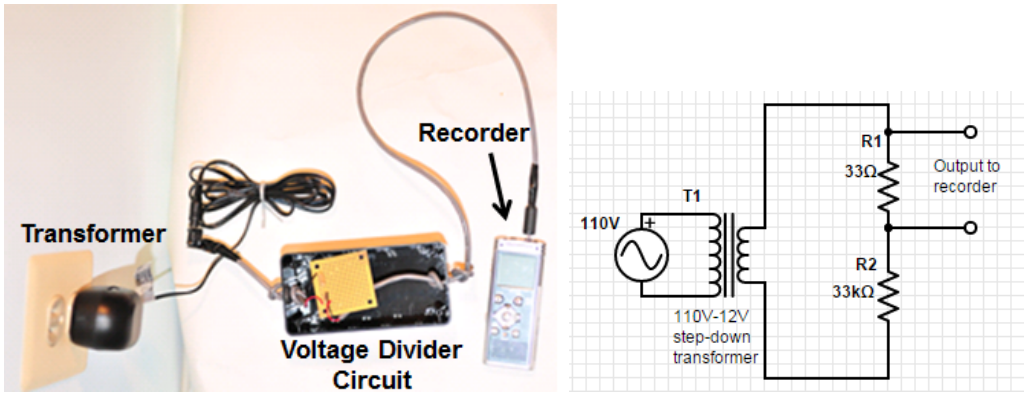


Figure 2.1: Framework of the FNET system [59]

System (FNET). The FNET is a power grid situational awareness tool that collects real-time, Global Position System (GPS) time-stamped measurements of grid reference data at the distribution level [25].

A framework for FNET is shown in Figure 2.1. The FNET system consists of two major components, which are the frequency disturbance recorders (FDRs) and the information management system (IMS). The FDRs are the sensors of the system; each FDR is an embedded microprocessor system that performs local GPS-synchronized measurements, such as computing the instantaneous ENF values over time. In this set-up, the FDR estimates the power frequency values at a rate of 10 records/s using phasor techniques [60]. The measured data is sent to the server through the Internet, where the IMS collects the data, stores it and provides a platform for the visualization and analysis of power system phenomena. More information on the FNET system can be found in [59, 61–63]. Another system similar to the FNET system, called the wide area management systems (WAMS), has been set up in Egypt, where the center providing the information management functions is at Helwan University [46, 47].



(a) (b)

Figure 2.2: (a) shows the reference recording measurement set-up that we use in our work, and (b) shows the circuit schematic.

Systems like FNET and WAMS offer tremendous benefits in power frequency monitoring, yet one does not need access to them in order to acquire ENF reference signals. An inexpensive hardware circuit can be built to record a power reference signal given access to an electric wall outlet. A step-down transformer is typically used to get the voltage from the wall outlet voltage levels down to a level that an analog-to-digital converter can capture. This is the approach that we have opted to use in our ENF work for acquiring the power reference signals that we need.

There is more than one approach to building the sensor hardware. In [64], an anti-aliasing filter is placed in the circuit along with a fuse for safety purposes. In [3], the step-down circuit is connected to a a BeagleBone Black board, via a Shmitt-trigger, that computes an estimate for the ENF signal directly.

In our work, the circuit is a simple voltage divider circuit, and it is connected to a digital audio recorder that records the raw power signal. The recorded digital signal is processed later using ENF estimation techniques, that are discussed in the next section, to extract the reference ENF signal. Figure 2.2 shows the circuit set-

up that we have used to collect our reference recordings[1]., and a schematic of the circuit[2].

2.3 Overview of ENF Estimation Approaches

As discussed earlier, accurately and reliably estimating ENF signals is key to the validity of the final results of an ENF-based application. This is especially important given that ENF traces found in media recordings tend to be of a low SNR, which necessitates the development of robust extraction approaches.

A necessary stage before estimating the changing instantaneous ENF value over time is pre-processing the signal. Typically, and as the ENF component is in a low frequency band, a low-pass filter, with proper anti-aliasing, can be applied to the ENF-containing signal to make the forthcoming computations of the estimation algorithms easier. For some estimation approaches, it also helps to band-pass the ENF-containing signal around the frequency band of interest, i.e., frequency band surrounding the nominal ENF value, or the frequency band surrounding an ENF harmonic where ENF traces are observed. This ENF-containing signal is then divided into consecutive overlapping, or non-overlapping, frames. A frequency estimation approach is applied on each frame to estimate its most dominant frequency around the nominal ENF value, or its harmonics. This frequency estimate would be the estimate of the instantaneous ENF value for this frame. Concatenating the

[1]We would like to thank Michael Luo for his assistance in building this circuit as part of his participation in the MERIT REU program at UMD in Summer 2012.

[2]The circuit schema was drawn using the online application found at: www.digikey.com/schemeit/.

frequency estimates of all the frames together forms the extracted ENF signal. The length of the frame, typically on the order of seconds, indicates the resolution of the extracted ENF signal. A trade-off typically exists here. A smaller frame size better captures the ENF variations but may result in poorer performance of the frequency estimation approach, and vice versa. The frame size used will also affect certain ENF-based applications, e.g., the temporal resolution in ENF-based time-of-recording estimation in media recordings will be limited by this frame size.

Broadly speaking, ENF estimation approaches fall into one of three categories: time-domain approaches, non-parametric frequency-domain approaches and parametric frequency-domain approaches.

2.3.1 Time-domain approach

The time-domain zero-crossing approach is fairly straightforward and is one of the few ENF estimation approaches that is not preceded by dividing the recording into consecutive blocks for individual processing. As described in [10], a band-pass filter with 49-51Hz cutoff (or equivalently a 59-61Hz cutoff) is first applied to the ENF-containing signal, without downsampling initially. This is done to separate the ENF waveform from the rest of the recording. Afterwards, the zero-crossings of the remaining ENF waveform are computed and the time differences between consecutive zero values is computed to estimate the instantaneous period of the power signal, and consequently used to compute the instantaneous ENF estimates.

2.3.2 Non-Parametric Frequency-Domain Approaches

Non-parametric approaches do not assume an explicit model for the data. The majority of these approaches are based on the Fourier analysis of a signal. The main methods considered here are based on the spectrogram and the time recursive iterative adaptive approach (TR-IAA).

Spectrogram: By far, the most commonly used non-parametric frequency-domain approach is a spectrogram-based approach, or equivalently, a periodogram-based approach or a Short Time Fourier Transform (STFT)-based approach. STFT is usually used for signals with a time-varying spectrum, which is the case of the ENF waveform captured in media recordings. After the ENF-containing signal is segmented into frames, each frame undergoes Fourier analysis to determine the strengths of the frequencies present. The spectrogram is then defined as the squared magnitude of the STFT, and is usually displayed as a two-dimensional intensity plot, with the two axes being time and frequency [27].

As the captured power signal in a frame can be considered to be a sinusoid of a frequency close to the nominal ENF value (or to a harmonic of the nominal ENF value), embedded in noise, the power spectral density (PSD) of it, estimated by the spectrogram, should ideally exhibit a peak at the frequency of the sinusoidal signal. Estimating this frequency well gives a good estimate for the ENF value of this frame.

The straightforward approach to estimating this frequency would be finding

the frequency that has the maximum spectral power component. Directly choosing this frequency as the ENF value, however, typically leads to loss in accuracy, because the spectrogram is computed for a finite number of discrete frequencies and the actual frequency with the maximum energy may not be among them. For this reason, typically, STFT-based ENF estimation approaches carry out further computations to obtain a more robust estimate. Examples of such operations are using quadratic interpolation or a weighted energy approach.

Maximum energy with quadratic interpolation: In this approach, we begin with finding the frequency that has the maximum spectral power component. Then, quadratic interpolation is used to fit the PSD points corresponding to the range about the discrete frequency with the maximum energy [24]. This interpolation is outlined below based on the approach in [65].

As the computed PSD of each frame is a function of the discrete frequencies, we can think of it as a function of frequency bin numbers, or simply indices. Denoting the index of the frequency with the maximum energy as k_{max}, we can define a coordinate system centered at $(k_{max}, 0)$. We take the log magnitude value of the PSD as $y(k)$. Using three points of the PSD, namely, at $k_{max}-1$, k_{max} and $k_{max}+1$, we can carry out quadratic interpolation on the parabola of the form: $y(k) = a(k-p)^2 + b$. Solving for the parabola peak p, we obtain the following expression.

$$p = \frac{1}{2} \frac{y(k_{max}-1) - y(k_{max}+1)}{y(k_{max}-1) - 2y(k_{max}) + y(k_{max}+1)}, \tag{2.1}$$

where $y(k)$ is defined as

$$y(k) = 20 \log_{10} |PSD(k)|. \tag{2.2}$$

The estimate of the peak location in bins that corresponds to the true frequency is $k_t = k_{max} + p$. The frequency estimate corresponding to k_t is $\frac{k_t f_s}{N}$, where f_s is the sampling frequency and N is the number of FFT points used in computing the spectrogram.

A drawback of this approach is that it is susceptible to outliers. If the maximum energy happens to be far away from the nominal frequency due to additive noise or interference from content, the subsequent estimation would be erroneous.

Weighted energy: The weighted energy approach makes use of the additional information that we have on the approximate location of the desired frequency (close to the nominal ENF value or its harmonics). As a result, the frequency estimates are more robust to outliers [4].

The weighted energy approach finds the frequency estimate $F(n)$, for the n^{th} frame, by weighing the frequency bins around the nominal ENF value (or the ENF harmonic of interest). The expression for frequency estimates is then given by the following equation [4]:

$$F(n) = \frac{\sum_{l=L_1}^{L_2} f(n,l)S(n,l)}{\sum_{l=L_1}^{L2} S(n,l)} \tag{2.3}$$

where L_1 and L_2 are the FFT indices of the boundary of the averaging region and $f(n,l)$ and $S(n,l)$ are the frequency and energy in the l^{th} frequency bin of the n^{th} time frame.

TR-IAA: Though spectrogram-based approaches are most commonly used, the authors in [24] advocate the use of a non-parametric, adaptive and high resolution technique known as the time-recursive iterative adaptive approach (TR-IAA). This

algorithm reaches the spectral estimates of a given frame by minimizing a quadratic cost function using a weighted least squares formulation. TR-IAA is an iterative technique that takes from 10 to 15 iterations to converge, where the spectral estimate is initialized to be either equal to the spectrogram or to the final spectral estimate of the preceding time frame [24,66]. This method is more computationally extensive than spectrogram-based techniques. After the convergence of the spectral estimate, a quadratic interpolation scheme similar to the first spectrogram-based approach discussed is used to estimate the frequency. It has been shown that the TR-IAA based approach provides slightly better frequency estimates than the spectrogram based approach, when the frame size is 20-30 seconds [24]. In our work on ENF, we typically use smaller frames of size around 5 seconds or less. As we consider such frame sizes, and due to the high computational costs of the TR-IAA based algorithm observed, we focus on spectrogram-based approaches for non-parametric methods in the comparison study in Section 2.4.

2.3.3 Parametric Frequency-domain Approaches

Parametric methods assume an explicit model for the signal and underlying noise. Due to such an explicit assumption about the model, the estimates obtained using parametric approaches are expected to be more accurate than those obtained by non-parametric approaches [67]. In the study shown here, we consider two of the most widely used parametric frequency estimation methods, which are based on the subspace analysis of a signal and noise model. In general, these methods

can be used to estimate the frequency of a signal composed of P complex frequency sinusoids embedded in white noise. ENF-containing signals can be passed through a band-pass filter centered around the ENF nominal frequency, or one of its harmonics. The resultant signal will consist of only one real sinusoid, making the value of P to be used equal to 2. The methods we study are here the MUltiple SIgnal Classification (MUSIC) method and the Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) method. In what follows, we briefly describe these two methods [67].

MUSIC: The MUSIC algorithm is a subspace-based approach to frequency estimation that relies on eigendecomposition and the properties between the signal and noise subspaces for sinusoidal signals with additive white noise. The algorithm first computes an $M \times M$ correlation matrix, where M is chosen to be larger than P, the number of anticipated complex exponentials. More specifically, for an N-point observed signal $x[n]$ where $N >> M$, we generate an $M \times N$ data matrix of the form:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}(1) & \mathbf{x}(2) & ... & \mathbf{x}(N-2) & \mathbf{x}(N-1) \end{bmatrix}^T \tag{2.4}$$

where $\mathbf{x}(n) = \begin{bmatrix} x(n) & x(n+1) & ... & x(n+M-1) \end{bmatrix}^T$.

An estimate of the correlation matrix $\hat{\mathbf{R}}_x$ can be computed by $\hat{\mathbf{R}}_x = \frac{1}{N}\mathbf{X}^H\mathbf{X}$, where the superscript H denotes the Hermitian of a matrix. Eigenanalysis is carried out on $\hat{\mathbf{R}}_x$ to find the vectors spanning the signal and noise subspaces. These subspaces are orthogonal to one another. The eigenvectors $\mathbf{q}_1, \mathbf{q}_2, ..\mathbf{q}_P$ correspond to the largest P eigenvalues that span the signal subspace and the remaining eigenvectors

$(\mathbf{q}_{P+1}, \mathbf{q}_{P+2}, ..\mathbf{q}_M)$ span the noise subspace.

MUSIC makes use of the orthogonality property of noise eigenvectors and steering vectors $\mathbf{v}(f_k)$, $1 \leq k \leq P$ (corresponding to the actual frequency components of the signal). Here $\mathbf{v}(f)$ is defined as:

$$v(f) = \begin{bmatrix} 1 & e^{j2\pi f} & e^{j4\pi f} & ... & e^{j2(M-1)\pi f} \end{bmatrix}^T \tag{2.5}$$

The pseudo-spectrum of MUSIC, $\bar{R}_{music}$, can then be computed as:

$$\bar{R}_{music}(e^{j2\pi f}) = \frac{1}{\sum_{m=p+1}^{M} |\mathbf{v}^H(f)\mathbf{q}_m|^2} = \frac{1}{|Q_m(e^{j2\pi f})|^2} \tag{2.6}$$

Due to the orthogonality property, the denominator should be zero at the frequencies of the signal.

Among various techniques studied in the statistical signal processing literature, RootMUSIC provides high precision at a moderate computational cost. It solves for the roots of the denominator directly [67]. The frequency estimates using this method are the arguments of the P roots closest to the unit circle. We opted to use RootMUSIC in our comparisons in Section 2.4.

ESPRIT: ESPRIT makes use of the rotational property between staggered subspaces that is invoked to produce the frequency estimates. In our case, this property relies on observations of the signal over two intervals of the same length staggered in time. ESPRIT is similar to MUSIC in the sense that they are both subspace-based approaches, but it is different in that it works with the signal subspace rather than the noise subspace. The implementation that we employ estimates the signal subspace from the data matrix $\mathbf{X}$ of Equation (2.4).

A Singular Value Decomposition (SVD) is applied to $\mathbf{X}$, giving:

$$\mathbf{X} = \mathbf{LSU}^H \tag{2.7}$$

where $\mathbf{L}$ is an $N \times N$ matrix of the left singular vectors, $\mathbf{S}$ is an $N \times M$ matrix with its main diagonal entries containing the singular values, and $\mathbf{U}$ is an $M \times M$ matrix of the right singular vectors. The singular values correspond to the square roots of the eigenvalues of the sample correlation matrix $\hat{\mathbf{R}}_x$ scaled by N, and the columns of $\mathbf{U}$ are the eigenvectors of $\hat{\mathbf{R}}_x$. These vectors form an orthonormal basis for the underlying M-dimensional vector space. More specifically, $\mathbf{U}$ can be written as $\mathbf{U} = [\mathbf{U}_s|\mathbf{U}_n]$, where $\mathbf{U}_s$ is the $M \times P$ matrix of right singular vectors corresponding to the singular values with the P largest magnitudes and $\mathbf{U}_n$ is the $M \times (M - P)$ matrix containing the remaining right singular vectors. The signal subspace can be partitioned into two smaller $(M - 1)$-dimensional subspaces as:

$$\mathbf{U}_s = \begin{bmatrix} \mathbf{U}_1 \\ \\ * \end{bmatrix} = \begin{bmatrix} * \\ \\ \mathbf{U}_2, \end{bmatrix} \tag{2.8}$$

where $\mathbf{U}_1$ and $\mathbf{U}_2$ correspond to the unstaggered and staggered subspaces, respectively. The relation between $\mathbf{U}_1$ and $\mathbf{U}_2$ can be written as:

$$\mathbf{U}_2 = \mathbf{U}_1\mathbf{Q}, \tag{2.9}$$

where $\mathbf{Q}$ is a $P \times P$ matrix. $\mathbf{Q}$ can be computed using a least squares method. Eigenanalysis can then be carried out on $\mathbf{Q}$. The frequency estimates can be extracted from the arguments of the eigenvalues of $\mathbf{Q}$. Denoting these eigenvalues by $\phi_k, 1 \leq k \leq P$, we can find the frequency estimates, $\hat{f}_k$, by:

$$\hat{f}_k = \frac{\angle\phi_k}{2\pi} \quad \text{with} \quad 1 \leq k \leq P. \tag{2.10}$$

It is worth noting that the accuracy of the estimates obtained using subspace methods can differ significantly depending on the parameter M chosen for the data matrix in Equation (2.4). It was shown in [68] that for estimating the frequency in a single sinusoid in white noise, the error variance is minimal when M is either $\frac{N}{3}$ or $\frac{2N}{3}$. For a general multiple signal case, a rule of thumb suggested in [69] is that M should be in the range $[\frac{2N}{5}, \frac{3N}{5}]$. So, before we carried out the experiments detailed in the next section, we tested the accuracy of MUSIC and ESPRIT on the range $M \in [\frac{N}{3}, \frac{2N}{3}]$ for the values of N that we intended to use. We found the value of M that gave the most accurate result for each case of N value, and subsequently used it in the experiments that follow. The metric used for determining accuracy is the correlation coefficient, which will be discussed in the next section.

2.4 Comparison of Estimation Approaches

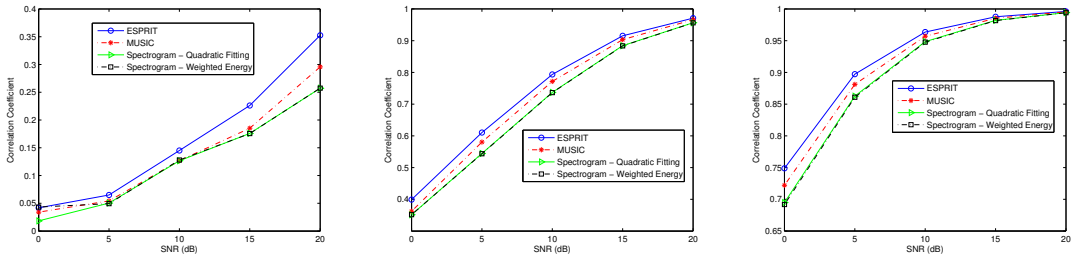
In this section, we compare the performance of some of the different frequency estimation approaches discussed in Section 2.3. In particular, we consider spectrogram-based approaches using maximum energy with quadratic interpolation, and using the weighted approach. We also consider the parametric approaches of MUSIC and ESPRIT. We present the experiments that we carry out and the criteria by which we compare the methods.

2.4.1 Experiments on Synthetic Signals

To facilitate a comparative study, we first generate synthetic signals with ground truth frequencies. In our model of the synthetic signal, the frequency of the signal changes on a frame-by-frame basis and the frequency estimation algorithms are applied to each frame. Since the ENF signal has a slowly-varying frequency, we attempt to capture the correlation of the frequencies in consecutive frames by first generating a random sequence of frequencies having a normal distribution of mean $\mu = 60$ and standard deviation $\sigma = 0.0133$. Then, to mimic the pseudo-periodic behavior of the ENF fluctuations [26], we pass this sequence through a filter of the form:

$$H(z) = \frac{1}{1 - 0.97z^{-1}}. \tag{2.11}$$

For one instance of this simulation, the synthetic signal is generated as a series of consecutive sinusoidal signals of frequencies corresponding to the generated sequence, each with a phase following a uniform distribution $U(0, 2\pi)$. We consider a sampling frequency of 441Hz, and add additive white gaussian noise (AWGN) to achieve different SNR levels. We also examine several frame sizes, ranging from 0.1 seconds (44 samples) to 1 second (441 samples). These frame sizes are smaller than the sizes needed for most ENF-based applications, but being able to accurately estimate ENF signals at such a high temporal resolution can improve the performance of most ENF-based applications and is crucial for certain ENF-based applications. An example of such an application is identifying the location-of-recording within a power grid, where ENF-related location specific variations may not be captured well



(a) 0.1-second frame (b) 0.5-second frame (c) 1-second frame

Figure 2.3: Synthetic Signals: Correlation Coefficient vs. SNR for very short frame sizes

if the frame size is larger than 1 second [29].

We carry out 100 simulation runs where each run has a different set of frequencies, phase angles, and additive noise. In each simulation run, the frequencies are estimated on a frame-by-frame basis using the frequency extraction methods discussed in Section 2.3. For the spectrogram-based approaches, the number of FFT points is chosen such that the frequency resolution is approximately 0.03Hz. For the weighted energy approach, the range considered for weighting is [59.8, 60.2]Hz. For the subspace methods, we use the experimentally optimized M value for the dimension of the data matrix based as mentioned in Section 2.3.3.

A direct way of comparing the frequency estimates is to subtract each sequence of estimates from the true sequence of frequencies and estimate the mean difference. This criterion does not fit our application because it may penalize some methods that tend to have an inherent bias in estimating the frequency, such as the weighted energy approach. As we are interested in measuring the similarity in the trends of the signals for most applications based on ENF signal analysis, these trends are better revealed using a correlation based metric.

More specifically, to measure the performance of different frequency estimation

methods, we use the cross-correlation coefficient between the frequency estimates and the true frequencies for different frame sizes and different SNR values. The results obtained are shown in Figure 2.3. The correlation coefficient of two sequences that do not match is close to zero, and the correlation coefficient for matching sequences is expected to be higher, and ideally closer to 1. From Figure 2.3, we can see that the frequency estimates worsen when more noise is added and the SNR decreases, which is to be expected. Also, when the frame size becomes too small, as with Figure 2.3(a), where it is equal to 0.1 second, all estimation techniques behave rather poorly, as the correlation coefficient achieved was around 0.35 in high SNR cases and only 0.05 in low SNR cases; the latter is barely differentiable from matching two unrelated ENF signals. For this reason, it is advisable not to use such small frame sizes whenever possible.

We also note that both spectrogram-based approaches give similar performances; the subspace-based approaches give better performances than the spectrogram-based approaches, with ESPRIT consistently outperforming MUSIC by a moderate margin. This is due to an explicit assumption on the sinusoidal signal model in the parametric approaches and this signal model matches the synthetic model well. As with ESPRIT outperforming MUSIC, our observation in the context of the ENF problem is consistent with the general result in the literature that ESPRIT gives slightly more accurate estimates than MUSIC for a similar computational cost [70].

2.4.2 Experiments on Power Grid Data Set

To validate the results obtained using synthetic data, we examine the performance on a power signal measurement available at [71]. This dataset also provides a sequence of reference frequencies computed using a Frequency Disturbance Recorder (FDR). FDRs estimate frequencies using phasor analysis and signal resampling techniques. The estimates are reported to provide a resolution of around ± 0.0005Hz [59].

We assume the power measurement to be noise-less for simplicity, and we add various levels of AWGN to it. We estimate the frequencies present for different frame sizes using the estimation techniques discussed in Section 2.3. We use the frequencies computed by the FDR as ground truth, and compute the correlation coefficient between this ground truth and the estimated sequence of frequencies. This correlation coefficient was computed after temporally aligning the ground truth frequency sequence with the estimated ones.

The results for the 1-second frame size case can be seen in Figure 2.4. The estimates obtained are worse than the estimates for synthetic data, which is understandable as the model we used was an idealization of the reality and the power measurements are likely not noise-free or perfect sinusoids. However, the order of accuracy remains consistent with the synthetic data comparison, with ESPRIT outperforming the other three methods, followed by MUSIC. Figure 2.4 shows similar trends and relative positions to those shown in Figure 2.3(c) except that the correlation coefficient values are lower.

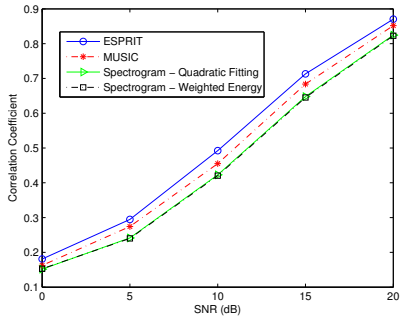


Figure 2.4: Power Grid Data Set: Correlation Coefficient vs. SNR for 1-sec frames

2.4.3 Matching ENF Signals from Audio and Power

A main forensic application involving ENF analysis requires matching an ENF signal extracted from an audio signal to that extracted from a power signal in order to determine the time of recording. Here, we carry out experiments to compare which of the four estimation methods gives the best matching between ENF signals that are known to be recorded at the same time.

More specifically, we make two simultaneous recordings of audio and power signals. The power signal is obtained from an electric outlet using the circuit set-up of Figure 2.2. The audio signal is obtained by recording the background noises in a room. The audio recording is expected to pick up ENF traces whose variations should match well the ENF traces present in the concurrently recorded power recording.

Prior to ENF extraction, we pass the audio signal through a bandpass filter centered around 60Hz, the nominal frequency, to remove as much noise as possible without affecting the frequency band of interest. We partition both the power signal and the filtered audio signal into overlapping frames, and compute the frequency

Table 2.1: Correlation coefficient values between the ENF signals extracted from the power and audio recordings.

Method	ESPRIT	MUSIC	Spectrogram – Quad. Interp.	Spectrogram – Weighted Energy
Correlation coefficient	0.978	0.977	0.970	0.954

estimates for each frame using the four ENF extraction techniques under study. In order to differentiate the various estimation methods, we compute the correlation coefficient between the audio and power estimates of each method.

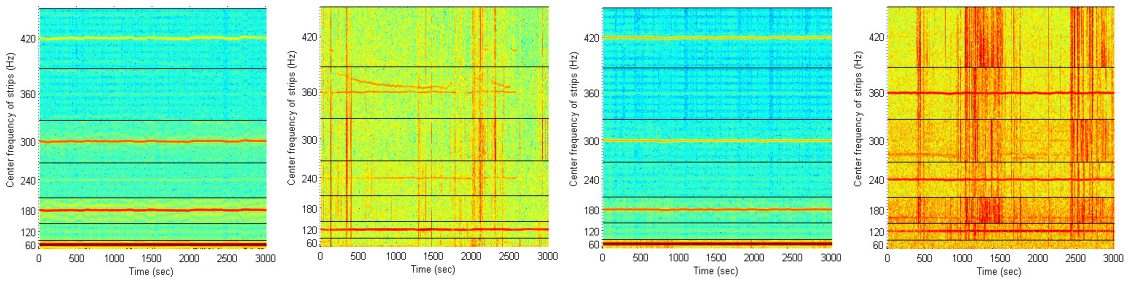
The results are shown in Table 2.1, where a higher value of the correlation coefficient suggests better matching between the ENF signals extracted from the audio and power signals. Although the values in Table 2.1 are close, they support the findings reached earlier that ESPRIT gives the best results followed by MUSIC. The results here also demonstrate that the spectrogram-based quadratic interpolation approach performs better than the weighted energy approach. This can be due to the restriction of the estimates of the weighted energy to a chosen range, which may not be true in the case of all signals. The quadratic interpolation approach has no such restraints.

All the estimation approaches discussed so far use the ENF traces found around either the nominal ENF value, or around a single harmonic of the nominal ENF value, to reach the final ENF signal estimate. In the next section, we explore a novel ENF estimation approach that exploits the presence of the ENF traces around more than one harmonic of the nominal ENF value to reach a more accurate and robust ENF signal estimate.

2.5 Proposed Spectrum Combining Estimation Approach

The validity of the results obtained in ENF-based studies depends strongly on how well the weak ENF signal is estimated from the available measurements. In power signals, the ENF appears around the nominal frequency of 50/60 Hz and at its harmonics. As mentioned earlier, most existing frequency estimation approaches for ENF signal extraction rely on the spectral band surrounding the nominal frequency, or on the spectral band surrounding one of the higher harmonics [15, 24].

We have observed that scaled versions of almost the same variations appear in many of the harmonic bands, although the ENF signal strength at different harmonic frequencies differs with recording environments and devices used. An example of this, which will be discussed further in Section 2.5.2.1, can be seen in Figure 2.5 for four different recording settings. Following this observation, we propose a low computational complexity approach to extract the ENF signal that strategically makes use of multiple spectral bands. We are inspired by a related problem in handling multipath in wireless communications that has led to a maximum ratio combining approach used in RAKE receivers [72], as well as by a harmonic extension of the classical MUSIC estimator [73]. Our proposed spectrum combining approach exploits traces of different ENF components appearing in a signal, and adaptively combines them based on estimates of the local SNRs to achieve a more robust and accurate estimate than that achieved by using only one component. We examine two variants of this approach, based on spectrogram and subspace frequency estimation techniques, respectively. The usefulness of this approach is especially prominent



(a) power recording - residential (b) audio recording - residential (c) power recording - workplace (d) audio recording - workplace

Figure 2.5: Spectrogram strips about the harmonics of 60 Hz for two sets of simultaneously recorded power and audio measurements.

when extracting weak ENF signals from audio/video files, which is challenging due to the presence of noise and media content.

In what follows, Section 2.5.1 explains the model considered and the proposed approach for spectrum combining, and Section 2.5.2 discusses the experiments conducted and analyzes the results obtained.

2.5.1 Model and Proposed Approach

2.5.1.1 Problem Formulation

As motivated earlier, we devise a technique to estimate the major frequency component of the power signal, through exploiting the presence of the ENF at the base frequency and its harmonics. We estimate, frame-by-frame over time, the deviation Δf_o of the base ENF from the nominal frequency f_o. The same variations of base and harmonic components from the nominal ENF values (up to a scaling factor) provide us with multiple observations on the base-band deviation Δf_o. To estimate Δf_o for a given frame, the estimation can be related to a highly

simplified model where we have L observations $Y_1, Y_2, ...Y_L$. The k^{th} observation is $Y_k = \alpha_k h(\Delta f_o) + N_k$, where $h(\Delta f_o)$ is a deterministic function of Δf_o representing the spectrum information of the ENF-containing signal that attains its maximum at the ENF component; α_k relates to the strength of the "signal component" around the k^{th} harmonic; the N_k's denote noise components that are assumed to be independently distributed, each following a normal distribution with mean zero and variance σ_k^2. To estimate the function values $h(\Delta f_o)$, a Maximum Likelihood Estimator (MLE) can be adopted as:

$$\hat{h}_{MLE}(\Delta f_o) = \frac{\sum_k (Y_k/\alpha_k).(\alpha_k^2/\sigma_k^2)}{\sum_k (\alpha_k^2/\sigma_k^2)}. \tag{2.12}$$

This suggests producing an ENF estimate by combining multiple base and harmonic spectral bands, each weighted according to its relative strength with respect to noise; the combined spectrum can then be processed to obtain Δf_o.

In the spectral domain, the ENF-containing signal can be considered to be a summation of impulses at the base ENF and its harmonics. Due to the recording mechanism and environment, additional frequency components may interfere around these bands, and we consider such interferences as noise around each harmonic. For a given frame, the observed power spectrum component, $P_{B,k}(f)$, contributed at the k^{th} harmonic band, analogous to Y_k in Equation (2.12), can be expressed as:

$$P_{B,k}(f) = A_k h_k(f) + P_{n,k}(f), \tag{2.13}$$

for $f \in [k\,(f_o - f_B)\,, k\,(f_o + f_B)]$, where f_B reflects the empirical support of ENF presence around the base frequency. Here, A_k denotes the magnitude of the energy

contributed by the frequency component close to kf_o; $P_{n,k}(f)$ denotes the independent noise component around the k^{th} harmonic, assumed white within the bandwidth of interest; $h_k(f)$ denotes an impulse-like function that attains its maximum at $f = k(f_o+\Delta f_o)$. In practice, we observe $h_k(f)$ as a peak energy concentration with a small spread. Assuming that the power signal contains L harmonics, we can write the power spectrum of the signal for a given frame as $P_{signal}(f) = \sum_{k=1}^{L} P_{B,k}\,(f)$.

To estimate the frequency deviation Δf_o for a given frame, the proposed spectrum combining approach first compresses and shifts the spectrum components to the nominal base range $[f_o - f_B, f_o + f_B]$, and then combines the components together. This is analogous to Equation (2.12), where the weighted summation can then be written as:

$$S(f) = \sum_{k=1}^{L} w_k \bar{P}_{B,k}\,(kf)\,, \tag{2.14}$$

where $\bar{P}_{B,k}$ denotes the normalized power spectrum component around the k^{th} harmonic. When applying Equation (2.14), the $h_k(f)$ components of Equation (2.13) are compressed along the frequency axis to become $h_k(kf)$ components; they should each have their maximum at $f = f_o + \Delta f_o$. The combining weight, w_k, in Equation (2.14) has been introduced to weigh the various harmonic spectral bands based on the SNR around the corresponding harmonic. This weight is analogous to the α_k^2/σ_k^2 in the numerator of Equation (2.12); the denominator in (2.12) can be seen as a normalization parameter for these SNR-based combining weights and can be incorporated into the definition of w_k. The frequency $f_{ENF} = f_o + \Delta f_o$ can be obtained by searching for the maximum in $S(f)$.

2.5.1.2 Determining Spectral Combining Weights

The combining weights are computed for a recording over a certain time duration (e.g., 30 minutes), and then recomputed for subsequent durations. This makes the weights adaptive to reflect the changing strength of the ENF traces in the harmonic bands over time. Taking the combining weight w_1 for the base band around f_o for a certain time duration as an example, we set w_1 as an estimate of the SNR in the band; $\hat{P}_{signal,1}/\hat{P}_{noise,1}$. We choose $f_B = 1\,\text{Hz}$, because in the US, the ENF fluctuates within 59.98 Hz and 60.02 Hz. The overall band for computing SNR would be $[59, 61]\,\text{Hz}$. Then, $\hat{P}_{signal,1}$ is the average power spectral density (PSD) within the band $[59.98, 60.02]\,\text{Hz}$ over the chosen time duration and $\hat{P}_{noise,1}$ is the average PSD in the bands $[59, 59.98]\,\text{Hz}$ and $[60.02, 61]\,\text{Hz}$ over the same time duration. The combining weights for other harmonic spectral bands are computed in the same manner, except for different sizes of the spectral bands considered.

2.5.1.3 Instantaneous ENF Estimation

After the w_k's are computed, the spectrum $S(f)$ can be computed using Equation (2.14). $S(f)$ can be seen as a weighted summation of shifted and compressed spectral bands. We explore the estimation of spectral bands through two methods: a spectrogram and a subspace-based "pseudo-spectrum". In the first method, the spectral harmonic bands are chosen from the spectrogram and shifted and compressed to compute $S(f)$. In the second method, the subspace-based frequency estimation technique MUSIC is used. As discussed in Section 2.3.3, the MUSIC

algorithm looks for P complex exponential frequencies in a signal, and computes a pseudo-spectrum [74]. The peaks in the pseudo-spectrum correspond to the dominant frequencies found in the signal. Here, we apply band-pass filters on the ENF-containing signal around the ENF harmonics. For each band-passed signal, the pseudo-spectrum can be computed for $P = 2$ (corresponding to the positive and negative ENF components), and the spectral band around the harmonic of interest is identified from the computed pseudo-spectrum and stored. After all harmonic bands are estimated, they are used to compute $S(f)$.

In practice, when computing $S(f)$, we shift the spectral bands to one of the higher bands rather than the base band, compressing or expanding as necessary, to make use of the wider range of variations available around higher harmonics for the same frequency resolution. In the results shown in Section 2.5.2, we shift the bands to the 240 Hz band. After $S(f)$ is computed, we search for its maximum through quadratic interpolation [65]. The ENF is set as the argument of the maximum. If $S(f)$ is defined around a higher spectral band than the base band, the solution is scaled accordingly.

2.5.2 Experiments and Results

2.5.2.1 Experimental Set-up

We carry out recordings in two different environment settings, one overnight in a residential setting (an apartment) and one during the day in a workplace setting (an office). Each set of recordings is composed of a power mains signal, used as

a reference signal, and an audio signal recorded concurrently and expected to pick up the ENF traces. The audio is recorded using a battery powered Olympus Voice Recorder WS-700M at a sampling rate of 44.1 kHz in MP3 format at 256 kbps [75]. All recordings are made in Maryland, which is part of the US Eastern Grid. The recordings are downsampled to 1000 Hz in WAV format to ease the computations. Their spectra are estimated for consecutive frames of 5 seconds long each. Sample spectrogram strips around the harmonics of the nominal ENF for each of the four recordings are shown in Figure 2.5.

From Figures 2.5(a) and 2.5(c), we can see that the power signal is almost noise free around the harmonics, and the ENF has a strong presence around the odd harmonics. Figure 2.5(b) shows that in the residential audio recording, the ENF is present strongly only around 120 Hz, with faint to no presence around the other harmonics. On the other hand, Figure 2.5(d) shows stronger presence for the ENF in the workplace audio signal around 240 Hz and 360 Hz; the noisy component appearing around 120 Hz is not the ENF as will be shown in Section 2.5.2.3. Interestingly, the ENF has almost no presence around the nominal 60 Hz in the audio signals. The frequency response of the built-in microphone of the Olympus recorder used to make these recordings ranges between 70Hz and 20kHz [75], which can explain this observation.

We estimate the ENF signals from the four recordings using our proposed spectrum combining approach. Table 2.2 shows sample values of combining weights computed at various harmonics considered for the four recordings. The combining weights shown are normalized, each expressed as a percentage of the sum of com-

Table 2.2: Sample values of spectral combining weights w_k's (The weights corresponding to the dominant harmonic bands are in italic)

Center Freq. (Hz)	**Power Residential**	**Power Office**	**Audio Residential**	**Audio office**
60	6.32	6.81	0	0
120	2.03	1.02	*75.75*	*15.69*
180	*18.79*	*19.75*	0	1.80
240	4.73	1.99	*13.77*	*32.00*
300	*28.88*	*29.15*	5.20	1.91
360	2.70	4.07	5.28	*44.69*
420	*35.61*	*36.16*	0	2.19
480	0.93	1.05	0	1.72

bining weights of all the spectral bands for its time duration. The resulting values conform with our earlier observations on the bands where the ENF has a strong presence. We also noticed that in some cases, the weight before normalization is less than 1, which implies that in such bands, the noise component is stronger than the signal component. We set such weights to zero before normalization and thus have them excluded from the summation of Equation (2.14).

2.5.2.2 Comparison Framework

To assess the performance of our proposed spectrum combining approach, we compare it to the conventional approach of using frequency estimation techniques on a single spectral band. Since our audio recordings show a strong presence of the ENF at 120 Hz, 240 Hz and/or 360 Hz, we generate ENF estimates based on the individual bands centered around these frequencies. We observe that isolated outliers may appear in the ENF estimates outside the known range of ENF variations ($[59.98, 60.02]$ Hz for the nominal band); we replace these outliers by the average of

the frequency estimates preceding and succeeding them. For each audio ENF signal estimate, we have a corresponding reference ENF signal estimate extracted from a power signal recorded simultaneously with the audio signal. We generate the audio ENF estimates twice, once using the spectrogram-based technique and once using MUSIC. For the reference power ENF, we use the average of the estimates obtained through the two techniques. As mentioned earlier, in applying both techniques on the signals, we find the maximum in the spectrogram or pseudo-spectrum through quadratic interpolation.

To compare between ENF estimates from an audio signal and a reference power signal, we split both ENF signal estimates into segments of 96 points each, corresponding to 8 minutes of data. We examine the performance in ENF estimation for time-of-recording authentication, a main ENF-based application [4, 31]. We consider a hypothesis testing framework:

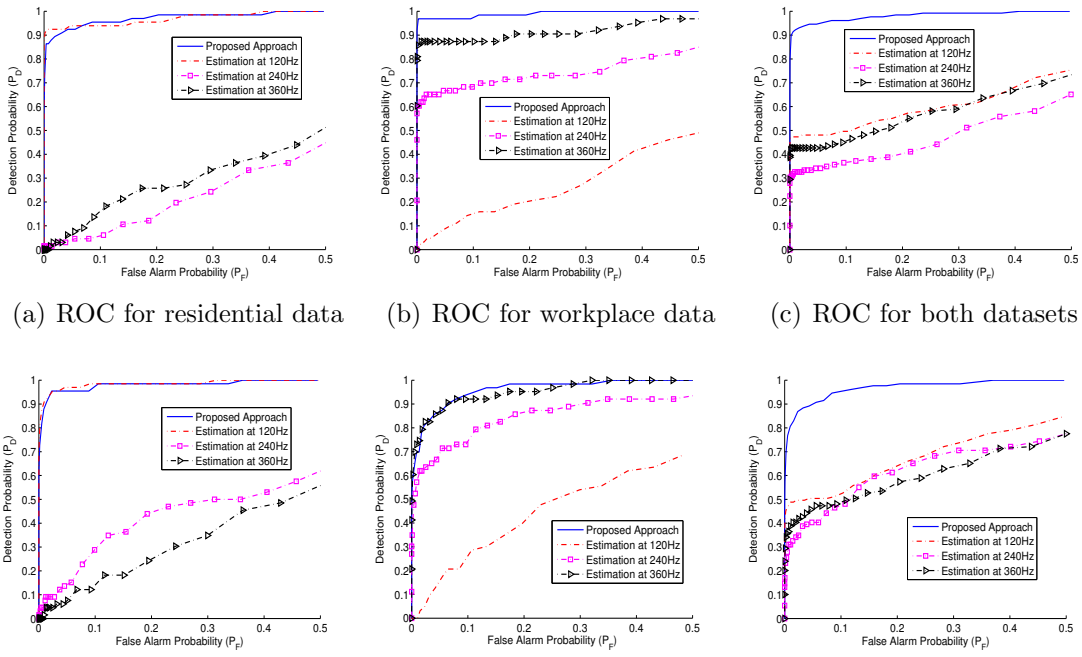
$$\begin{cases} H_0 : \text{segments were recorded at different times.} \\ H_1 : \text{segments were recorded simultaneously.} \end{cases}$$

We find the correlation coefficient between all combinations of segments from both ENF signals, and apply thresholding to decide on H_0 vs. H_1. Figure 2.6 shows the Receiver Operator Characteristic (ROC) curves for the cases studied.

2.5.2.3 Results and Discussions

Figure 2.6 shows that the spectrum combining approach outperforms the approaches for individual bands; it behaves comparably to estimation around the

"dominant harmonic" if present, i.e., 120 Hz in the residential recordings and 360 Hz in the workplace recordings. Figures 2.6(b) and 2.6(e) demonstrate the ability of the proposed approach to suppress spurious peaks appearing at certain harmonics due to noise and distortions: the poor ROC curve from estimation around 120 Hz reveals that the frequency component in that band, seen in Figure 2.5(d), is not the ENF. It could be the result of stray electromagnetic fields with complex spectra found in the workplace (office) setting [15]. The proposed approach was unaffected and was able to leverage the true ENF components at 240 Hz and 360 Hz to achieve a good ENF estimate. Figures 2.6(c) and 2.6(f) demonstrate the proposed approach's robustness and its ability to adapt to unpredictable changes in ENF strengths at harmonics in cases where the dominant harmonic varies across the length of a recording.



(a) ROC for residential data (b) ROC for workplace data (c) ROC for both datasets

(d) ROC for residential data (e) ROC for workplace data (f) ROC for both datasets

Figure 2.6: ROC curves for matching ENF estimates from audio signals to those estimated from reference power signals. (a)-(c) show spectrogram-based results, and (d)-(f) shows MUSIC-based results.

Comparing the two variants of the proposed approach, we can see that the spectrogram-based method yields better results than the MUSIC-based method. As the MUSIC pseudo-spectrum is computed through a rather sophisticated procedure and does not have a physical meaning as direct as the spectrogram does, the observations from MUSIC may thus deviate more from the model discussed in Section 2.5.1.1 than the spectrogram-based observations do. This would affect the effectiveness of the spectrum combining, and explain the discrepancy in the performance between the two variants.

2.6 Chapter Summary

In this chapter, we have studied different parametric and non-parametric frequency estimation approaches as applied to extracting the ENF signal from media recordings in high temporal and frequency resolution. We have conducted experiments on synthetic data and experimental data under different noise conditions to evaluate the performance of the studied methods using a correlation coefficient based metric. Our results have demonstrated that the ESPRIT-based parametric frequency estimation method provides the best results for ENF matching using correlation coefficient, especially for cases where the temporal resolution is high, i.e., estimating instantaneous ENF values for frames of size 1 second. When using frames of larger size, the computational costs of the parametric approaches may not be worth the expected improvements in accuracy and one may opt to use non-parametric spectrogram-based approaches instead.

We have also proposed a novel spectrum combining approach for extracting the ENF signal from multimedia recordings. The approach makes use of the ENF around the different harmonics of 50/60 Hz rather than only one harmonic. This is achieved through a weighted summation of multiple spectral bands from around the harmonics, weighted according to the local SNR in each band. Our experiments have shown that the proposed approach can achieve more accurate and robust estimates than the conventional approach for ENF estimation. Of the two variants presented, the improvement in performance was more visible in the non-parametric spectrogram-based variant over the parametric MUSIC-based variant.

Chapter 3

Factors Affecting Capture of ENF Traces in Recordings

3.1 Chapter Introduction

As we have established, ENF traces can be captured in audio and video recordings made in areas where there is electrical activity. The majority of ENF research so far has broadly focused on two areas: developing approaches to accurately extract the ENF signal from media recordings [24, 30, 32, 36, 38, 48], and examining novel ENF-based applications or improving on their performance [8,29,37,41,43,59]. Amid all this work towards extracting and using the ENF signal, an essential research question remains without a solid answer: In what kind of recording situations can we be confident that ENF traces will be captured in media signals? Answering this question can be informative for the true applicability of the ENF signature, and can be beneficial towards the development of ENF-based applications.

The ENF traces are an imprint of power grid activity on media recordings, so a basic requirement for the capture of ENF traces is that there should be some sort of electrical activity in the place of recording. A distinction can be made between two types of recordings: recordings made using recorders connected to the power mains and recordings made using battery-powered recorders. In the case of the former, it is generally accepted that ENF traces will appear in the resultant recording, due to electromagnetic interference resulting from the recorder's connection to the power mains [3]. It is in the case of the latter that the situation becomes more complex.

In this chapter, we carry out a study exploring factors that can affect the capture of ENF traces in media recordings. Our focus here is on *audio* recordings made using *battery-powered* recorders. First, we summarize the results of studies in the literature that have addressed this issue. Next, we examine further factors that can promote or hinder the capture of ENF traces. Our explorations include two types of factors: (1) those related to the recording environment and (2) those related to the recording device used.

The rest of this chapter is organized as follows. In Section 3.2, we present a summary on the results of the studies previously done towards understand the factors affecting the capture of ENF traces in audio recordings. In Sections 3.3 and 3.4, we present our explorations into the effect of the recording environment and the recording device, respectively, on the capture of ENF traces. In Section 3.5, we conclude the chapter and outline avenues for future work.

3.2 Overview of previous work

Understanding the factors that promote or hinder the capture of ENF traces in media recordings can help us gain a better understanding on the situations in which ENF analysis is applicable. It can also inform the way we develop ENF-based applications. In this section, we summarize results shown in previous studies pertaining to the factors affecting the capture of ENF traces in audio recordings made by battery-powered recorders.

Broadly speaking, the capture of ENF traces can be affected by several factors that can be divided into two categories: factors related to the environment in which the recording was made and factors related to the recording device used to make the recording. Interaction between different factors may as well lead to different results. For instance, electromagnetic fields in the place of recording promote the capture of ENF traces if the microphone is *dynamic* but not in the case where the microphone is *electret*. Table 3.1 shows a sample of factors that have been studied in the literature for their effect on the capture of ENF traces in audio recordings [3, 14, 33].

Overall, the most common cause of the capture of ENF traces in audio recordings is the acoustic mains hum, which can be produced by mains-powered equipment in the place of recording. Recently, the hypothesis that this background noise is a carrier of ENF traces was confirmed in [3]. Experiments carried out in an indoor setting suggested a high robustness of ENF traces. These traces were present in a recording made 10 meters away from a noise source emitting ENF-containing noise and located in a different room.

Table 3.1: Sample of factors affecting the capture of ENF traces in audio recordings made by battery-powered recorders

	Factors	**Effect**
Environmental	Electromagnetic (EM) fields	Promote capture of ENF traces in recordings made by *dynamic* microphones but not in those made by *electret* microphones.
	Acoustic mains hum	Promotes capture of ENF traces; sources include fans, power adaptors, lights, and fridges.
	Electric cables in vicinity	Not sufficient for capture of ENF traces.
Device-related	Type of microphone	Different microphone types have different reactions to the same sources, e.g., to EM fields.
	Frequency band of recorders	A recorder may not be capable of recording at low frequencies, e.g., at around 50/60Hz.
	Internal compression by recorder	Strong compression, e.g., Adaptive Multi-Rate, which can limit capturing of ENF traces.

In general, and as discussed in Chapter 2, ENF traces are not restricted to appear in the frequency band surrounding the nominal 50/60Hz band, but can also appear in bands surrounding the higher harmonics of the nominal ENF value, i.e., around integer multiples of 50/60Hz. When examining signals containing ENF traces, it is common to observe scaled versions of almost the same variations appearing at different harmonic frequencies. The specific harmonics at which the ENF traces appear and the strength of the captured traces at each harmonic may differ from one recording to another [32, 48]. Therefore, the locations and corresponding strengths by which the ENF traces appear should be considered as well when examining factors affecting the capture of ENF traces in media.

In the following sections, we present our explorations on factors that may influence the embedding of ENF traces in audio recordings due to the environment in which the recording was made and due to the recorder device being used.

3.3 Explorations on environment-related factors

As established in Section 3.2, a main source of ENF traces in audio recordings is the acoustic mains hum. Knowing this, it can be useful to examine the way in which the acoustic mains hum can behave as a physical sound wave in a recording environment [76]. This can help understand its effects on the manner in which ENF traces are embedded in an audio recording.

The acoustic mains hum is a sound signal, and sound is a mechanical longitudinal wave. It is a disturbance that travels through a medium, transporting energy from one location to another through a series of interacting particles. It is a longitudinal wave because the particles, e.g., air molecules, vibrate in a direction parallel to the direction of propagation of the wave. The sound wave can then be characterized by compressions and rarefactions, which are physical regions where particles are compressed together and regions where particles are spread apart, respectively. Sound waves move at a speed v that depends on the medium in which they travel. In air, this speed is 343m/s at a temperature of 20°C. A sound wave can generally be characterized by its frequency f and wavelength λ, which can be related to its speed through:

$$v = f \cdot \lambda \tag{3.1}$$

The acoustic mains hum can be seen as the summation of several sound waves of varying strength (amplitude) at frequencies that are multiples of the nominal ENF value. Using a speed of sound of 343m/s and Equation (3.1), Table 3.2 shows the expected wavelengths of the sound waves at different ENF harmonics. These

Table 3.2: Wavelengths of sound waves at different ENF harmonics in air at a temperature of 20°C

Frequency	**Wavelength**	**Frequency**	**Wavelength**
50Hz	6.86m	60Hz	5.72m
100Hz	3.43m	120Hz	2.86m
150Hz	2.29m	180Hz	1.91m
200Hz	1.72m	240Hz	1.43m
250Hz	1.37m	300Hz	1.14m
300Hz	1.14m	360Hz	0.95m
350Hz	0.98m	420Hz	0.82m

wavelength values, ranging from 0.8m to 7m, help put in perspective the movement of these sound waves in a location where we expect an audio recording to capture ENF traces.

In the remainder of this section, we show results on experiments done in different environment conditions that demonstrate how the factors related to the recording environment can affect the way ENF traces are captured in a recording.

3.3.1 Effect of wave interference

When a sound wave, such as the acoustic mains hum, is propagating in a medium, such as air, it will reach a point where it hits a boundary, i.e., an interface between two media. There are four possible behaviors a wave can exhibit at a boundary: reflection (bouncing off the boundary), diffraction (a change in direction of the wave as it passes through an opening or around a barrier in its path), transmission (crossing of the boundary into the new medium), and refraction (occurs with transmission and is characterized by a change in speed and direction). Typically, depending on the differences between the two media and the properties

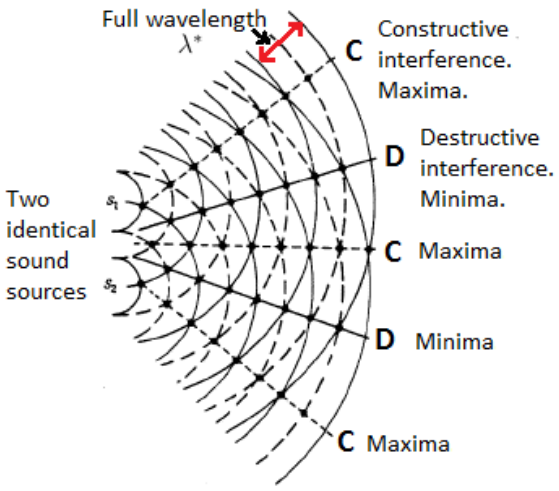


Figure 3.1: Diagram showing the result of interference of two identical sound waves. Original image (before modification) from [77] .

of each individual medium, the wave propagation behavior at the interface will be different. For instance, in the case of a sound wave propagating in air, hitting a hard smooth surface results in more reflection than hitting a soft surface [76].

The reflection of sound waves can lead to reverberations or echoes. The reflected sound waves can interfere with the original sound waves, which will result in different areas in the room with different interference sound signals. The two extremes of these interferences are the *constructive* interference and the *destructive* interference.

Figure 3.1 shows a theoretical diagram of how waves emitted from two point sources are expected to interfere with each other. In this diagram, two sources are emitting identical signals of wavelength λ^*, which is the distance between any two solid curves, or any two dotted curves. The distance between a solid curve and a dotted curve is the half-wavelength $\lambda^*/2$. Here, a solid curve denotes a compression, i.e., an area of high pressure where the medium's particles are compressed together,

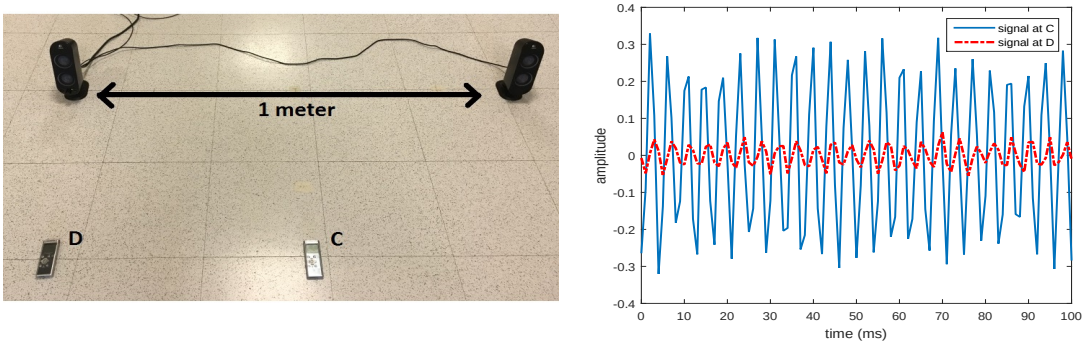
and a dotted curve denotes a rarefaction, i.e., an area of low pressure where the medium's particles are spread apart. Locations of expected constructive and destructive interference can be seen in the figure denoted by C's and D's, respectively.

Constructive interference is created in locations where compressions from the two signals meet, creating an area of even higher pressure, followed by a meeting of rarefactions from the two signals, making the pressure in the area even lower. The overall effect is that the combined signal becomes stronger. Destructive interference is the opposite of constructive interference. It occurs at locations where a compression meets a rarefaction, thereby canceling each other out and resulting in little movement of the medium's particles.

In a given location where there are equipment connected to the power mains, the acoustic mains hum may be emitted from different sources. These multiple versions of this signal, along with their reflections, would lead to various levels of interference in the environment of recording. This suggests that in certain environments where ENF traces are expected to occur, e.g., a room where there is electrical activity, ENF traces may not appear in a recording made in this environment or may appear at different strengths depending on the specific location of the recorder within the environment.

Guided by the diagram of Figure 3.1, we carried out a verification experiment[1] to examine how ENF traces expected to be captured in an audio signal may be canceled out, or strengthened, depending on the recorder's specific location within the environment. We generated a synthetic tone signal at 240Hz and emitted it

[1]We would like to thank Yingxue Wang for her assistance in carrying out this experiment.



(a) Experiment setup

(b) Sample of obtained recordings

Figure 3.2: (a) shows the experimental set-up for the sound wave interference experiment we carried out. (b) shows samples of the recordings made at locations C and D.

from two speakers placed 1 meter apart. Figure 3.2(a) shows the layout we used. We placed two Olympus recorders [75], one at a central location C expected to exhibit constructive interference, and one at another location D expected to exhibit destructive interference.

We show samples of the recorded audio signals in Figure 3.2(b). Although the signal recorded at location D in theory should be a zero signal, the presence of a non-negligible amplitude can be explained by somewhat constructive interference from reflected versions of the original two signals. In a more controlled recording setting, the amplitude of the signal recorded at D should be much lower. However, we can see that the signal recorded at location C is stronger, i.e., has a larger amplitude of around 0.25, than the signal recorded at location D, which has a lower amplitude of around 0.03. This is consistent with what we would expect given our understanding of sound wave interference.

In a practical scenario where a recorder is recording an acoustic signal carrying ENF traces, there may be several sources in the room emitting signals carrying

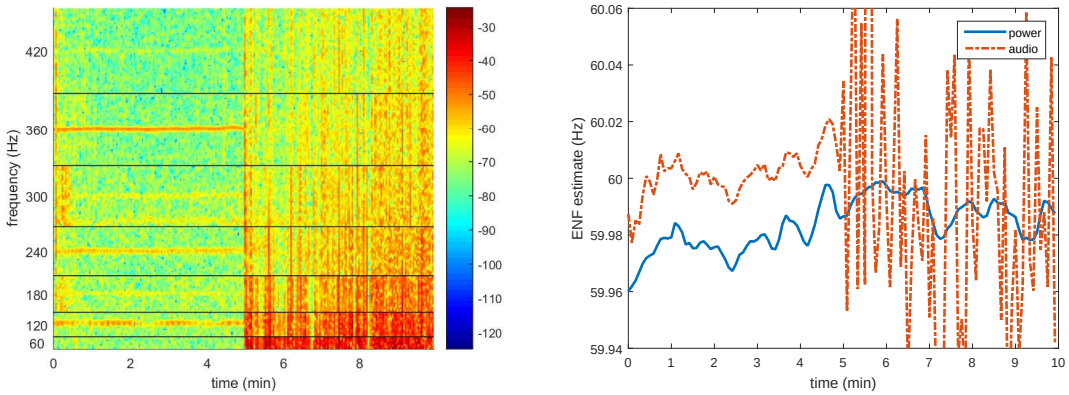
ENF traces, and some of these signals will be reflected as well. All these signals will interfere with one another to form the signal recorded by the recorder. The discussions in this section suggest that the particular location of the recorder within the location of recording will have an effect on how the ENF traces are captured.

3.3.2 Effect of recorder movement

In this section, we consider the effect of moving a recording on the way ENF traces are captured in the resulting recording.

We carried out an experiment where we made a 10-min audio recording using an Olympus recording in an office setting in Maryland, where the nominal ENF value is 60Hz. The Olympus recorder was stationary during the first half of the recording and then was continuously moved during the second half of the recording. Figure 3.3(a) shows the spectrogram strips about the ENF harmonics for the recorded signal. We can see that, along with ENF traces around the 120Hz and 240Hz harmonics, there is a prominent ENF trace around the 360Hz harmonic in the first half of the signal. After the 5-minute mark, however, the ENF traces become distorted and indistinguishable.

Figure 3.3(b) shows the ENF signal extracted from the audio recording (from around the 360Hz component) and the ENF signal extracted form a simultaneously recorded reference power recording. We can see that the audio ENF signal matches well with the power signal in the first 5 minutes, but not afterwards, which is when the audio recorder was moving. The correlation coefficient between the audio and



(a) Spectrogram strips

(b) ENF estimates

Figure 3.3: (a) Spectrogram strips about the harmonics of 60Hz from an audio recording where the recorder starts moving at the 5-min mark, and (b) ENF signals extracted from the audio recording and a simultaneously recorded reference power recording.

power ENF signals was as high as 95% in the first half of the recording.

This experiment demonstrates how moving an audio recorder can compromise the captured ENF traces. A possible explanation for this observation can be that it is an effect of changes in the pressure of sound waves at the recorder's microphone while it is being moved.

As seen in this section, it is not recommended to move a recorder while making a recording expected to capture ENF traces. It is highly likely that the resulting recording will contain compromised ENF traces. This is not a desired effect when carrying out ENF analysis as estimating the minute variations in the instantaneous value of the ENF is important for the validity of the subsequent ENF-based applications.

3.3.3 ENF traces across different locations

In what follows, we aim to demonstrate and visualize how the strength of captured ENF traces can change across the length of a recording as the recording location is being changed. To this end, we use an Olympus recorder to make a single 1-hour long recording, where the recorder is moved every 2 minutes to a new location. While making this 1-hour long audio recording, we concurrently record a 1-hour long reference power recording. As seen in Section 3.3.2, such a reference recording can serve as a guide towards understanding how well ENF traces appear in the audio recording.

While making the audio recording, the Olympus recorder was moved to 25 different locations in the second floor of the Kim Engineering Building (KEB) at the University of Maryland, College Park. The trajectory followed by the recorder can be seen in Figure 3.4. Each number denotes a 2-minute stop made by the recorder. In most of the cases, a stop is equivalent to the recorder being placed in one location in a room/corridor. However, in the case of stops 10-15 and 16-21, we have placed the recorder in 6 different locations in two different rooms, Room 2111 and Room 2107, respectively. This was done in an effort to further examine how the specific location within an environment can have an effect on the presence of ENF traces, following the discussions in Section 3.3.1.

Figure 3.5 shows the spectrogram strips about the ENF harmonics for the hour long audio signal. The numbers in Figure 3.5 denote the stop where the recorder was stationed while that portion of the recording was being made. From this figure,

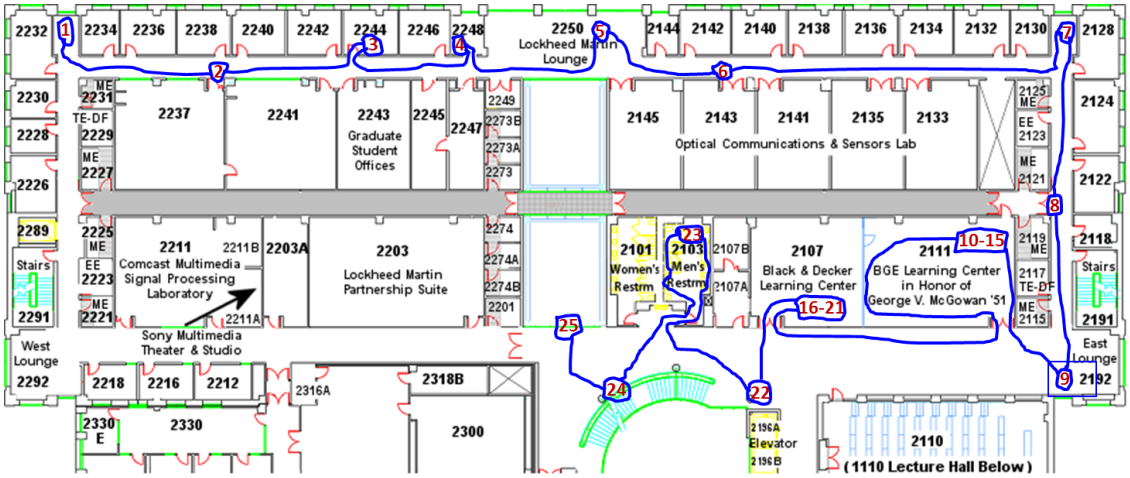


Figure 3.4: Floor plan of 2nd floor in KEB showing trajectory of recorder and stops made while recording. Each number denotes a 2-minute stop for the recorder. Original map obtained from [78]

we can see that there are three major harmonics around which the ENF traces are appearing strongly: 120Hz, 240Hz, and 360Hz. We also note that, in between stops as the recorder is being moved from one stop to another, there is a significant noise and no dominant ENF traces can be observed. This is similar to the observations seen in Section 3.3.2.

We can visibly see that the strength of the ENF component can vary from one recording location to another. The 360Hz component, in particular, seems to be strong in only a handful of locations, such as locations 4, 7, and 8, and almost non-existent in others, such as in locations 16-21. To explain this, we can note that all these different rooms/corridors contain different equipment that emit different audio signals carrying ENF traces. Also, some differences can be attributed to constructive/destructive interferences among the signals carrying ENF traces. An indication of this is that location 10 has a visibly weaker 240Hz components than

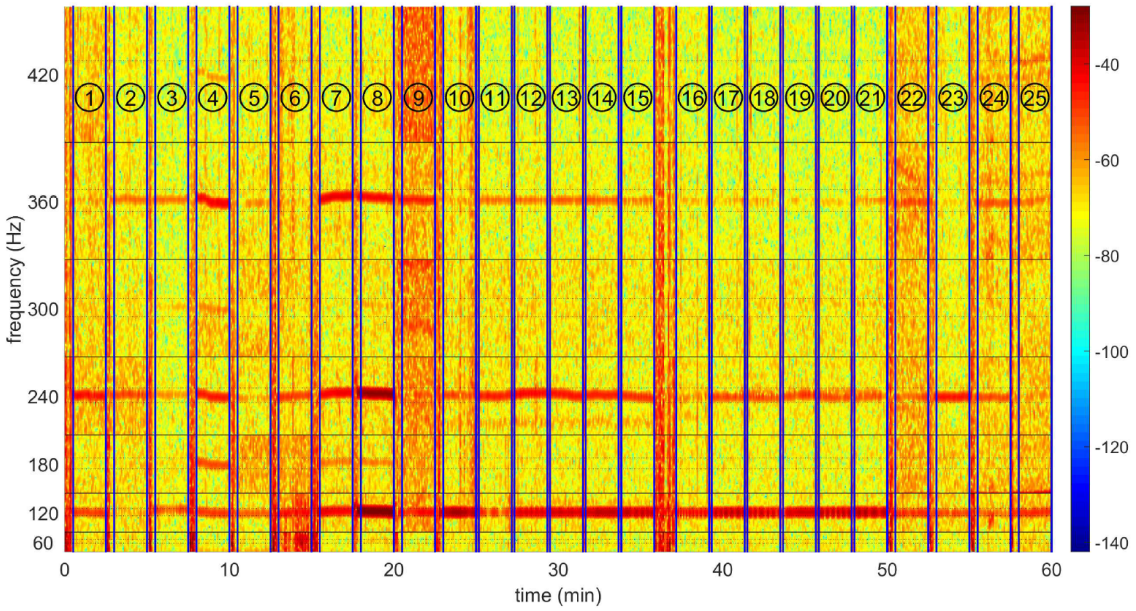
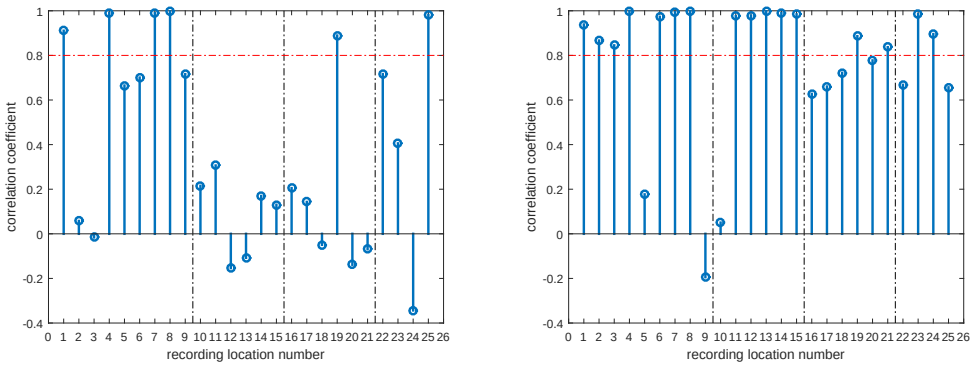


Figure 3.5: Spectrogram strips about ENF harmonics for the obtained 1-hour long recording. The numbers denote the location of the recorder while the recording was being made. Stops 10-15 were in the same room, and Stops 16-21 were in the same room. The blue vertical lines denote the separation between the recorder being placed in a certain location, and it being moved to a different location.

the rest of the locations 11-15 in the same room.

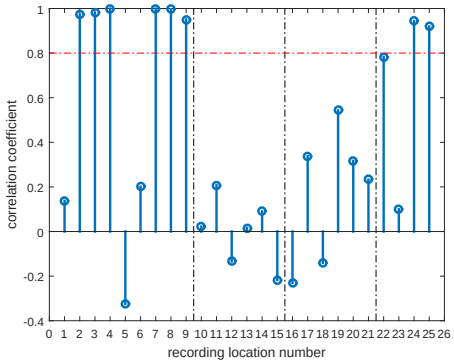
As mentioned earlier, while making this 1-hour long audio recording, we were recording a concurrent 1-hour long reference power recording. To assess how well the ENF traces are captured at each location, we examine the correlation coefficient between the ENF signal extracted from an audio clip around a certain harmonic with the reference ENF signal extracted from the corresponding reference power clip. For this set of recordings, we extracted ENF signals for frames of 10-seconds long with 50% overlap.

Figure 3.6 shows the correlation coefficient values obtained for all the locations considered for the cases of ENF signals extracted solely from around the 120Hz, the 240Hz, or the 360Hz harmonic. The red horizontal line refers to a 0.8 correlation coefficient value, which is the value we have chosen as the lower bound for an ac-



(a) case of 120Hz

(b) case of 240Hz



(c) case of 360Hz

Figure 3.6: Correlation coefficient values obtained for all the location cases considered for ENF signals extracted from around solely either the 120Hz, 240Hz, or 360Hz harmonic. Each correlation coefficient value is computed between an extracted ENF signal about an ENF harmonic with its corresponding reference power ENF signal segment. Vertical black borders are placed around locations of recording that belong to the same room, i.e., locations 10-15 and locations 16-21, respectively.

ceptably high correlation coefficient value. Recording clips achieving lower than this value are considered to have either not captured ENF traces at the particular harmonic or captured them weakly.

Examining Figure 3.5, we can see that there is a strong component at 120Hz throughout locations 10-21, even though this does not reflect in high correlation coefficients in most of these locations in Figure 3.6(a). It would seem that this component is not the ENF component and could be the result of stray electromagnetic

spectra in the location of recording. For the case of the 120Hz component in Figure 3.6(a), we can see that the only locations that give a high correlation coefficient are 1, 4, 7, 8, 19, and 25. Location 19 is the sole location in Room 2107 to show prominent ENF at 120Hz.

Examining Figure 3.6(b), we can see that most of the clips show high correlation coefficient values at 240Hz, thus exhibiting strong ENF traces. Again, we see another example of a sole recording in one room that does not exhibit similar behavior as the rest, i.e., location 10 in Room 2111, which confirms our earlier observation on the 240Hz component begin captured weakly at location 10 and strongly at locations 11-15.

Examining Figure 3.6(c) confirms our earlier observation that the 360Hz ENF component is strong at locations 4, 7, and 8. We can also see that the visibly faint 360Hz components at locations 2, 3, 9, 24, and 25 in Figure 3.5 were enough in these cases to yield good ENF signal estimates.

The results of this experiment provide further evidence that the environment and the specific location within it in which a recording is made have an effect on the captured ENF traces.

3.4 Explorations on device-related factors

In this section, we show results of experiments that suggest that the recorder (receiver) used to make an audio recording can have an effect on the way ENF traces are captured in the resulting recording. As discussed earlier, ENF traces are

not restricted to appear in the frequency band surrounding the nominal 50/60Hz band, but can also appear in bands surrounding the harmonics of the nominal ENF value. As such, when examining whether ENF traces are present in a recording, we examine frequency bands surrounding not only the nominal ENF values, but its harmonics as well.

In the experiments that follow[2], we concurrently record a reference power recording while carrying out an audio recording. We examine the spectrogram strips about the ENF harmonics of the audio signal to identify if ENF traces were possibly captured. To confirm that the traces observed were indeed due to the ENF, and not due to some other possible environmental signals, we compare the audio ENF variations to the ENF variations observed in the concurrently made power recordings. It is through this approach that we can confirm that the traces found about the ENF harmonics for the recordings shown in this section were indeed due to the ENF signal. Following this, the metric that we used to judge the strength of ENF presence in this section is a weight that estimates the local SNR at each ENF harmonic. We explain how we arrive at this estimate in Section 3.4.1. Then, we show the results of our experiments on device-related factors in Section 3.4.2.

3.4.1 Computation of local SNR estimate

ENF traces will appear in an audio recording when a signal carrying these traces, e.g., the acoustic mains hum is captured in it. We can express this signal

[2]We would like to thank Steven Gambino for his assistance in carrying out the experiments shown in this section.

$p(t)$ as a summation of sinusoids whose frequencies are time-varying around the harmonics of the nominal ENF value:

$$p(t) = \sum_{k=1}^{L} A_k \sin\left(2\pi k(f_0 + \Delta f(t))t + \phi\right), \tag{3.2}$$

where f_0 is the nominal ENF value equal to 50Hz or 60Hz depending on the grid, $\Delta f(t)$ is the instantaneous deviation of the ENF from the nominal value, and A_k is the amplitude of the k^{th} sinusoidal harmonic component denoting the strength of the particular ENF component.

Here, we aim to obtain an estimate of the local SNR at particular ENF harmonics, expressed in decibels (dB). We denote the local SNR at the k^{th} harmonic by:

$$SNR_k = 10\log\frac{P_{signal,k}}{P_{noise,k}} = 10\log\frac{A_k^2}{P_{noise,k}}, \tag{3.3}$$

where $P_{signal,k}$ and $P_{noise,k}$ are the powers of the signal and noise components, respectively, at the k^{th} harmonic.

To estimate the SNR of a certain ENF component contained in a signal, we examine the spectrogram of the audio signal, which gives estimates of the power spectral density (PSD) at different frequencies over time. We first divide the signal into consecutive frames. For each frame, we find the value of the peak of the spectrogram surrounding the multiple of the nominal ENF value we are interested in; this is our estimate of $P_{S+N,k} = P_{signal,k} + P_{noise,k}$ as it corresponds to the signal peak superimposed over the noise present. Denoting the frequency at which this peak appears as $f_{peak,k}$, we compute our estimate for $P_{noise,k}$ as the average of the PSD values corresponding to the frequencies in the ranges of $[f_{peak,k} - \Delta_1, f_{peak,k} - \Delta_2]$

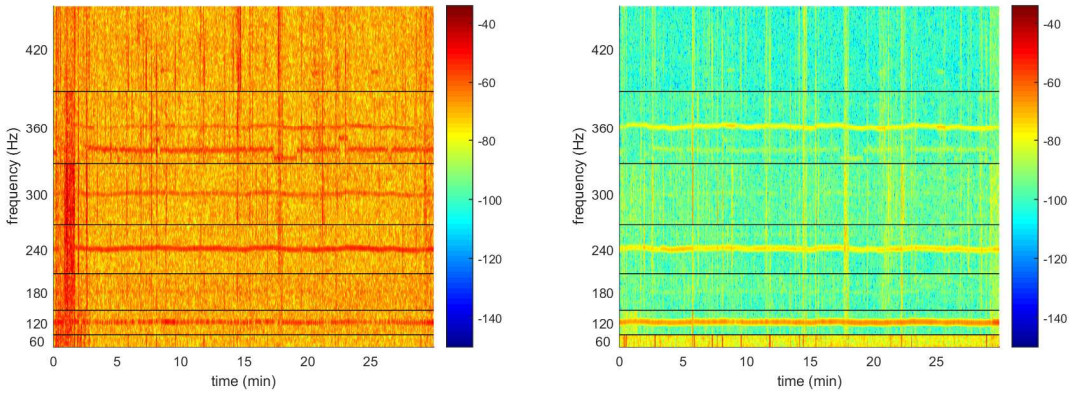
and $[f_{peak,k} + \Delta_2, f_{peak,k} + \Delta_1]$. In our work, we have empirically chosen $\Delta_1 = 2$Hz and $\Delta_2 = 0.5$Hz. After computing the $P_{signal,k}$ estimate as $P_{S+N,k} - P_{noise,k}$, we can compute the SNR_k estimate using Equation (3.3). Through this approach, we can estimate the SNR_k value for each frame. We can also compute the average SNR_k value over a certain time period as the mean of the SNR_k values of the frames within this time period.

3.4.2 Experiments and Results

In what follows, we present the results of two sets of experiments that we carried out, which suggest that the recording device used can have an effect on (i) the strength by which ENF traces are captured in a recording, and (ii) the ENF harmonics around which the traces can appear. As mentioned earlier, before we carry out the analysis on the strengths of the ENF traces observed, we ascertain that the observed traces are indeed due to the ENF by comparing them to traces observed in concurrently recorded reference power recordings. The experiments discussed here were carried out at the University of Maryland, College Park, where the nominal ENF value is 60Hz.

3.4.2.1 Difference in ENF strength

In this experiment, we carry out three sets of simultaneous recordings by two different receivers. Each recording is 30 minutes long. The first receiver is the built-in microphone of an Olympus 700-M audio recorder [75], and the second receiver is



(a) Olympus recording (b) B&K recording

Figure 3.7: Spectrogram strips about the harmonics of 60Hz from the first set of simultaneously recorded audio measurements in the same environment made by different recorders, Olympus and B&K.

a Brüel & Kjær (B&K) microphone (Type 4191) [79]. Figure 3.7 shows spectrogram strips of the recordings surrounding ENF harmonics for one set of the recordings. Here, we can see that both recorders capture prominent ENF traces at around 120Hz and 240Hz, and the B&K recording captures strong ENF traces at around 360Hz.

When comparing the 240Hz strips between the two spectrograms in Figure 3.7, we can see that the ENF traces captured by the Olympus recording have a red color while those of the B&K recording have a yellow color. Examining the color bars of the spectrograms, this indicates that the Olympus recording's strip has a higher PSD value than that of the B&K recording. However, a more telling indicator on the strength of the captured ENF traces is their respective local SNR.

We compute the SNR estimates around 120Hz, 240Hz, and 360Hz using the approach described in Section 3.4.1. The results, shown in Table 3.3, show that for each of these ENF components present in the audio signals, the component in the B&K recording has a higher SNR estimate than its counterpart in the Olym-

Table 3.3: SNR estimates (in dB) in set of simultaneous recordings made by different recorders, Olympus and B&K, in the same setting

SNR	**Set 1**		**Set 2**		**Set 3**	
around	**Olympus**	**B&K**	**Olympus**	**B&K**	**Olympus**	**B&K**
120Hz	8.39	26.41	18.47	23.23	19.23	26.47
240Hz	11.86	17.34	8.85	14.17	8.92	12.80
360Hz	6.93	15.61	6.93	7.07	14.25	20.79

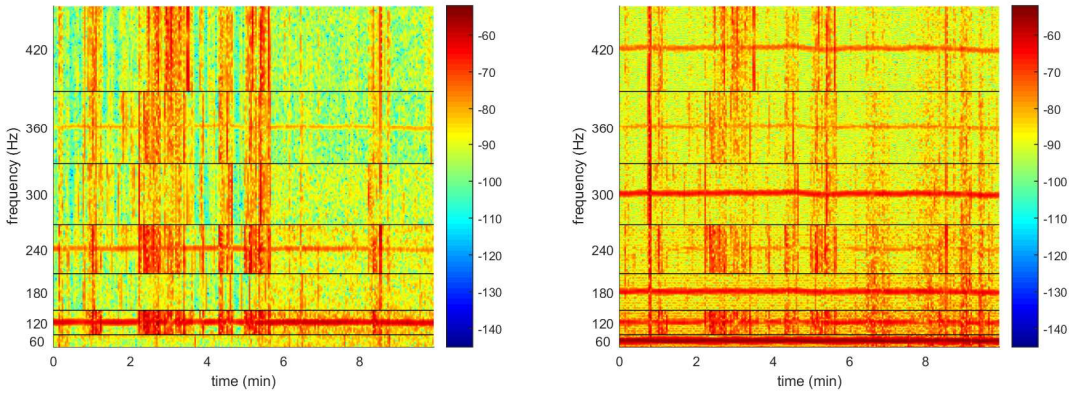
pus recording. Given that each set of recordings was made in the same location, this suggests that using the B&K recorder would yield stronger ENF traces, which demonstrates that the choice of receiver can have an affect the strength by which the ENF traces are captured.

3.4.2.2 Difference in ENF harmonics captured

In this experiment, we carry out a set of two simultaneous recordings using different receivers in the same environment. Here, one of the receivers is the B&K microphone (Type 4191) used in the experiments of Section 3.4.2.1, and the second receiver is a Max4466 microphone from Adafruit [80]. Figure 3.8 shows the spectrogram strips about the ENF harmonics for the recordings that were made.

The results of this experiment yield an interesting observation: Even though both recordings were made in the same location, the B&K microphone captures the ENF traces at odd harmonics of 60Hz, which is consistent with the results of Section 3.4.2.1, while the Max4466 captures the ENF traces at even harmonics of 60Hz.

To get a numerical perspective on the difference in the local strengths of the captured ENF traces between the two recordings, Table 3.4 shows the SNR estimates



(a) B&K recording

(b) Max4466 recording

Figure 3.8: Spectrogram strips about the harmonics of 60Hz from the set of simultaneously recorded audio measurements in the same environment made by different recorders, B&K and Max4466.

Table 3.4: SNR estimates (in dB) in set of simultaneous recordings made by different recorders, B&K and Max4466, in the same setting

SNR around	**60Hz**	**120Hz**	**180Hz**	**240Hz**	**300Hz**	**360Hz**	**420Hz**
B&K	5.10	16.44	3.75	10.46	3.65	9.58	3.64
Max4466	19.05	9.67	16.22	5.61	18.38	8.73	12.61

computed for the ENF traces observed in each of the two recordings. The results shown in this table and the observed strips in Figure 3.8 suggest that the choice of recorder used to make a recording can have an effect on the ENF harmonic at which the ENF traces appear.

3.4.2.3 Discussion

In each of the two experiments shown in this section, we have used two different recorders in the same environment to record the background noise carrying ENF traces. Examining the results of these experiments by themselves, we can hypothesize that the choice of recorder used to make an audio recording can influence

the strength of the ENF traces captured and the harmonic locations at which these ENF traces appear.

In Section 3.3, however, we have seen that the specific locations within an environment where an audio recorder is placed can have an effect as well on the strengths/locations by which ENF traces are captured in the resulting recording. It is therefore plausible that such factors may have played a role as well in the results shown in this section, in addition to factors related to the recorders being used. This demonstrates how understanding better how ENF traces are captured in a recording is a complex problem, which would require further study to reach a more confident conclusion.

3.5 Chapter Summary

The ENF signal has been under study in the research community in recent years for its ubiquitous nature as it can be captured intrinsically by media recordings. Research work has been done to propose approaches to extract it and explore its subsequently interesting applications in information forensics and security. There is a need, however, for further research into the factors and conditions that can promote or hinder the capture of ENF traces in media recordings. This would help us gain a stronger understanding on the situations where the use of the ENF signal can be applicable, and could possibly help us design certain protocols that can increase the likelihood of the capture of ENF traces in recordings.

Through the study shown in this chapter, we have seen that the choice of

recorder used to carry out a recording may have an effect on the strength by which ENF traces are captured and the ENF harmonics at which they are present. Factors related to the environment also play a role: the presence of different sources of waves carrying ENF traces and the interference between these waves and their reflected versions can affect the strength of the captured ENF traces at specific locations within the environment, and moving the recorder while making recording will likely compromise the captured ENF traces.

We have seen in this study that understanding the factors that affect the capture of ENF traces in recordings is not a straightforward problem, as more than one factor is likely contributing to the final state of the captured ENF traces. In the future, it would beneficial to examine factors such as those explored in this chapter, and others, in exhaustive and more controlled experiments that would help achieve a stronger understanding on the situations that promote or hinder the capture of ENF traces and ultimately understand better the real-world applicability of the ENF signal. A similar study can be done with video recordings as well where the role of factors such as internal video camera operations and the behavior of light waves in an environment can be explored.

Chapter 4

ENF-Based Region-of-Recording Identification for Media Signals

4.1 Chapter Introduction

As discussed in Chapter 1, several forensics applications based on the use of ENF signals have been proposed. The ENF has been shown to be useful for detecting tampering or modifications in a multimedia signal [41, 43], and for helping estimate or validate the time-of-recording of the multimedia signal as well as its location-of-recording across grids or within a certain grid [4, 10, 27]. Applications such as these typically need either the knowledge of the grid-of-origin, or concurrent power references from a set of possible grids to identify the grid-of-origin. Aside from high computational costs of exhaustive searches, it may not always be possible to have the needed concurrent references at hand. In this chapter, we present our proposed novel application that seeks to estimate the grid in which an ENF-containing signal

was recorded, without a need for concurrent power references.

We have collected power and audio recordings from eleven different grids around the world[1]. Upon extracting the ENF signals from these recordings, we have noticed that there are differences between them in the nature and manner of their variations. We hypothesized that processing an ENF signal to extract its statistical features may facilitate the identification of the grid, and consequently the region, in which it was recorded. Following this, we devise a machine learning implementation that learns the characteristics of ENF signals from different grids, and uses it to classify ENF signals in terms of their regions-of-recording. Such a system that identifies the grid (region) in which a multimedia signal was recorded, without needing concurrent power references to compare with, can be very important for multimedia forensics and security. It paves the way to identify the origins of such videos as those of terrorist attacks, ransom demands, or child pornography and exploitation [81]. It also substantially reduces computational complexity and facilitates the determination of time/localization information when concurrent references are available by first narrowing down the likely region before a detailed search in time alignment with the proper references can be carried out. In other words, if an investigator is given a video of unknown time and location information, he/she can first use our approach to estimate the grid in which the video was recorded. Following this, the reference data available for the estimated grid can then be used to carry out further forensic operations, such as time-of-recording authentication

[1]We would like to thank Imad Atshan, Yunfang Feng, Berk Gurakan, Jad Hajj-Ahmad, Jana Hajj-Ahmad, Shan He, Wenjun Lu, Michael Luo, Ashwin Swaminathan, and Avinash Varna for their assistance in collecting power and audio recordings from various power grids.

and finer localization estimation [10, 27, 29].

In what follows, we also explore the effect of the type of training data used on the testing results obtained. In particular, we make the distinction between "clean" ENF data extracted from power recordings and "noisy" ENF data extracted from audio recordings. We examine different training scenarios for building multi-conditional learning systems to determine the favorable system set-up to use given the nature of training data available and the expected testing scenarios.

The rest of this chapter is organized as follows. Section 4.2 describes our location-dependant ENF dataset, and examines the differences between the collected ENF signals from different power grids. Section 4.3 presents the proposed features for the machine learning implementation, and discusses the results obtained for building a multiclass region-of-recording classifier. Section 4.4 examines different cases of multi-conditional learning systems. Section 4.5 provides further discussions on the performance of our proposed systems, and Section 4.6 concludes the chapter.

4.2 Location-Dependent ENF Database

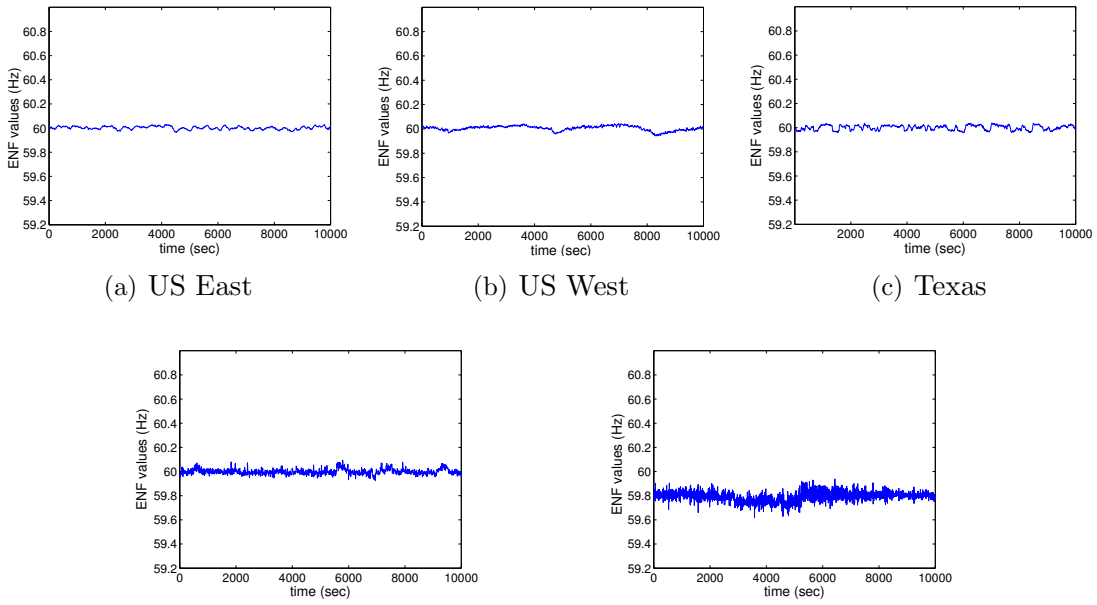
We describe in this section the ENF database, which we have established and will use in this chapter, containing ENF signals from different power grids. We also discuss the observed statistical differences between the ENF signals from different grids to provide intuitions for the features chosen for the machine learning implementation of the region-of-recording classifier.

4.2.1 Database Description

We have collected power and audio recordings from eleven different grids. Among them, five grids have a nominal ENF of 60Hz, and six grids have a nominal ENF of 50Hz. For the 60Hz grids, we have recordings from major North American grids: Eastern North America (or US East for short), Western North America (or US West for short), Texas and Quebec, as well as from aboard a cruise ship that was sailing along the North Atlantic coast of the United States. Among the 50Hz grids, we have recordings from grids in China (Beijing area), India (North Indian grid), Ireland, Lebanon, Tenerife (the largest of the Spanish Canary islands) and Turkey.

For each ENF-containing recording, we divide the signal into non-overlapping frames of length 5 seconds each. We make use of the spectrogram-based spectrum combining approach discussed in Chapter 2 to estimate the dominant frequency that is the instantaneous ENF for each frame. This approach uses the frequency components surrounding multiple harmonics of the ENF to achieve more robust estimates [32]. After computing the instantaneous ENF for each frame, we arrive at the ENF signals for our ENF-containing recordings. Sample ENF signals extracted from power recordings from the different grids can be seen in Figures 4.1 and 4.2.

Upon examining the ENF signals obtained from the different grids, we observe several differences that can be exploited to extract meaningful features for grid classification. We plan to extract features from equally sized ENF signal segments corresponding to recordings of length on the order of minutes. For the results shown



(a) US East (b) US West (c) Texas

(d) Quebec, Canada (e) Cruise Ship

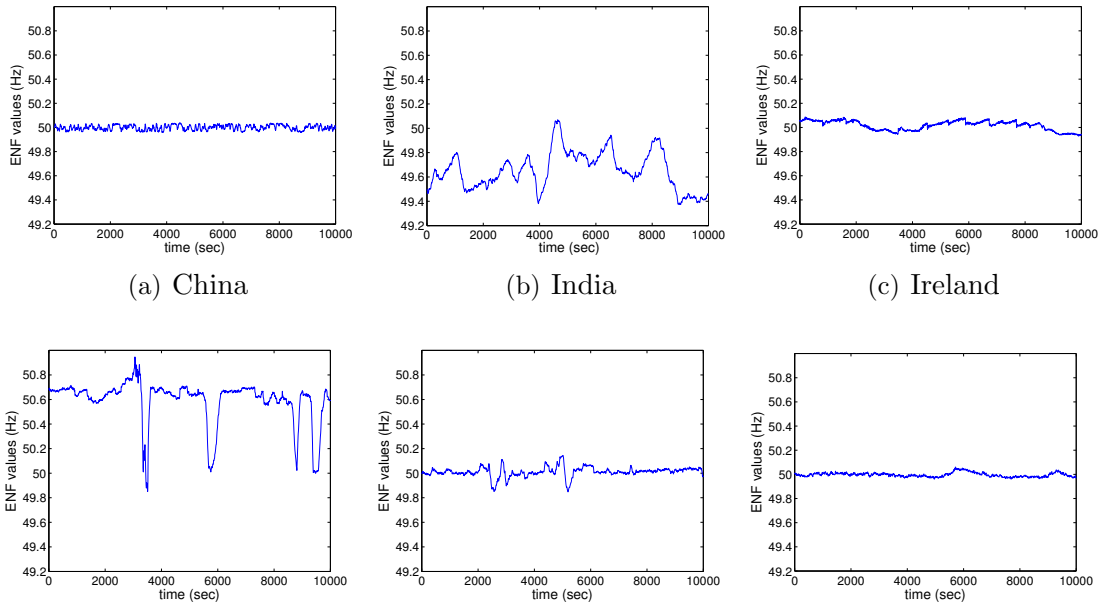
Figure 4.1: Sample ENF signals extracted from power recordings from the 60Hz grids.

in this chapter, we empirically choose 8 minutes as our segment length. Since we are estimating instantaneous ENF for frames of 5 seconds long, this means that our ENF signal segments are of length $S = 96$ samples. Figure 4.3 shows the number of available ENF signal segments, or *examples*, for each grid given this choice of S. These examples will be used for training and testing our machine learning systems.

4.2.2 Comparison of ENF Signals from Different Grids

We examine here the statistical differences observed between ENF signals from different grids. This study motivates a set of features to adopt for our machine learning implementation, which will be discussed in Section 4.3.1.

Examining Figures 4.1 and 4.2, we observe several differences between the ENF signals originating from different grids. The first discerning feature is the mean of an



(a) China (b) India (c) Ireland

(d) Lebanon (e) Tenerife (f) Turkey

Figure 4.2: Sample ENF signals extracted from power recordings from the 50Hz grids.

ENF signal. We can easily tell if a signal belongs to a 50Hz or 60Hz grid depending on how close its temporal average is to either 50Hz or 60Hz. Among signals whose means are similar, there are also some notable differences. For instance, among the 60Hz ENF signals, the Cruise Ship's ENF signal mean falls below 60Hz; and among the 50Hz ENF signals, Lebanon's ENF signal mean can be seen to be above 50Hz most of the time.

The ENF signals also differ in terms of the nature of their variations. Our data shows that among the 60Hz grids, the ENF fluctuations in the US signals seem to be the most controlled, with US East and Texas ENFs showing high similarity in the manner of their variations; the US West ENF appears to drift more before returning to the nominal value. Quebec's ENF exhibits relatively more variations than the US ENFs. The Cruise Ship ENF exhibits the most variations among the 60Hz signals

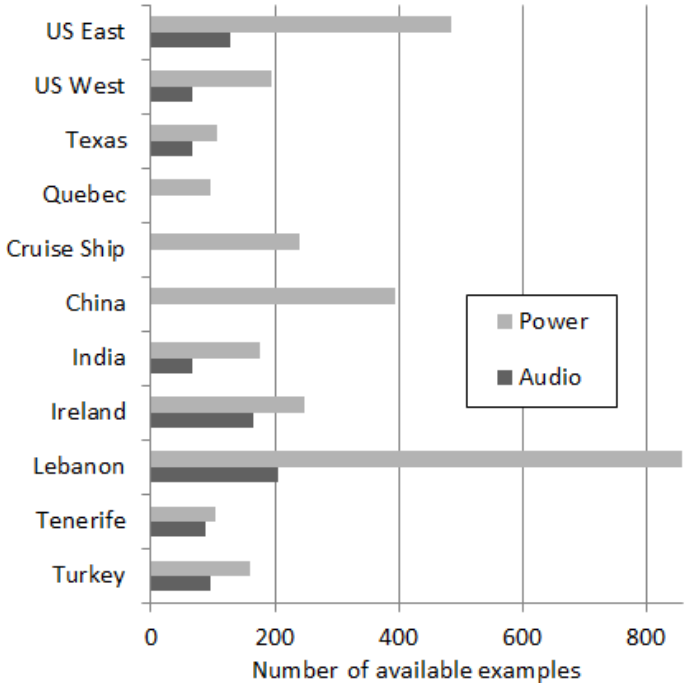


Figure 4.3: The number of available examples (ENF signal segments of size $S = 96$ samples) per grid.

that we have collected so far. Among the 50Hz ENF signals, Ireland and Turkey's ENF signals appear similar and somewhat controlled, although Ireland's ENF shows a tendency to drift before returning to its nominal value. The ENF of China tends to vary at a different rate than the ENFs of Ireland and Turkey. Lebanon's ENF has frequent outliers that drop around 1Hz, a characteristic that does not appear in the other ENF signal samples. The ENF for the Spanish island of Tenerife seems inconsistent as well, at times appearing to be similar to the ENF of Turkey, and at other times exhibiting variations of larger magnitude. India's ENF has a larger range of drift as compared with most other grids in our dataset.

To understand these different ENF variations between grids, we recall that the ENF changes due to load changes in the power grid. The control mechanism reacts to such changes by adjusting the power generation to regulate the ENF and bring

it back towards its nominal value. Different power grids may have different control mechanisms as well as power supply capabilities, therefore affecting the effectiveness and manner in which they are controlling the ENF variations. Typically, larger power grids with more abundant power generation capabilities tend to have smaller frequency variations [1]. Our observations on ENF signals reflect these general characteristics of power grids: the US grids have better control mechanisms and power resources than most other grids that we have observed, so their range of ENF variations is very small (around ± 0.02Hz). The large grids of US and China exhibit smaller variations than the other smaller grids. Lebanon has a very small grid and limited power resources, which is reflected in its grid's large ENF variations. Tenerife is an island and has a small grid, which can explain the inconsistency in its ENF characteristics. Although the Northern Indian grid from which we collected our Indian ENF is fairly large in size, the large ENF variations observed for India may be attributed to limitations in the power resources and control mechanisms governing the grid.

4.3 Multiclass Region-of-Recording Classification

In this section, we explain the proposed multiclass region-of-recording system. The first component of the system is a feature extractor, which extracts the features discussed in Section 4.3.1 for each training example. The second component is the multiclass classifier. We examine different implementations in Section 4.3.2, and present the results in Section 4.3.3.

4.3.1 Feature Extraction and Analysis

After examining the empirical differences among grids in Section 4.2.2, we now discuss quantitative features that can be extracted. We consider having a set of ENF signal segments, $s[n]$'s, of fixed size $S = 96$ samples from candidate power grids. These ENF signal segments are extracted from ENF-containing recordings that are 8 minutes long each, as mentioned in Section 4.2.1.

Following the observations in Section 4.2.2, we adopt, as features, the mean of an ENF segment, as well as the variance of the segment and its dynamic range (the maximum ENF value minus the minimum ENF value). These features are good candidates to facilitate location classification.

To develop other features, we apply a transformation to the ENF segment and then examine the statistical properties of the transformation as potential features. More specifically, we consider Wavelet signal analysis to study the ENF signal segments at multiple time-frequency resolutions. We apply an L-level dyadic wavelet decomposition, where each level provides an approximation to the original signal and the detailed variations at the respective level of resolution [82, 83]. We compute the variances of the high-pass band of each decomposition level (the details) and also the variance of the lowest time-frequency band (the approximation) as candidate features. These wavelet-based features would help us capture the differences in the subtle variations of the ENF among different grids. The wavelet function that we have used to generate the features in this implementation was the discrete approximation of the Meyer wavelet.

Table 4.1: Proposed feature components

Index	**Features**
1	Mean of ENF segment.
2	log(variance) of ENF segment.
3	log(range) of ENF segment.
4	log(variance) of approximation after L-level wavelet analysis (L =9).
5-13	log(variance) of nine levels of detail signals computed through L-level wavelet analysis from coarser to finer (L =9).
14-15	AR(2) model parameters a_1 and a_2.
16	log(variance) of the innovation signal after AR(2) modeling.

Complementing the wavelet features, we extract a set of features obtained from a statistical modeling of the ENF signal. Following recent work that has proposed an autoregressive (AR) model of order 2 for ENF signals [26, 31], an ENF segment $s[n]$ would be modeled as:

$$s[n] = a_1 s[n-1] + a_2 s[n-2] + v[n] \tag{4.1}$$

The original study was made on ENF signals from the United States, but the idea can be extended to examine ENF signals from other grids. We consider three feature values from this AR modeling: the two AR parameters resulting from modeling, a_1 and a_2, and the variance of the model's innovation signal $v[n]$. The AR parameters entail the manner of how samples of $s[n]$ relate with one another, and the variance of the innovation signal is an indicator of how well the signal can be fitted into the AR(2) model (if normalized by the overall signal variance). These features have the potential to help distinguish ENF signals in terms of how well they can fit such an auto-regressive model and in what manner.

The feature components that we use for location classification are summarized in Table 4.1. We apply a log operator on the range and variance feature values to focus on their orders of magnitude and potentially enhance the separability between the final feature values. As mentioned earlier, the feature values are extracted from ENF signal segments of size $S = 96$ samples. The computed feature values are normalized to the range of $[-100, 100]$ by a linear scaling, whereby the k^{th} feature value in a training example is normalized according to the other feature values in position k in all training examples. The normalization parameters are stored and later applied to the testing examples to normalize them. Equations (4.3)-(4.4) summarize the process of normalization for $k = 1, 2, ..., 16$.

$$\mu_k = \frac{1}{M}\sum_{j=1}^{M}\left(\frac{1}{N_j}\sum_{i,l_i=j} f_i[k]\right) \tag{4.2}$$

$$f_i'[k] = f_i[k] - \mu_k \tag{4.3}$$

$$f_i''[k] = 100 \times \frac{f_i'[k]}{\max_i |f_i'[k]|} \tag{4.4}$$

Here, we assume to have M classes, each having N_j examples, with $j = 1, 2, ..., M$, and $\sum_{j=1}^{M} N_j = N$. We denote the label and feature vector of an example i by l_i and f_i, respectively, for $i = 1, 2, ... , N$, and output the final normalized result f_i''.

Figures 4.4 and 4.5 show scatter plots of sample normalized feature values for instances of training data from different grids. We can see that feature points from the same grid tend to form a cluster, whose center is generally separate from the centers of the clusters of other grids. In Figure 4.4, we see that when represented by the variance of three wavelet coefficients at different detail levels, the Cruise Ship and Quebec grids form clusters that have almost no overlap with the US clusters,

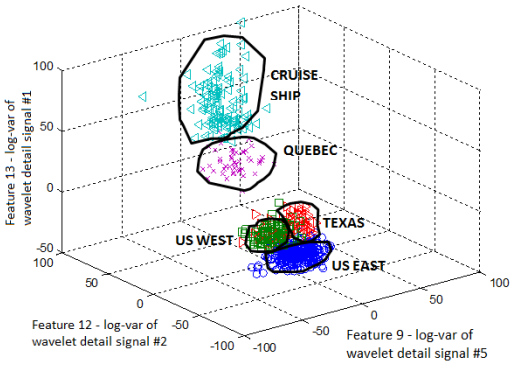
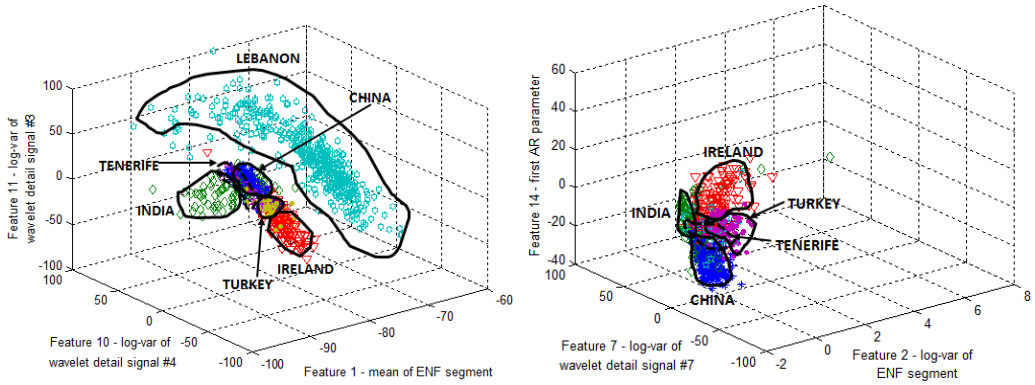


Figure 4.4: Sample feature values for training data instances from 60Hz grids.

and the US clusters appear to have little overlap among themselves.



(a) (b)

Figure 4.5: Sample normalized feature values for training data instances from 50Hz grids.

In Figure 4.5(a), we see that the feature points from Lebanon form a distinct cluster from the clusters of the other 50Hz grids. Examining the other five 50Hz classes, we see in Figures 4.5(b) that four of them (China, India, Ireland and Turkey) can form clusters that have small overlap with one another. The cluster of the fifth class, Tenerife, however, seems to have more overlap with these four classes than they have among each other.

Figures 4.4 and 4.5 suggest that it is possible through a linear classifier to identify such grids as Lebanon and several 60Hz grids. Given the overlapping nature

of other clusters, however, it would be better to build a classifier for region-of-recording classification that is non-linear, which would have a better chance at providing good separation between data from different regions.

4.3.2 SVM Classifier

a) Choice of supervised learning mechanism: We use a Support Vector Machine (SVM) to build the location classifier. In our implementation, we make use of the LIBSVM library [84]. This library implements a multiclass SVM that uses the following approach: For a total of M classes, the system trains $\binom{M}{2}$ binary classifiers; each binary classifier is trained on one of the $\binom{M}{2}$ possible pairs of classes, learning to differentiate between the respective two classes. When testing the trained system, we pass the testing example through all binary classifiers, and assign votes to each possible class based on which class emerges as the winner from each binary classification task. The final winning class is the class with the largest number of votes. For a testing example, the LIBSVM implementation also provides M probability (confidence) values, where the j^{th} probability value gives the probability that the testing example belongs to the j^{th} class.

The numbers of available examples for training and testing that we have for each grid are shown in Figure 4.3, with the segment size S of 96 samples. Due to logistical and resource constraints in collecting recordings from various grids around the world, the data that we have is imbalanced: We have significantly more recordings from some grids versus others, and we do not have audio recordings from

all the grids considered. This imbalance in training data can create overfitting or bias problems when testing the system. If a system is trained on a dataset where the majority of the training examples belong to one class, it tends to be more biased in the testing scenario to assign the testing examples to this majority class [85]. To tackle this issue, we use a variant of SVM called the weighted SVM, which is supported by LIBSVM.

SVM implementations usually include a fixed cost value C, which controls the penalty on making a mistake while classifying an example. The weighted SVM addresses the issue of imbalanced data through assigning different cost values for examples from different classes. The larger class has a smaller cost value than the smaller class, which means that the penalty for making a mistake on an example from the smaller class would be larger [85]. Here, with M classes, the cost for class j that has N_j training examples would be $w_j \cdot C$ where:

$$w_j = \frac{N_{min}}{N_j}, \text{ for } j = 1, 2, ..., M \text{ and } N_{min} = \min_j N_j. \tag{4.5}$$

In our implementations, we use the non-linear Radial Basis Function (RBF) kernel for our SVMs. Using the LIBSVM library, two important parameters need to be chosen for the RBF kernel: the cost parameter C and a parameter γ, that relates to how far the influence of a single training example reaches. For each SVM classifier we train, we select these two parameters through cross-validation.

b) Systems to be trained: As shown in Figure 4.3, the data that we have are of two main types: ENF segments extracted from either power recordings or audio recordings. Generally, ENF segments extracted from power recordings are cleaner

signals with high signal-to-noise ratio (SNR), while the ENF segments extracted from audio recordings can be noisy. This would affect the quality of the feature values extracted. To gain a better understanding on the effect of ENF data types on the performance of a region-of-recording machine learning system, we train three SVM systems with different sets of training data and compare the testing results. These three systems are shown in Table 4.2.

Table 4.2: Description of the trained SVM systems

System	**Num. of classes**	**Training dataset**
I	$M = 11$ classes	Only ENF signals extracted from power recordings.
II	$M = 8$ classes	Only ENF signals extracted from audio recordings.
III	$M = 11$ classes	ENF signals extracted from both power and audio recordings, assuming that signals of both types from the same grid belong to the same class.

As mentioned earlier, for each testing example, the LIBSVM implementation provides a probability value giving its confidence in its decision on the region-of-recording. We make use of this feature when exploring three scenarios for our trained systems. In the first scenario, we take the SVM system's decision as the final decision. In the second and third scenarios, if the confidence is lower than a set threshold (e.g., 0.6), we advance this example to the next stage. For the second scenario, this next stage assigns the testing example a decision of "None of the Above". For the third scenario, we subject the example with such initial low confidence to a final binary SVM classifier trained on the two classes that received the highest confidences in the first stage.

Table 4.3: Accuracies for different systems averaged over 20 rounds.

Training	**Testing**	**No second stage**	**with "None of the Above"**	**with binary classifiers**
I: power	power	85.6%	77.3%	**88.4%**
	audio	36.5%	28.0%	37.0%
II: audio	power	47.5%	36.7%	48.6%
	audio	78.0%	51.9%	**84.3%**
III: power + audio	power	83.7%	71.7%	**87.5%**
	audio	74.7%	56.2%	**81.7%**

4.3.3 Results and Discussions

a) Results overview: We present here the results obtained for testing the data on the three systems of Table 4.2 for the three different scenarios discussed above: taking the SVM system's decision as final, or using one of the two second stages (including a "None of the Above" option or using a final binary classifier). We empirically choose 0.6 as the threshold of the confidence value for a testing example to be advanced to the second stage. For each system, we show the power and audio testing results separately to understand how well the system is suited for each type of data.

We first divide all the available data into six groups so that each group has approximately the same number of examples from every grid. Then, we train each of our three SVM systems for 20 rounds, considering in each round a different combination of three groups out of the six as training data and the remaining three groups as testing data. The results shown here are averaged over the results of these 20 training/testing rounds.

Table 4.3 shows the accuracies achieved for testing the different systems using the second stage options discussed in Section 4.3.2. To compute these accuracies, we

first compute the identification accuracy for each class, i.e., for class j, the identification accuracy is equal to the ratio of the correctly identified testing examples from this class to the total number of testing examples from this class. The accuracies in Table 4.3 report the averages of these identification accuracies across all the classes. We opted for this accuracy measure to avoid biasing our results by the performance of classes with a larger number of testing examples.

In Table 4.3, we first notice that for each training/testing combination, the highest accuracy is consistently achieved in the scenario where we use binary classifiers in the second stage, followed by the baseline scenario where there is no second stage, and the lowest accuracy is when we use the "None of the Above" option. To understand these results, we consider the scenario where we do not have a second stage. If we opt to disqualify the decision of any testing example in case of low confidence, the overall accuracy drops because we are losing the correct decisions that have a low confidence. If, on the other hand, we opt to use the binary classifiers, we give the examples that were on the borderline of making a correct decision an opportunity to rectify the decision.

To compare the testing results of the different systems, we consider the accuracies resulting from using binary classifiers, as this is currently our best-case scenario. We can see that testing the power data on a system trained on power data (i.e., Systems I and III) results in a high testing accuracy of about 88%, while testing this power data on the system trained only on audio data (i.e., System II) results in a low accuracy of only about 49%. Similarly, testing the audio data on a system trained on audio data (i.e., Systems II and III) results in high accuracies within

Table 4.4: Confusion matrix for power ENF testing data on System I – Accuracies(in %) averaged over 20 rounds.

	Testing Classes	Num. of examples	US East	US West	Texas	Queb.	Cruise Ship	Chi.	Ind.	Ire.	Leb.	Tene.	Turk.
Power	**US East**	242	**98.9**	0.8	0.3	-	-	-	-	-	-	-	-
	US West	97	1.5	**94.3**	4.2	-	-	-	-	-	-	-	-
	Texas	53	5.8	23.0	**71.2**	-	-	-	-	-	-	-	-
	Quebec	47	-	-	-	**100.0**	-	-	-	-	-	-	-
	Cruise Ship	120	-	-	-	-	**100.0**	-	-	-	-	-	-
	China	197	-	-	-	-	-	**96.2**	0.2	0.2	-	1.6	1.8
	India	90	-	-	-	-	-	4.6	**90.3**	-	0.3	4.4	0.4
	Ireland	125	-	-	-	-	-	0.2	0.1	**95.6**	-	2.1	2.0
	Lebanon	429	-	-	-	-	-	-	0.2	-	**99.8**	-	-
	Tenerife	53	-	-	-	-	-	17.9	9.6	13.3	-	**37.7**	21.5
	Turkey	81	-	-	-	-	-	3.0	0.5	4.7	-	3.2	**88.6**

Table 4.5: Confusion matrix for audio ENF testing data on System II – Accuracies (in %) averaged over 20 rounds.

	Testing Classes	Num. of examples	US East	US West	Texas	India	Ireland	Lebanon	Tenerife	Turkey
Audio	**US East**	65	**92.4**	4.8	2.8	-	-	-	-	-
	US West	34	7.5	**91.2**	1.3	-	-	-	-	-
	Texas	33	4.7	3.5	**91.8**	-	-	-	-	-
	India	34	-	-	-	**77.1**	0.3	6.5	13.5	2.6
	Ireland	82	-	-	-	-	**89.6**	0.4	3.6	6.4
	Lebanon	103	-	-	-	3.0	0.4	**92.3**	3.8	0.5
	Tenerife	45	-	-	-	9.0	14.2	4.3	**61.3**	11.2
	Turkey	48	-	-	-	0.4	12.4	2.1	6.1	**79.0**

81-85%, while testing this audio data on a system not trained on audio data (i.e., System I) results in a low accuracy of only 37%. This shows that it is desirable to incorporate, into the training process, data of the same set of signal conditions that the system would be anticipated to have in testing.

b) Close examination of results: To understand these results better, we examine the confusion matrices of the three systems considered. As Systems I and II proved to be ineffective for classifying data on which they are not trained on, we forgo the testing of audio data on System I or power data on System II. The confusion matrices for the three systems are shown in Tables 4.4, 4.5 and 4.6, respectively. In each of these tables, the labels of the rows denote the actual grid (region) and condition of the signals tested (power or audio), while the labels of the columns denote the predicted

Table 4.6: Confusion matrix for power and audio ENF testing data on System III – Accuracies (in %) averaged over 20 rounds

	Testing Classes	Num. of examples	US East	US West	Texas	Queb.	Cruise Ship	Chi.	Ind.	Ire.	Leb.	Tene.	Turk.
Power	**US East**	242	**99.2**	0.6	0.2	-	-	-	-	-	-	-	-
	US West	97	4.3	**91.9**	3.8	-	-	-	-	-	-	-	-
	Texas	53	15.9	23.4	**60.7**	-	-	-	-	-	-	-	-
	Quebec	47	-	-	-	**100.0**	-	-	-	-	-	-	-
	Cruise Ship	120	0.1	-	0.1	-	**99.8**	-	-	-	-	-	-
	China	197	-	-	-	-	-	**95.6**	0.2	0.2	-	3.2	0.8
	India	47	-	-	-	-	-	4.3	**91.7**	-	0.2	3.6	0.2
	Ireland	125	-	-	-	-	-	-	0.1	**93.7**	0.4	3.4	2.4
	Lebanon	429	-	-	-	-	-	-	-	-	**100.0**	-	-
	Tenerife	53	-	-	-	-	-	14.1	13.6	8.5	0.3	**44.4**	19.1
	Turkey	81	-	-	-	-	-	2.5	0.2	4.9	-	7.0	**85.4**
Audio	**US East**	65	**85.6**	5.5	8.7	0.2	-	-	-	-	-	-	-
	US West	34	4.1	**92.8**	3.1	-	-	-	-	-	-	-	-
	Texas	33	4.8	2.9	**91.8**	-	0.5	-	-	-	-	-	-
	India	34	-	-	-	-	-	-	**79.8**	0.7	2.0	16.3	1.2
	Ireland	82	-	-	-	-	-	-	-	**89.2**	-	1.8	9.0
	Lebanon	103	-	-	-	-	-	0.3	5.9	0.2	**91.6**	1.9	0.1
	Tenerife	45	-	-	-	-	-	0.2	16.7	16.9	-	**42.5**	23.7
	Turkey	48	-	-	-	-	-	-	1.3	12.9	0.9	4.7	**80.2**

grid by the system. The entries in the diagonals of the tables, highlighted in bold face, show the correct identification accuracy for each class. The tables show how well the testing examples from each grid and signal condition were classified when applied to our trained systems. We include the number of testing examples available for each class to highlight the difference in the number of examples available for each class to help present the proper context for the corresponding testing percentages.

As mentioned earlier, the accuracies that we are examining here are the ones resulting from using binary classifiers in the second classification stage. We can see that by incorporating the mean ENF value as a feature, the 50Hz signals are never mistaken for 60Hz signals, and vice versa.

Considering the power signals, we can see that the correct identification accuracies in Tables 4.4 and 4.6 fall in the high range of 90-100% for all signals except those of Texas, Tenerife and Turkey. Among the 60Hz signals, we can see that the Quebec and Cruise Ship signals are notable for their consistently near-perfect identi-

fication rates, due to the clear distinction in the range and nature of their variations as compared with the more controlled US signals. US East and US West signals can be mistaken for each other or for Texas, which is understandable given the close similarity between them in control mechanisms and power resources. Texas signals have notably lower identification rates (71.2% and 60.7%), being mistaken for the other US signals often. Texas is a smaller grid than the other two US grids, and as mentioned earlier, this can at times result in relatively larger and less predictable frequency variations, while at other times share similar behavior as other US grids. This would make it more difficult to define the properties of the Texas ENF signals, and would confuse them with the properties of the larger US grids.

With regards to the 50Hz signals, the high accuracies achieved for China, Lebanon, India and Ireland can be attributed to the general separability of their feature values from the features values of other 50Hz signals, as discussed in Section 4.2. The identification accuracy for power signals for Turkey falls in the lower range of 85-89%. The identification accuracy is the lowest in the case of Tenerife, being 38% for System I and 44% for System III. These low values for Tenerife seem to reflect the observations made in Section 4.2. The Tenerife ENF is inconsistent in its statistical properties, as exemplified in Figure 4.2(e); and from the limited amount of data that we have collected, its feature values from various data acquisition sessions exhibit the most overlap with the feature values from other 50Hz grids, as exemplified in Figure 4.5. This would negatively affect the identification accuracies of signals from other grids that become mistaken to be from Tenerife. It is likely that if Tenerife had not been included as a training class, the identification accuracy

for Turkey, for instance, would have been higher.

Considering the audio signals in Tables 4.5 and 4.6, we can see a similar trend of correct identification accuracies among grids to the trend observed with the power signals, although the values generally are lower. The range for correct classification rate for the 60Hz US signals is 85-93%, and that of Ireland and Lebanon is 89-93%. It is notable among the 60Hz signals that Texas has a good identification accuracy, when compared with the results on power recordings. It seems that the noisier nature of the audio ENF allowed better identification for Texas signals, while affecting somewhat negatively the performance on US East signals in System III. As noted in Section 4.2, there is a high similarity in ENFs between US East and Texas ENFs, and this might cause the system to have difficulty differentiating one from the other. The identification accuracies of India and Turkey are in the lower range of 77-81%, and Tenerife achieves the lowest identification accuracies (61% and 43%). Again, the presence of Tenerife as a training class with inconsistent properties can explain its low identification accuracies, and can affect the correct identification for other grids, i.e., India and Turkey.

The general drop in identification accuracies with the audio signals as compared with the power signals can be explained by the nature of the audio ENF signals. The ENF signals extracted from audio signals are more susceptible to noise than those extracted from power signals, and the amount of noise and distortions affecting the ENF signal estimates can be different even within signals of the same class due to different recording conditions. This can create confusions for the machine learning system when defining the class boundaries and could lead to more

mistakes in identification. Another reason is the fewer amount of audio data available to us in our established database.

Comparing Systems I and II with System III, we can see from Table 4.3 that the correct identification accuracies for System III are, with few exceptions, similar to or slightly lower than their counterparts in Systems I and II. System III defines classes as a mixture of audio and power ENF signals, which means that signals belonging to one class may have a larger range of differences from one another due to their varying noise levels.

Overall, by incorporating training examples of multiple conditions, we have developed a machine learning based system (System III) that achieves a high accuracy on identifying the region/grid-of-origin of ENF signals extracted from both power and audio recordings. Meanwhile, if the test signal's noise condition is known a priori or can be estimated well, classification performance may have some further improvement by employing a system that is well trained on signals from its corresponding conditions (Systems I and II).

4.4 Noise Adaptation using Multi-Conditional Learning

We have observed in Section 4.3 that a mismatch in the training and testing conditions can lead to lower accuracy values for correct identification of the region-of-recording of ENF signals. As shown in Table 4.3, the percentage of correctly identified ENF signals extracted from audio recordings is around 81-85% when testing on a system trained on audio ENF data (System II or III), and drops to 48% when

the system used for testing is only trained on power ENF data (System I). In this section, we explore a multi-conditional learning system that can adapt to changes in the noise environment between the training and testing data. Such multi-conditional learning systems have been used in speech technology literature, in which the problem of varying noise conditions is prevalent for such tasks as speaker recognition and speech understanding [86, 87].

4.4.1 System Model

The proposed system model assumes to have a set of K different noise conditions for the training data. We start with a separate SVM multiclass classifier, such as the ones discussed in Section 4.3, for each noise condition. The data needed for training can be datasets coming from known separate noise conditions or can be generated synthetically. For instance, if we have clean ENF signals extracted from power recordings, we can add to them synthetic white Gaussian noise (WGN) to obtain several sets of the same ENF signals at various SNR levels. We use each set to generate feature values that will be used to build its corresponding SVM classifier.

When subjecting a testing example to the system, each SVM classifier gives M confidence values, with the j^{th} value denoting the probability that the testing example belongs to the j^{th} class. To reach a final decision for the testing example from the results of the K classifiers, we use a Bayesian framework. Such a framework requires the knowledge of the likelihood of observing the testing example for each noise condition.

We denote the K training datasets, referring to the K noise conditions, by ϕ_i, with $i = 1, 2, ..., K$. We represent the outputs of the K classifiers to a testing example $\mathbf{x}$ in a $K \times M$ matrix, where the component $p_{i,j}$ in the $(i,j)^{th}$ entry denotes the (estimated) probability that $\mathbf{x}$ belongs to class j assuming it belongs to the noise condition ϕ_i. Following this, we can express the probability of the final decision D that $\mathbf{x}$ belongs to class j as:

$$p(D = j|\mathbf{x}) = \sum_{i=1}^{K} p(D = j|\phi_i, \mathbf{x})p(\phi_i|\mathbf{x}) \tag{4.6}$$

$$= \sum_{i=1}^{K} p_{i,j}(\mathbf{x})p(\phi_i|\mathbf{x}). \tag{4.7}$$

Here, $p(\phi_i|\mathbf{x})$ can be obtained using the Bayes formula as:

$$p(\phi_i|\mathbf{x}) = \frac{p(\mathbf{x}|\phi_i)p(\phi_i)}{\sum_{i'=0}^{K} p(\mathbf{x}|\phi_{i'})p(\phi_{i'})}, \tag{4.8}$$

where $p(\phi_i)$ is the prior probability of noise condition i.

Assuming that $p(\phi_i)$ is uniform for all i, we can combine Equations (4.7) and (4.8) to write:

$$p(D = j|\mathbf{x}) \propto \sum_{i=1}^{K} p_{i,j}(\mathbf{x})p(\mathbf{x}|\phi_i) \tag{4.9}$$

The decision rule $\delta(\mathbf{x})$ for $\mathbf{x}$ can now be expressed as:

$$\delta(\mathbf{x}) = \arg\max_j p(D = j|\mathbf{x}) \tag{4.10}$$

$$= \arg\max_j \sum_{i=1}^{K} p_{i,j}(\mathbf{x})p(\mathbf{x}|\phi_i). \tag{4.11}$$

We obtain the $p_{i,j}(\mathbf{x})$ values from the K SVM classifiers. For the $p(\mathbf{x}|\phi_i)$ values, we learn the conditional distributions of $\mathbf{x}$ for each of the K noise conditions

using the Gaussian mixture models (GMM) approach. GMMs are universal approximations of densities, i.e., given a sufficient number of mixture components, they can approximate any distribution [88]. This will allow us to compute the $p(\mathbf{x}|\phi_i)$ values.

4.4.2 Experimental Setup

For this experiment, we start with the same dataset as in the experiment of Section 4.3, for which the number of examples available per grid for a single training/testing round is shown in Figure 4.3. We exclude using the data of grids from which we do not have audio recordings, i.e., Quebec, Cruise Ship and China. In addition, we exclude the data from Tenerife due to the high variability and unpredictability of its ENF observed in this very limited amount of data.

Our experiment explores the benefits of the noise adaptation approach using multi-conditional learning presented in Section 4.4.1 as compared with the previous approach of Section 4.3 of training one SVM classifier without modifying the training data or accounting for different noise conditions. Given a set of training data, we first build a baseline system of a multiclass classifier using weighted SVM. We then add synthetic noise to our training data to generate the noisy training datasets, and build the system of K classifiers described in Section 4.4.1. We train four types of multi-conditional systems as listed in Table 4.7, which differ by the design of the K considered noise conditions.

For all SVM classifiers built in this experiment, we do not make use of a second

Table 4.7: Description of trained multi-conditional systems

Type	**Training data type**	**Num. of classifiers**	**Description of training data for "noise" conditions ϕ_i's, $i = 1, ..., K$**
1	power	K =6	Power ENF signals for $i = 1$ and power ENF signals with added synthetic noise to achieve SNRs {20, 15, 10, 5, 0}dB for $i = 2, ...,$ 6, respectively.
2	audio	$K = 6$	Audio ENF signals for $i = 1$ and audio ENF signals with added synthetic noise to achieve SNRs {20, 15, 10, 5, 0}dB for $i = 2, ...,$6, respectively.
3	power + audio	$K = 7$	Type 1 conditions for i =1, ...,6 and audio ENF signals for i =7.
4	power + audio	K =12	Type 1 conditions for i =1, ..., 6 and Type 2 conditions for $i = 7, ...,$12.

stage (such as using additional binary classifiers for borderline cases) in order to keep the focus on examining the merits of noise adaptation. We train data from seven grids, namely, US East, US West, Texas, India, Ireland, Lebanon, and Turkey, using the features listed in Table 4.1.

4.4.3 Results and Discussions

The results of our experiment are shown in Table 4.8. The accuracies shown are averaged over 20 training/testing rounds, following the same procedure of data preparation and accuracy computations as in Section 4.3.3.

The major advantage of noise adaptation appears when testing ENF signals extracted from audio recordings on a system that has only ENF signals from power recordings to work with. When we do not adapt for noise, the identification accuracy is only 46%; when we use noise adaptation, the identification accuracy of audio ENF data rises to 74%. In this case, the training system has learned the

Table 4.8: Results of testing on multi-conditional systems for seven grids

Training	**Testing**	**Noise adaptation?**	**Accuracy**
power	power	No	90.6%
		Yes (Type 1)	86.8%
	audio	No	46.4%
		Yes (Type 1)	**74.4%**
audio	power	No	51.0%
		Yes (Type 2)	49.4%
	audio	No	81.9%
		Yes (Type 2)	73.4%
power + audio	power	No	88.4%
		Yes (Type 3)	87.0%
		Yes (Type 4)	86.3%
	audio	No	83.3%
		Yes (Type 3)	75.6%
		Yes (Type 4)	76.4%

feature characteristics of noisy ENF signals (through the K classifiers trained on synthetically noisy power ENF signals), and thus works better at identifying the region-of-recording of noisy audio ENF signals than the baseline system that has only learned the characteristics of clean power ENF data.

In the meantime, we notice that the noise adaptation procedure has little effect on testing ENF signals extracted from power recordings. In all such cases, the identification accuracy decreases by about 2-4% in the noise adaptation case from the results on our baseline systems. This is understandable as the power ENF data is generally clean and has high SNR, so testing on a system trained on ENF data of different noise conditions offers little benefit to it, nor does it do much harm especially if one of the noise conditions is the "clean" condition. Similarly, testing the ENF signals extracted from audio recordings on noise adaptation systems where audio training data is originally available provides no improvement on testing this

same data on baseline systems whose training data contained audio ENF data.

Overall, we can see that the noise adaptation approach can provide a better alternative system in cases of mismatch between the noise conditions of training and testing data. In particular, the identification accuracy for the region-of-recording of noisy ENF signals extracted from audio recordings is substantially improved in the case where our training data is limited to clean ENF signals extracted from power recordings.

4.5 Further Discussions

In this section, we provide further discussions on the performance of our proposed systems. We examine the effect of dimensionality reduction schemes in Section 4.5.1 and discuss grids with varying power profiles in Section 4.5.2.

4.5.1 Dimensionality Reduction

Dimensionality reduction schemes are known to be helpful to facilitate efficient implementations of machine learning in many applications involving a high dimension of features. To examine their effects on our problem, we have experimented with different dimensionality reduction schemes, specifically, the Fisher's Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) [89]. Using LDA, we reduce the dimensionality of our data from the original dimensionality of 16 to $M-1$, where M is the number of classes. Using PCA, we examine different dimensions ranging from $M-1$ to 15. We compare whether or not either of

Table 4.9: Accuracies (averaged over 20 rounds) for single stage classification systems with different dimensionality schemes

Training	**Testing**	**No dimensionality reduction**	**LDA**	**PCA**
I: power	power	**85.6%**	84.9%	85.0% for 14 dims
	audio	36.5%	36.7%	37.9% for 13 dims
II: audio	power	47.5%	12.5%	49.7%for 10 dims
	audio	**78.0%**	13.1%	70.7% for 9 dims
III: power	power	**83.7%**	81.5%	81.2% for 15 dims
+ audio	audio	**74.7%**	74.4%	72.9% for 14 dims

the dimensionality reduction schemes improves on the testing results of the trained systems.

Table 4.9 shows the accuracies achieved for testing the different systems discussed in Section 4.3.2, with no second stage (including a "None of the Above" option or using binary classifiers). For PCA, we show the best accuracy achieved among the dimensions considered. We can see that in most cases, the accuracies of either LDA or PCA are similar to or slightly worse than the accuracies achieved without dimensionality reduction. This is likely due to having a small dimension of only 16 features to begin with, thus the effect from dimensionality reduction in conjunction with SVM with radial basis kernel is not significant.

Such dimensionality reduction schemes may be useful in future work, where the number of feature components considered could be much larger, and where a stronger need to choose a small amount of useful features for discrimination could arise.

4.5.2 Grids with Varying Profiles

A potential factor impacting the effectiveness of ENF-based location classification is the variability of a grid's ENF attributes with time. The power profile of a certain grid may exhibit different types of behaviors, depending on the time of a day or the season of a year. For our dataset, we have made sure to collect data from both daytime and nighttime, and when possible, we have tried to collect data from different times in a year. The power profile may also change with time due to changes that might occur to the power system, such as changing the regulating control mechanisms or the generation/supply capabilities. In addition, this profile may be sensitive to changes in the load, depending on its particular operation procedures. These changes can be handled by periodically incorporating new training data to update the classifier.

In order to examine the effect of a grid's changing ENF profile, we carry out the following study. We have noted earlier that Lebanon's ENF has frequent large outliers that deviate from its more stable states; an example of this can be seen in Figure 4.2(d). In a new experiment, we manually divide the ENF signals from Lebanon into two categories, one denoted by *Lebanon-stable* and the second by *Lebanon-volatile*. The first category contains Lebanon ENF signal segments that do not have large outliers, and the second contains those that do.

We repeat the training/testing experiment for the scenario that has given the best results in Section 4.3, which was System I trained on power data and tested on power data. We now train on twelve classes, replacing the original Lebanon

Table 4.10: Confusion matrix for testing Lebanon data in the new setting – Accuracies (in %) averaged over 20 rounds

Testing Classes	**Lebanon Stable**	**Lebanon Volatile**	**Other**
Lebanon-stable	96.6	3.3	0.1
Lebanon-volatile	20.5	79.3	0.2
Other	-	0.0056	99.99

class with the two classes, Lebanon-stable and Lebanon-volatile. The testing results for the Lebanon classes can be seen in Table 4.10. In our results, the Lebanon-stable ENF testing data were identified 96.6% of the time as Lebanon-stable, and 3.3% as being Lebanon-volatile. The Lebanon-volatile ENF testing data were identified 79.3% of the time as being Lebanon-volatile, and 20.5% of the time as being Lebanon-stable. This shows that when using a 12-class multi-class classifier, the majority of the Lebanon-volatile data being "wrongly" classified are actually classified to the correct grid (i.e., Lebanon). This reflects that even for periods or seasons with volatile behavior, it still contains attributes seen at other times for the same grid.

If we consider a classification of a Lebanon ENF example to be correct if it is classified to either the Lebanon-stable or the Lebanon-volatile category, this brings the correct classification rate of the Lebanon ENF testing data to 99.8% in this experiment. The average correct classification rate for this experiment over the data from all eleven grids is 87.6%, which is comparable to the accuracy of the previous classifier in Table 4.3 (88.4%).

This study suggests one approach for tackling the possible variation of ENF grid characteristics with time. A complementary approach, as mentioned earlier,

would be to periodically collect new reference data from the classes a classifier is being trained on, and use them to update the database used for training the system. This would be especially useful in cases of changes in power management plans and equipment.

Another extension to our work is to identify the country-of-recording of a media signal. Large countries, such as the US, China and India, tend to have more than one interconnected power grid within the country. We have seen this in our study as we included three US grids in our experiments (US East, US West and Texas). If the goal of a forensic study were to be identifying the country-of-origin of a recording, our proposed approach can be used as a first step in estimating the grid-of-origin of an ENF-containing signal. Following that, a testing result pointing to a grid in a country would suggest the origin of the recording to be that country.

Examining the US results in Tables 4.4, 4.5, and 4.6, this gives us 100% country-of-origin identification accuracy in nearly all cases. The accuracy is only slightly lower when testing audio-ENF on System III, which was trained on both power and audio ENF segments. For this system, US East testing examples are identified as originating from the US 99.8% of the time, and the corresponding identification accuracies for US West and Texas are 100% and 99.5%, respectively. Even though these same-country grids are not interconnected and do not have the same ENF variations, they likely share similar equipment and control mechanisms, making their ENF variations more similar to one another than to those from grids of other countries. Further experiments with a larger number of grids can certainly provide more validation to these results.

A case of interest is when one interconnected grid spans several countries, such as the case of the synchronous grid of Continental Europe. If the goal of a forensic study was to identify the country-of-recording of a media signal, ENF signals from the European grid would fall under one class in a grid-of-origin machine learning implementation. Once the classifier determines a media signal belongs to the European grid, further ENF-based intra-grid location analysis may be done to estimate the country-of-origin [29].

Furthermore, due to logistical and resource constraints in collecting the recordings, as mentioned earlier, we had imbalanced data in our studies. This resulted in having much more recordings from certain grids, such as Lebanon and US East, compared to a much smaller number of recordings in some other grids, such as Tenerife. In an ideal situation, we would like a large number of recordings from all grids studied in order to better capture the variability in the characteristics of ENF signals.

4.6 Chapter Summary

In this chapter, we have presented a machine learning based system that can identify the grid-of-origin of an ENF signal without needing concurrent power references. ENF signals from different power grids display different statistical characteristics, which can be exploited to identify the power grid from which they originated. These differences are attributed to the size of power grids and the techniques and available power/energy resources by which the grids are controlled and operated.

We have presented and compared three machine learning systems that are trained on identifying the origin of ENF signals embedded in clean power signals and/or noisy audio signals from different grids. We were able to achieve an average accuracy of 88.4% on identifying ENF signals extracted from power recordings from eleven candidate power grids, and an average accuracy of 84.3% on identifying ENF signals extracted from audio recordings from eight candidate power grids.

In addition, we have explored using multi-conditional systems that can adapt to cases where the noise conditions of the training and testing data are different. This approach was able to improve the identification accuracy of noisy ENF signals extracted from audio recordings by around 28% when the training dataset is limited to clean ENF signals extracted from power recordings.

This work presents a new capability of using ENF signals for multimedia forensics, by identifying the grid (region) of origin of an audio/video recording via extracting and classifying its ENF signal. This work can also reduce substantial computational complexities in traditional ENF-based time/location analysis problems where there is a large number of potential grids-of-origin for a media recording being studied.

As a first work of its type, there is room for further improvement. Because we have been constrained by the available resources for data collection in this first work, we anticipate that the future collection of more ENF data from different times can help examine the resilience of the extracted features over time. Another direction is to explore additional features and incorporate advanced feature selection approaches to improve the learning system. We hope that further work can build upon this first

work to achieve higher performance for a large number of power grids.

The work discussed in this chapter has been used as the basis of the 2016 edition of the Signal Processing Cup (SP Cup), a global undergraduate competition organized by the IEEE Signal Processing Society. In this third edition of the SP Cup, this competition engaged participants from nearly thirty countries. 334 students from 23 countries formed 52 teams that registered for the competition. Among them, more than 200 students in 33 teams turned in the required submissions by the open competition deadline in January 2016. The top three finalist teams were invited to attend the IEEE Conference on Acoustics, Speech, and Signal Processing in Shanghai, China in March 2016, where they presented their final work in front of a panel of judges. More information on the SP Cup 2016 can be found in [90].

Chapter 5

Exploiting Power Signatures for Camera Forensics

5.1 Chapter Introduction

In this chapter, we explore a novel application of the ENF signal that is targeted at characterizing the video camera producing an ENF-containing video. This can be particularly useful in scenarios where there is a need to verify that a suspect owns a camera that produced a suspicious video. The approach proposed in this chapter, which is inspired by our work on flicker forensics to be discussed in Chapter 6, works within a completely nonintrusive scenario where solely analyzing an ENF-containing video can shed some light on its origins.

Our focus here is on videos recorded using the widely used group of video cameras equipped with Complementary Metal-Oxide-Semiconductor (CMOS) image sensors that employ a *rolling shutter*. Unlike a camera employing a *global shutter*

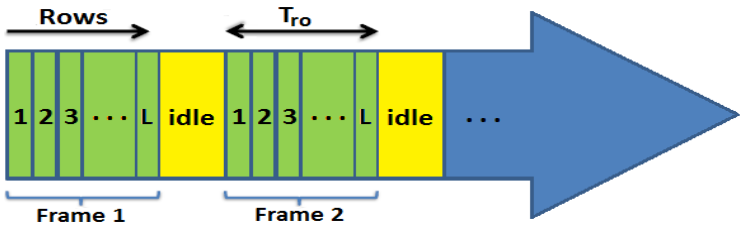


Figure 5.1: Timing of rolling shutter sampling: the L rows of a frame are sequentially exposed, followed by an idle period before proceeding to the rows of the next frame.

that acquires all the pixels of a video frame at the same time, a camera employing a rolling shutter acquires a video frame one row at a time. Although this sequential read-out mechanism of rolling shutter has traditionally been considered detrimental to image/video quality due to its accompanying artifacts, recent work has shown that it can be exploited with computer vision and computational photography techniques to produce interesting results [91, 92]. Recent work on ENF, for instance, has made use of the rolling shutter towards improved ENF extraction from videos [8, 31, 36].

Figure 5.1 illustrates the timing for image acquisition with rolling shutter. Each row of the frame is sequentially exposed to light followed by an idle period before proceeding to the next frame [36]. The amount of time during which a camera acquires the rows of a video frame, which we denote by the read-out time T_{ro}, is specific to the camera and is a value that is not typically mentioned in its user manual or specifications list. In this chapter, we characterize the camera producing an ENF-containing video by estimating its T_{ro} value.

The rest of this chapter is organized as follows. Section 5.2 explains our model for signal capture and the proposed approach for estimating the read-out time of a camera recording an ENF-containing video; Section 5.3 discusses the experimental set-up and results; and Section 5.4 summarizes this work.

5.2 Model and Proposed Approach

ENF traces are embedded in the visual track of video recordings due to the near-invisible flickering of electric lighting, i.e., through the changing intensity of electric lighting captured by the camera. The electric light intensity relates to the supplied electric current via a power law thus making its nominal frequency twice the nominal ENF value, i.e., 120Hz in North America and 100Hz in most other parts of the world. Following this, the electric light signal can be modeled as a sinusoid:

$$x(t) = A_{\mathrm{e}} \sin(2\pi \tilde{f}_{\mathrm{e}} t + \phi) \tag{5.1}$$

where $\tilde{f}_{\mathrm{e}} := f_{\mathrm{e}}(t)$ represents a variable frequency, corresponding to the ENF component that fluctuates around 100/120Hz, and A_{e} and ϕ are the magnitude and phase, respectively.

In what follows, we model the light intensity signal captured in a video by a camera [93]. Then, we describe our proposed approach to estimate the camera's read-out time.

5.2.1 Modeling the Captured Signal

The process in which a camera acquires a video can be seen as a two-step process. First, integration of photons happens over a duration ΔT, which is related to the camera's *shutter speed*. Second, the camera samples the resulting integration signal.

The integration phase can be modeled as a convolution of the light signal $x(t)$ with a rectangular integration window $h(t)$ whose Fourier transform can be written as $H(f) = \Delta T \operatorname{sinc}(f\Delta T)$. The Fourier transform of the signal obtained is then:

$$Y(f) = X(f) \cdot H(f) \tag{5.2}$$

$$= \frac{A_\mathrm{e}}{2}\left[e^{-j\phi}\delta(f-\tilde{f}_\mathrm{e}) + e^{j\phi}\delta(f+\tilde{f}_\mathrm{e})\right] \cdot H(f) \tag{5.3}$$

$$= \frac{A_\mathrm{e}\Delta T}{2}\operatorname{sinc}(\tilde{f}_\mathrm{e}\Delta T)\left[e^{-j\phi}\delta(f-\tilde{f}_\mathrm{e}) + e^{j\phi}\delta(f+\tilde{f}_\mathrm{e})\right] \tag{5.4}$$

This allows us to write $y(t)$ as $\tilde{A}\sin(2\pi\tilde{f}_\mathrm{e}t + \phi)$, where $\tilde{A} = \frac{A_\mathrm{e}\Delta T}{2}\operatorname{sinc}(\tilde{f}_\mathrm{e}\Delta T)$.

To model the sampling phase of the camera's acquisition of the light intensity signal, we first need to define the camera's sampling rate. To begin with, we write the camera's frame rate f_c as $f_\mathrm{c} = 1/T_\mathrm{c}$, where T_c is the frame period. T_c includes the period of time T_ro required to sample the L rows of a frame and possibly an additional idle time period. Since the camera being considered in this chapter employs a rolling shutter, each row is sampled at a different time, so the sampling rate that we are interested in is not f_c, but rather $f_\mathrm{s} = 1/T_\mathrm{s}$, where T_s is the time between subsequent row read-outs. For modeling purposes, we assume that if the camera read-out is performed continuously at a rate of f_s for the entire frame period T_c, i.e., in a case where there is no idle time, then the camera would in principle be able to read-out M rows for the duration of T_c where $M \geq L$, with L being the actual number of rows in a frame. Following this, we can express the camera's sampling rate f_s as:

$$f_\mathrm{s} = \frac{1}{T_\mathrm{s}} = \frac{M}{T_\mathrm{c}} = \frac{L}{T_\mathrm{ro}}. \tag{5.5}$$

In this setting, M is unknown, but L can be found by examining the video height in the video's metadata.

We denote the sampled signal by $s[n]$ for $n \in \mathbb{N}$, which can be written as:

$$s[n] = y(nT_{\mathrm{s}}) = \tilde{A}\sin\left(2\pi\tilde{f}_{\mathrm{e}}T_{\mathrm{s}}n + \phi\right) \text{ for } n \in \mathbb{N}. \tag{5.6}$$

The intensity value $s[n]$ is the light intensity captured by all the pixels in the n^{th} row. To make the relation clearer, we write n as $n = kM + l$, where k and l are the frame and row indices, respectively, such that $k \in \{0, 1, 2, ..., F-1\}$ and $l \in \{0, 1, 2, ..., M-1\}$, with F being the number of frames in the video.

Replacing n by $kM + l$ in Equation (5.6), and using Equation (5.5), we obtain:

$$\begin{aligned} s[k,l] &= \tilde{A}\sin\left(2\pi\tilde{f}_{\mathrm{e}}\frac{T_{\mathrm{c}}}{M}kM + 2\pi\tilde{f}_{\mathrm{e}}T_{\mathrm{s}}l + \phi\right) &(5.7)\\ &= \tilde{A}\sin\left(2\pi\frac{\tilde{f}_{\mathrm{e}}}{f_{\mathrm{c}}}k + 2\pi\frac{\tilde{f}_{\mathrm{e}}}{f_{\mathrm{s}}}l + \phi\right). &(5.8)\end{aligned}$$

Since a video camera's frame rate f_{c} typically falls in the range of 24–60Hz, we would have $\tilde{f}_{\mathrm{e}} > f_{\mathrm{c}}$. To account for aliasing, we write $\tilde{f}_{\mathrm{e}}$ as:

$$\tilde{f}_{\mathrm{e}} = \tilde{f}_{\mathrm{a}} + mf_{\mathrm{c}}, \text{ where } m \in \mathbb{N} \text{ and } \tilde{f}_{\mathrm{a}} \in [-f_{\mathrm{c}}/2, f_{\mathrm{c}}/2]. \tag{5.9}$$

We can now write $s[k,l]$ as:

$$\begin{aligned} s[k,l] &= \tilde{A}\sin\left(2\pi\frac{\tilde{f}_{\mathrm{a}}}{f_{\mathrm{c}}}k + 2\pi mk + 2\pi\frac{\tilde{f}_{\mathrm{e}}}{f_{\mathrm{s}}}l + \phi\right) &(5.10)\\ &= \tilde{A}\sin\left(2\pi\frac{\tilde{f}_{\mathrm{a}}}{f_{\mathrm{c}}}k + 2\pi\frac{\tilde{f}_{\mathrm{e}}}{f_{\mathrm{s}}}l + \phi\right) &(5.11)\\ &= \tilde{A}\sin\left(\tilde{\omega}_{\mathrm{a}}k + \tilde{\omega}_{\mathrm{b}}l + \phi\right), &(5.12)\end{aligned}$$

where $\tilde{\omega}_{\mathrm{a}} = 2\pi\tilde{f}_{\mathrm{a}}/f_{\mathrm{c}}$ is expressed in *radians/frame* and $\tilde{\omega}_{\mathrm{b}} = 2\pi\tilde{f}_{\mathrm{e}}/f_{\mathrm{s}}$ is expressed in *radians/row*.

5.2.2 Proposed Approach

In this section, we first describe the *vertical phase method* on which we based our proposed approach. We then explain how we adapt the *vertical phase method* to our ENF-based approach in a practical setting.

5.2.2.1 Vertical Phase Method

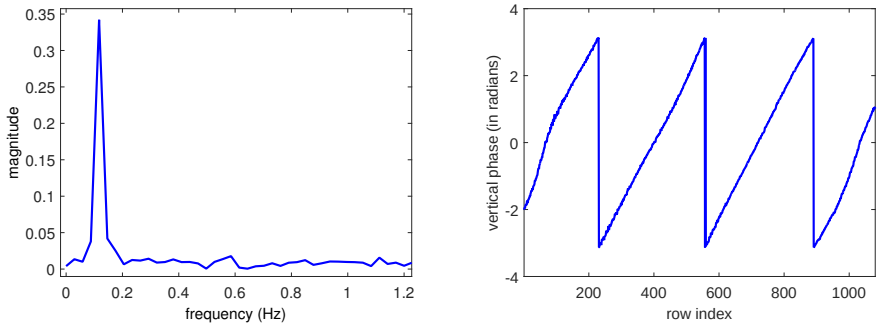
This method examines the evolution of the embedded intensity signal, $s[k, l]$, over frames, and computes the vertical radial frequency $\tilde{\omega}_{\mathrm{b}}$ to aid the estimation of the read-out time T_{ro}. By exploiting $\tilde{\omega}_{\mathrm{b}}$'s relation to the delay between ENF traces in adjacent rows, we can estimate it through analyzing the phase shift in the discrete-time Fourier transforms (DTFTs) of the row intensity signals.

For simplicity, we first assume that the time-varying parameters involved, namely, $\tilde{\omega}_{\mathrm{a}}, \tilde{\omega}_{\mathrm{b}}, \tilde{f}_{\mathrm{a}}$ and $\tilde{f}_{\mathrm{e}}$, are all constant at their respective nominal values. We will relax this assumption in Section 5.2.2.2.

The first step is obtaining an estimate for the aliased frequency $\tilde{\omega}_{\mathrm{a}}$. To do that, we examine the following modification of Equation (5.12):

$$s_{\mathrm{l}^*}[k] = \tilde{A} \sin\left(\tilde{\omega}_{\mathrm{a}} k + \tilde{\omega}_{\mathrm{b}} l^* + \phi\right). \tag{5.13}$$

Here, we fix the row index l to a certain value l^*, and the resulting signal $s_{\mathrm{l}^*}[k]$ as a function of the frame index k is a sinusoid of frequency $\tilde{\omega}_{\mathrm{a}}$. An estimate of $\tilde{\omega}_{\mathrm{a}}$ can then be obtained by finding the frequency that shows a peak in the Fourier transform of $s_{\mathrm{l}^*}[k]$. We can equivalently find an estimate of $\tilde{f}_{\mathrm{a}}$ by using $\tilde{f}_{\mathrm{a}} = \tilde{\omega}_{\mathrm{a}} f_{\mathrm{c}}/2\pi$



(a) (b)

Figure 5.2: Results of applying *vertical phase method* on a video taken by the back camera of an iPhone 5. (a) shows the Fourier transform of s_{l^*} exhibiting peak close to the expected $\tilde{f}_{\mathrm{a}}$, (b) shows the linear vertical phase.

and the known frame rate f_{c}. An example can be seen in Figure 5.2(a), where a peak of the Fourier transform is visible close to the expected $\tilde{f}_{\mathrm{a}}$.

The next step is to obtain an estimate of $\tilde{\omega}_{\mathrm{b}}$. To do that, we compute the value of the DTFT of $s_{\mathrm{l}}[k]$ at $\tilde{\omega} = \tilde{\omega}_{\mathrm{a}}$ for each case of $l \in \{0, 1, ..., L-1\}$. We compile the values into a vector of size L, resulting in $S_{\tilde{\omega}_{\mathrm{a}}}[l]$ that we denote as the *vertical Fourier transform*. The phase component $\Phi_{\tilde{\omega}_{\mathrm{a}}}[l]$ of $S_{\tilde{\omega}_{\mathrm{a}}}[l]$ can be written as:

$$\Phi_{\tilde{\omega}_{\mathrm{a}}}[l] = \tilde{\omega}_{\mathrm{b}} l + \phi. \tag{5.14}$$

An example of this *vertical phase* can be seen in Figure 5.2(b), wrapped between $[-\pi, \pi]$. After *unwrapping* this vertical phase, $\tilde{\omega}_{\mathrm{b}}$ can be estimated from the slope using linear regression.

Now that we have an estimate of $\tilde{\omega}_{\mathrm{b}}$, we can compute the T_{ro} estimate. Examining the definition of $\tilde{\omega}_{\mathrm{b}}$, and using Equation (5.5), we can write it as:

$$\tilde{\omega}_{\mathrm{b}} = 2\pi \cdot \frac{\tilde{f}_{\mathrm{e}}}{f_{\mathrm{s}}} = 2\pi \cdot \frac{\tilde{f}_{\mathrm{e}}}{L/T_{\mathrm{ro}}}. \tag{5.15}$$

Algorithm 1 Proposed approach to compute T_{ro} estimate.

1: Pre-process the video for analysis.
2: Find if the nominal frequency $\bar{f}_{\text{e}}$ is 100Hz or 120Hz.
3: Assign the origin frequency $f_{\text{o}} := k\bar{f}_{\text{e}}$, where $k := 1$.
4: Compute the aliased frequency of f_{o} as: $f_{\text{a,o}} = f_{\text{o}} - mf_{\text{c}}$, such that $m \in \mathbb{N}$ and $f_{\text{a,o}} \in [-f_{\text{c}}/2, f_{\text{c}}/2]$.
5: Find the frequency in $f_{\text{a,o}}$'s vicinity with the ones with the most linear vertical phase, and estimate the corresponding slope $\hat{\omega}_{\text{b}}$.
6: **if** the vertical phase is sufficiently linear,
7: Compute T_{ro} as: $T_{\text{ro}} = (L \cdot \hat{\omega}_{\text{b}}) \,/\, \left(2\pi k\bar{f}_{\text{e}}\right)$.
8: **else**
9: Assign $f_{\text{o}} := k\bar{f}_{\text{e}}$, where $k := k + 1$.
10: Go to Line 4.

Thus, we can use the known values of the frame height and the nominal ENF value to estimate the read-out time T_{ro} via:

$$T_{ro} = \frac{L \cdot \tilde{\omega}_{\text{b}}}{2\pi\tilde{f}_{\text{e}}}. \tag{5.16}$$

5.2.2.2 Adapting to a Practical Setting

Based on the above method, we now discuss how to modify it to obtain the T_{ro} value from the embedded ENF traces in a practical setting. We first need to account for the time-varying nature of the parameters that were assumed constant in the previous discussion. We also need to account for the fact that the embedded ENF traces may not always be strong enough around the nominal 100/120Hz value. In practice, it is not uncommon for the ENF traces to be more strongly captured at higher harmonics of the nominal frequency than at the nominal value [32, 48].

The steps of our proposed approach are outlined in Algorithm 1. The first step is to prepare the ENF-containing video for analysis. The pre-processing operations involved here can vary depending on the video at hand, with the goal of making

the embedded ENF detectable with high signal-to-noise ratio (SNR). A number of these enhancement operations have been discussed in [8]. Examples of such operations include identifying static regions in the videos that are more favorable for ENF extraction, compensating for camera motion, and compensating for brightness changes caused by a camera's automatic brightness control mechanism. After pre-processing, for each frame of the video, we obtain a 1-D vector of size L, a *frame signal*, where the l^{th} entry corresponds to the contents of the frame's l^{th} row.

Next, we must ascertain whether the nominal ENF value, $\bar{f}_{\mathrm{e}}$, is 100Hz or 120Hz. This can be done by examining the time-frequency content in the recorded video. To do so, we connect the frame signals from consecutive frames and compute the Fourier transform of the resulting signal. We use the *nominal row sampling rate*, defined as the product of the video's frame rate, f_{c}, and the frame height, L, as a unit of reference corresponding to the Fourier transform's frequency axis. Plotting the Fourier transform would then reveal peaks at the frequency values of $\tilde{f}_{\mathrm{e}}$ shifted by multiples of f_{c} [36]. If these peaks appear close to $120 + n \cdot f_{\mathrm{c}}$, then $\bar{f}_{\mathrm{e}} = 120$Hz, and if they appear close to $100 + n \cdot f_{\mathrm{c}}$, then $\bar{f}_{\mathrm{e}} = 100$Hz ($n \in \mathbb{N}$).

Assigning $f_{\mathrm{o}} := \bar{f}_{\mathrm{e}}$ and via Line 4 of Algorithm 1, we compute the aliased frequency $f_{\mathrm{a,o}}$ where we expect to find the ENF traces. If the ENF were constant at the nominal value, we would proceed to compute the vertical phase at $f_{\mathrm{a,o}}$ and the corresponding slope. As the time-varying ENF is likely not to be at its nominal value during the recording of the video, the corresponding aliased frequency might not lie at the calculated $f_{\mathrm{a,o}}$. To account for this, we sample frequencies in the vicinity of $f_{\mathrm{a,o}}$ and compute their corresponding vertical phases. Among the candidates, we

select the most linear vertical phase, where linearity is assessed based on the root mean square (RMS) error of the linear regression.

Ideally, the estimated slope of the vertical phase can reveal the read-out time using Equation (5.16). However, if the vertical phase is not linear enough, and the regression RMS error is not small enough, the final estimate will be incorrect. We have empirically found that a threshold of 0.04 for the regression RMS error is a good cut-off value to avoid obtaining erroneous estimates. This may happen when the ENF is not strongly captured at the nominal value. In such a case, the ENF, if present, may be more reliably captured at higher frequencies than at the nominal value, so we assign f_{o} to be the next harmonic of $\bar{f}_{\mathrm{e}}$ and repeat the procedure.

If a low RMS error cannot be achieved for several iterations, it may become necessary to improve the pre-processing operations [8].

5.3 Experiment and Results

In this section, we describe the experiment carried out to test the proposed approach, and discuss the results obtained.

5.3.1 Experimental Set-up

We have recorded short videos (30-75 seconds long each) using five different video cameras in environments where there is electric lighting in Maryland, USA. The aim of the experiment is to analyze each of the videos using the proposed approach of Section 5.2.2 and estimate the read-out time T_{ro} of the video's camera.

In order to evaluate the accuracy of our ENF-based T_{ro} estimates, we need to compute ground truth values for the T_{ro} values of the cameras at hand. We have employed a protocol described in [94] for this purpose. Along with the camera to be characterized, this protocol requires a Liquid Crystal Display (LCD) screen, and a photo-dioide equipped circuit that takes as an input a light signal and records it as a digital signal.

As will be elaborated more about in Chapter 6, LCD screens do not emit light on their own, but rather are equipped with a *back-light* that emits a light signal passing through the screens array of liquid crystal cells to produce images. This light signal is a periodic signal with frequency f_{bl}. We use the photo-dioide equipped circuit to record the back-light signal of an LCD screen, and then analyze the Fourier transform of the recorded digital signal to estimate the screen's back-light frequency f_{bl}. Afterwards, using a video camera that we wish to characterize, we take a short video, of about one minute long, by camcording a uniformly grey screen displayed on the LCD screen with the now known f_{bl} value. In this scenario, the f_{bl} is analogous to the ENF value $\tilde{f}_{\text{e}}$, and the video being taken by the camera will capture the back-light signal in a similar way to how the camera would have captured the electric light signal carrying ENF traces. The benefit here is that, in this controlled setting, the captured signal, i.e., the flicker signal, has a high SNR. This will allow us to use the *vertical phase method* of Section 5.2.2.1, where we replace $\tilde{f}_{\text{e}}$ by the obtained f_{bl}, to obtain a highly confident estimate for the camera's T_{ro} value that we can use as ground truth for subsequent experimental evaluations.

Table 5.1: Cameras used in our experiments

Camera ID	**Model**	**L**	**T_{ro} (ms)**
1	Sony Cybershot DSC–RX 100 II	1080	13.4
2	Sony Handycam HDR–TG1	1080	14.6
3	Canon SX230–HS	240	18.2
4	iPhone 5 front camera	720	22.9
5	iPhone 5 back camera	1080	27.4

We have carried out this protocol using two LCD screens on the five cameras at our disposal. Table 5.1 shows the full details for the cameras.

5.3.2 Results and Discussions

We have applied the proposed approach of Section 5.2.2.2 on the videos taken by the five cameras in Table 5.1. For the videos of Cameras 1 and 2, we have found good T_{ro} estimates based on the ENF traces of the second harmonic, while for Cameras 3, 4, and 5, we have found good T_{ro} estimates based on the ENF traces of the base nominal frequency.

Table 5.2 shows the results obtained for the five videos. We can see that we have obtained excellent T_{ro} estimates for all the cases, with the relative error being within 1.5%.

Table 5.2: Estimated T_{ro} values of considered videos using our proposed approach

Camera ID	1	2	3	4	5
Expected T_{ro} (ms)	13.4	14.6	18.2	22.9	27.4
Estimated T_{ro} (ms)	13.60	14.75	18.41	23.13	27.63
Relative Error (%)	1.5	1.0	1.2	1.0	0.8

We have clearly benefited from having short videos in this experiment, as the US ENF generally remains well controlled and does not vary much within such a short time window. In the case where longer videos are to be analyzed, it would be advisable to divide the videos into shorter segments and analyze each separately so as not to be affected negatively by the changing ENF value over time.

5.4 Chapter Summary

In this chapter, we have presented an ENF-based forensic application, whereby we are able to analyze an ENF-containing video to characterize the camera that produced the video. This is done by estimating the camera's read-out time, or the time needed to read one frame, which is typically less than the frame period for the commonly used cameras equipped with rolling shutter. We have tested our proposed nonintrusive approach on short ENF-containing videos taken using five different cameras, where we have seen high performance in estimating read-out times.

This work shows the potential for the ENF traces captured in a video to characterize the camera producing the video. It can provide corroborating evidence in cases where a video is linked to a suspect owning a certain camera. In future work, we plan to examine a wider range of cameras to investigate the variability in read-out time values, and thus better understand the broad applicability and performance of this approach. We also plan to investigate further video camera characteristics that can be extracted based on analyzing the captured ENF traces.

Chapter 6

Flicker Forensics for Camcorder Piracy

6.1 Chapter Introduction

Movie piracy remains a major concern today that jeopardizes the sustainability of the entertainment industry. Unauthorized disclosure of copyrighted material prior to theatrical or DVD/Blu-ray releases significantly harms box office revenues. To address this risk, it is a common practice for content owners to rely on cryptography-based content protection techniques such as Conditional Access Systems (CAS) or Digital Rights Management (DRM) to secure multimedia content along the distribution pipeline [52]. Nevertheless, such protection eventually has to be lifted to present the content to the end-user, and a pirate can then place a camera in front of the screen showing the content to record a pirated copy of the movie.

A mitigating strategy consists of embedding forensic watermarks within the rendered content, which can survive the digital-analog-digital conversion [53]. As a

result, when a pirated copy surfaces on an unauthorized distribution platform, it is possible to recover the underlying watermark identifier and trace it back to the user or device from which the piracy originated [54]. This piracy deterrence mechanism is already in place in professional environments (e.g., for reviewing pre-theatrical release movie screeners or in digital cinemas [95]). Due to a recent specification by the motion picture industry that mandates the use of forensic watermarking for ultra high definition content [96], these traitor tracing watermarks are also likely to soon reach the homes of consumers.

In this context, movies are likely to be displayed on the widely used Liquid-Crystal-Display (LCD) screens, and it is therefore relevant to evaluate what kind of distortion may appear when recording such a display. Early works on camcorder piracy recently focused on the ability to recognize this type of piracy through the use of discriminating features. Camcorder piracy can indeed be revealed by the presence of tell-tale visual artifacts. Examples of such artifacts include the presence of a luminance flicker due to the interplay between the screen and the camcorder [58], the presence of combing artifacts due to the interlaced display of the screen [55], the presence of global motion indicating that the camera is hand-held [56], ghosting artifacts due to the integration of several frames by the camcorder [57], a statistical deviation of the color bias and saturation [97], a statistical deviation of the edges orientation due to the capture geometry [98], blurry edges due to the recapture [99], and the presence of Moiré due to spatial aliasing [100]. These techniques rely on the design of features affected by these visual artifacts and they are then fed to a state-of-the-art classification tool.

Figure 6.1: Flicker artifact when recording an LCD screen displaying a uniformly gray frame with a camcorder.

In this work, we focus on a single visual artificat due to camcorder piracy, namely, the *luminance flicker*. This artifact appears when the camcording process involves the use of an LCD screen. Although camcorded videos made with other types of display screens are out of the scope of this chapter, related works indicate that the flicker artifact also occurs in other piracy scenarios [58]. As depicted in Figure 6.1, this flicker signal is routinely incarnated by dark and bright stripes that scroll up or down the recaptured video. Our objective is to go a step further into understanding this visual artifact, beyond the binary classification of "camcorded" vs. "not camcorded", a problem studied early on to improve digital cameras [101]. In Section 6.2, we first review the internal mechanisms of an LCD screen and a camcorder at the origin of the flicker and derive a parametric model to describe this periodic signal. We then derive a piracy identity that connects some settings of the pirate devices and a parameter of the flicker signal. We discuss in Section 6.3 various estimation techniques to recover this flicker parameter from camcorded videos in practice. In Section 6.4, we report experimental results that demonstrate that it is

possible to accurately link camcorded video content to their associated pirate devices based on the piracy identity. Next, in Section 6.5, we discuss how to recover the shape of the flicker signal present in a camcorded video and illustrate that it can help in recognizing different back-light technologies of the LCD screen. In Section 4.6, we summarize our findings and outline research directions for future work.

6.2 Modeling the Flicker Signal

The flicker signal originates from the interplay between the LCD screen and the camcording that essentially yields some aliasing. For simplicity, our model breaks down the acquisition pipeline into three stages: (i) the emission of a back-light signal by the screen, (ii) the integration of the light emitted by the screen with a sensor of the camcorder, and (iii) the sequential sampling of the different rows of a video frame. For reference, Table 6.1 enumerates the parameters that we have used to model the luminance flicker.

Table 6.1: Parameters used in the modeling of the flicker signal

Parameter	**Description**
f_{bl}	frequency of back-light signal
T_{bl}	period of back-light signal
θ	duty cycle of back-light signal
T_{ss}	camcorder's integration period related to its shutter speed
f_{c}	frame rate of camcorder
T_{c}	frame period of camcorder
f_{s}	sampling rate of camcorder
T_{s}	sampling period of camcorder
T_{ro}	read-out time of camcorder
T_{idle}	idle time of camcorder, equal to $T_{\mathrm{c}} - T_{\mathrm{ro}}$
R	number of rows in video frame
R^{+}	assumed number of rows in video frame if $T_{\mathrm{idle}} = 0$

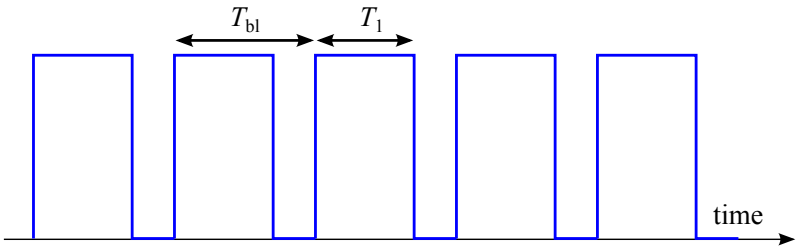


Figure 6.2: Simplified model of the back-light signal as a periodic rectangular signal with period T_{bl}, shown here with a duty cycle $\theta = T_{\text{l}}/T_{\text{bl}} = 0.7$.

6.2.1 LCD Screens and Back-light

Nowadays, most TV sets and computer screens incorporate an LCD screen. The image is formed by the light that is let through by an array of liquid crystal cells, each cell encoding a pixel of the image [102]. Each individual crystal cell can be tuned by changing the electric potential applied at the bounds of the cell. As a result, the liquid crystal lets more or less light pass, thereby producing the luminance of the corresponding pixel. The colors of the image are obtained by interleaving a color filter in between the liquid crystals and the surface of the screen. In other words, by design, LCD screens need a source of light, defined as the *back-light*, to illuminate the liquid crystal array from behind.

In practice, the back-light is a periodic signal whose frequency f_{bl} is high enough to be imperceptible to the human eye, typically within the range of 120Hz to 1kHz. As depicted in Figure 6.2, the back-light signal $b(t)$ is assumed to be periodic with a frequency $f_{\text{bl}} = 1/T_{\text{bl}}$ and a duty cycle $\theta = T_{\text{l}}/T_{\text{bl}}$:

$$b(t) = \Pi_{T_{\text{l}}}(t) * \sum_{i=-\infty}^{+\infty} \delta(t - i\,T_{\text{bl}}), \tag{6.1}$$

where $\delta(.)$ is the unit impulse function, $*$ is the convolution operator, and $\Pi_T(.)$ is

the following rectangular signal:

$$\Pi_T(t) = \begin{cases} 1 & \text{if } -T/2 < t < T/2, \\ 0 & \text{otherwise.} \end{cases} \tag{6.2}$$

Our empirical observations confirmed that such a model for the back-light signal is not unrealistic, even if it may be more or less rectangular depending on the back-light technology. The intensity of the back-light can be adjusted by tuning the duty cycle θ offering various brightness settings for the screen. The spectrum of the back-light signal in the Fourier domain is then given by:

$$B(f) = \theta \sum_{i=-\infty}^{+\infty} \text{sinc}(i\,\theta)\,\delta(f - i\,f_{\text{bl}}). \tag{6.3}$$

6.2.2 Light Integration with a Camcorder

Camcorders have an array of sensors that are in charge of converting captured photons into electrical charges, which are eventually translated as pixel values. A sensor is exposed to light for a given period of time T_{ss}, which is related to the shutter speed for the camcorder. All photons reaching the sensor during this integration phase are converted into electrical charges and are accumulated. At the end of the integration period, the current is discharged to be translated as a pixel value.

If such a sensor is directly hit by the back-light signal defined in Equation (6.1), the amount of light accumulated over the integration period T_{ss} is given by:

$$a(t) = b(t) * \Pi_{T_{\text{ss}}}(t). \tag{6.4}$$

In the Fourier domain, this convolution becomes a multiplication, and the spectrum is given by:

$$A(f) = \theta\, T_{\text{ss}} \sum_{i=-\infty}^{+\infty} A_i\, \delta(f - i\, f_{\text{bl}}), \tag{6.5}$$

where $A_i = \text{sinc}(i\,\theta)\ \text{sinc}(i\, f_{\text{bl}}\, T_{\text{ss}})$. In other words, the spectrum reduces to a collection of frequency rays regularly spaced according to the back-light frequency f_{bl}. As a result, the integration signal $a(t)$ can be written as a sum of cosines at the harmonics of the back-light:

$$a(t) = 2\,\theta\, T_{\text{ss}} \sum_{i=0}^{+\infty} A_i\ \cos(2\pi\, i\, f_{\text{bl}}\, t). \tag{6.6}$$

6.2.3 Sampling with a Rolling Shutter

Most camcorders commercially sold today use Complementary Metal-Oxide-Semiconductor (CMOS) sensors and rely on a *rolling shutter* [103, 104]. In contrast to global shutters that acquire a whole frame at once by discharging all sensors in one go, a rolling shutter captures each row of the frame sequentially (e.g., from top to bottom) as depicted in Figure 6.3. As a result, each row of the image is exposed to a different portion of the back-light and the integrated luminance will therefore differ for different rows, resulting in the *flicker* artifact, expressed as a more or less visible variation of luminance along the vertical axis in the recorded video, as illustrated earlier in Figure 6.1.

The frame acquisition period $T_{\text{c}} = 1/f_{\text{c}}$ is the sum of the readout time T_{ro}, required to acquire all R rows of the frame, and an idle period T_{idle} [36]. For modeling purposes, we assume that the camcorder is able to read $R^{+} \geq R$ rows at its full

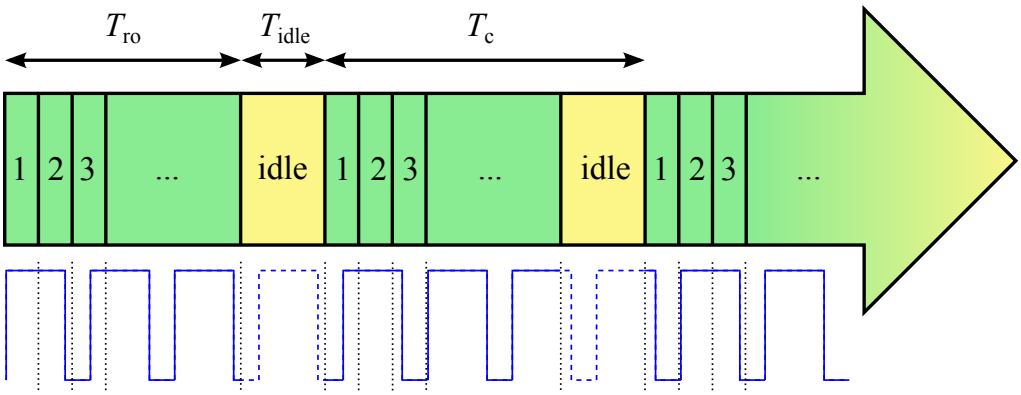


Figure 6.3: Rolling shutter illustrated. The rows of a video frame are acquired sequentially and followed by an idle transition period (T_{idle}) before moving to the next frame. The back-light signal is depicted below to hint that a camcorder sensor integrates a different portion of the back-light signal for each row, thereby possibly creating visible artifacts.

capacity and that only R of them are retained in forming the final video frames while others are discarded. In other words, the sampling rate of the camcorder is given by:

$$f_{\text{s}} = \frac{1}{T_{\text{s}}} = \frac{R^{+}}{T_{\text{c}}} = \frac{R}{T_{\text{ro}}}. \tag{6.7}$$

For simplicity, we only consider below the flicker component resulting from the first cosine component in Equation (6.6) (i.e., $i = 1$). We also disregard the DC component (i.e., $i = 0$) as it is unlikely to create the variation of luminance that we are trying to explain. The sampling of the back-light signal $b(t)$ at the rate f_{s} therefore yields the following discrete signal $f[n]$:

$$f[n] = A_1 \cos\left(2\pi\, f_{\text{bl}}\, n T_{\text{s}}\right), \text{ for } n \in \mathbb{N}. \tag{6.8}$$

This discrete signal can be seen as the recorded luminance for each row when the camcorder records a uniformly gray image. The cosine in Equation (6.8) clearly indicates that sensors associated to different rows will output different luminance values.

To appreciate the spatio-temporal nature of this model, let us rewrite the sampling index n as a combination of the frame index k and the row index $r \in \{0, 1, \ldots, R^+ - 1\}$, namely, $n = kR^+ + r$. Equation (6.8) can then be rewritten as:

$$f[k, r] = A_1 \cos\left(2\pi\, k\, \frac{f_{\mathrm{bl}}}{f_{\mathrm{c}}} + 2\pi\, r\, \frac{f_{\mathrm{bl}}}{f_{\mathrm{s}}}\right). \tag{6.9}$$

In practice, the sampling rate f_{s} is usually much larger than the Nyquist sampling rate of a regular back-light signal with frequency f_{bl}. On the other hand, a camcorder typically operates at a frame rate f_{c} between 24 and 60 frames-per-second (fps), which is lower than a regular back-light frequency f_{bl}. We define the aliased frequency f_{a} as follows:

$$f_{\mathrm{bl}} = f_{\mathrm{a}} + m\, f_{\mathrm{c}}, m \in \mathbb{N}, \tag{6.10}$$

where m is chosen such that $f_{\mathrm{a}} \in [-f_{\mathrm{c}}/2, f_{\mathrm{c}}/2]$. Equation (6.9) then simplifies to:

$$f[k, r] = A_1 \cos(k\, \omega_k + r\, \omega_r), \tag{6.11}$$

where $\omega_k = 2\pi f_{\mathrm{a}}/f_{\mathrm{c}}$ is a temporal radial frequency expressed in radians-per-frame and $\omega_r = 2\pi f_{\mathrm{bl}}/f_{\mathrm{s}}$ is a vertical radial frequency expressed in radians-per-row. A phase term θ can be added to the argument of the cosine in Equation (6.11) to account for the absence of synchronization between the back-light of the screen and the shutter of the camcorder.

6.2.4 Discussion

The derivations detailed in the previous sections provide some insight on the visual artifacts, such as the ones depicted in Figure 6.1, produced when a camcorder

records a screen that displays a uniformly gray image. For each row of the frame, the associated light-capturing sensors of the camcorder see a different portion of the back-light signal and thus produce different luminance levels due to the duty cycle of the back-light. When such captured rows are displayed on top of one another, one can notice luminance variations along the vertical axis whose frequency is given by the term ω_r in Equation (6.11).

In addition, as mentioned earlier, camcorders usually have a frame rate that is notably lower than the frequency of the back-light of the LCD screen. As a result, camcording for piracy is prone to temporal aliasing. Visually, it makes the vertical luminance variation pattern scroll up/down the screen as indicated by the term ω_k in Equation (6.11). This low frequency temporal variation usually makes the flicker much more noticeable. In such cases as when the back-light frequency f_{bl} is a multiple of the frame rate f_{c} of the camcorder, the temporal radial frequency ω_k is a multiple of 2π and the flicker appears static in the recorded video.

While this model has been derived using simplifying assumptions, empirical observations have shown that the model holds in practice with regular video content. Earlier modeling efforts based on empirical measurements with reference test signals yielded the same model [105]. Interestingly, by using Equation (6.7) and the definition of the vertical radial frequency ω_r, it is possible to establish the following mathematical identity:

$$\Pi_{\text{flicker}} := T_{\text{ro}}\, f_{\text{bl}} = \frac{R\,\omega_r}{2\pi}, \tag{6.12}$$

which links such characteristics of the pirate devices as the read-out time T_{ro} of the

camcorder and the back-light frequency f_{bl} of the LCD screen, together with an intrinsic property of the flicker signal present in the video signal, the vertical radial frequency ω_r This *LCD-camcorder piracy identity* opens avenues to identify pirate devices from the analysis of the pirated videos that they produced [94].

6.3 Estimation of the Vertical Radial Frequency

On the right-hand side of the piracy identity given by Equation (6.12), the number of rows R is readily available from a frame of the pirated video but the vertical radial frequency ω_r of the flicker needs to be estimated. In practice, it is a challenging task since the flicker signal has a much lower energy than the pirated video content. As a result, the video content is likely to interfere with the estimation process. In what follows, we describe a methodology to estimate ω_r based on an analysis of the phase of some temporal discrete Fourier transform (DFT) coefficient at the aliased frequency f_{a} for different rows of a frame. We also present a refinement of this estimation technique to leverage on the harmonics of the flicker, and an alternate estimation strategy to cope with static flicker when the back-light frequency f_{bl} of the screen is a multiple of the frame rate f_{c}.

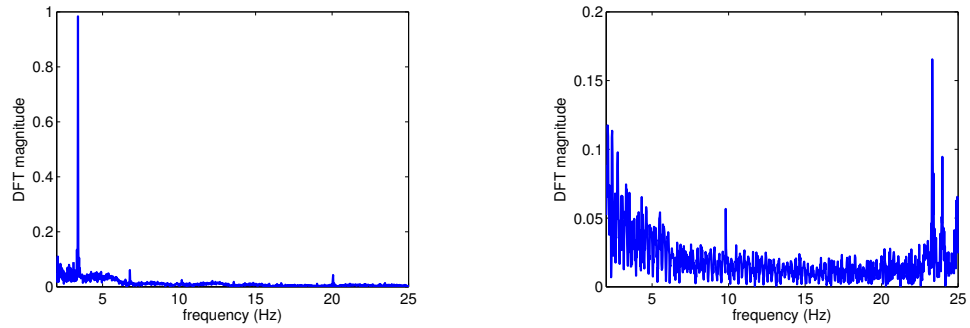
6.3.1 Flicker Phase Method

For each frame of a pirated video sample, the average luminance of each row is given by:

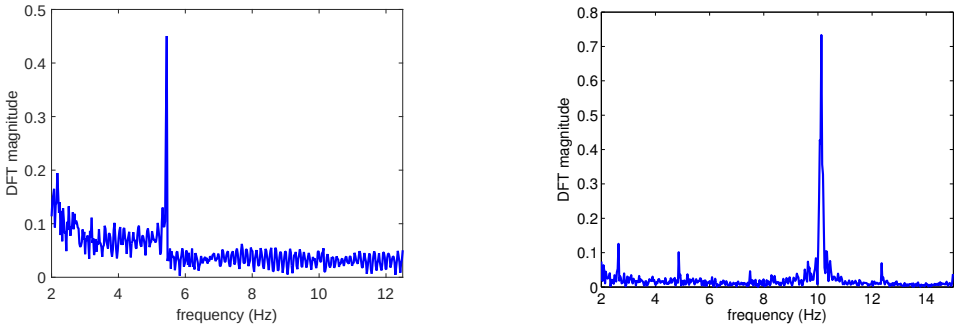
$$\begin{aligned} \mathbf{s}[r,k] &= \frac{1}{W}\sum_{c=1}^{W}\mathbf{p}[c,r,k] \\ &= \frac{1}{W}\sum_{c=1}^{W}\big(\mathbf{v}[c,r,k]+\mathbf{f}[c,r,k]\big), \end{aligned} \tag{6.13}$$

where c, r and k are the column, row and frame indices, respectively, and W is the number of pixels in a row of a frame of the pirated video sample $\mathbf{p}$, which is composed of a component $\mathbf{v}$ inherited from the original camcorded video and a flicker component $\mathbf{f}$. The virtue of this horizontal averaging operation is that it improves the signal-to-noise ratio (SNR) of the flicker versus the video content itself. The first term in Equation (6.13) aggregates a large number of pixels and thus reduces the interference due to the content. Due to the rolling shutter mechanism, for a given row, the sensors associated with each pixel are exposed to the same portion of the back-light signal and the second term in Equation (6.13) thus consolidates the flicker component.

Since the flicker signal has been modeled as a cosine function in Equation (6.11), one would expect the magnitude $\mathbf{R}[r,\omega]$ of the DFT of the row average $\mathbf{s}[r,k]$ along the time axis to feature a peak close to the aliased radial frequency ω_k. Figure 6.4 illustrates such temporal spectra $\mathbf{S}[r^*,\omega]$ for an arbitrarily chosen row r^* with alternate combinations of LCD screens and camcorders. In these figures, the x-axis has been mapped to Hz using the knowledge of the frame rate f_c, and the estimated



(a) $f_{\mathrm{a}} \approx$ -3.39Hz (b) $f_{\mathrm{a}} \approx$ -23.30Hz



(c) $f_{\mathrm{a}} \approx +5.43$Hz (d) $f_{\mathrm{a}} \approx +10.13$Hz

Figure 6.4: Magnitude of the DFT of row average $\mathbf{s}[r^*, k]$ at row r^* as a function of frame index k for several camcorded recordings of samples of the *Wall-E* video using various combination of LCD screens and camcorders.

aliased temporal frequency f_{a} is indicated for reference. All spectra have a more or less salient peak at the ground truth aliased frequency f_{a}. The presence of this anomalous peak could, for example, be used to detect camcorder piracy [58].

According to Equation (6.11), the phase $\Phi_{\omega_k}[r]$ of the DFT coefficients $\mathbf{S}[r, \omega_k]$ is given by $r\omega_r + \theta$, where θ is some constant phase shared across rows. In other words, the *vertical phase* of the flicker is expected to evolve linearly along the rows, with a slope equal to the vertical radial frequency ω_r. Empirical observations reported in Figure 6.5 corroborate this overall linear trend, if we ignore the visualization effect due to the modulo-2π operator. The undesired modulo-2π wrapping in

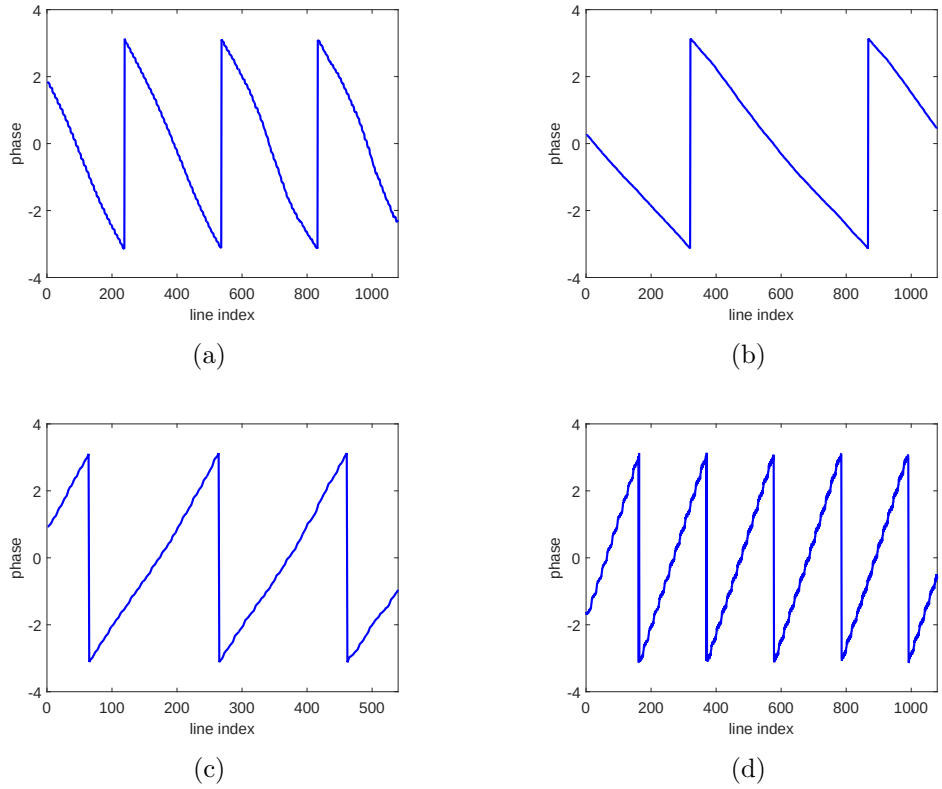


Figure 6.5: Evolution of the flicker phase, computed using the DFT coefficients $\mathbf{S}[r, |\omega_k|]$, with the row index in the video frame. The measurements have been extracted from camcorded recordings of the *Wall-E* video samples using different camcorder–screen pairs.

the phase can be compensated for as follows:

$$\Psi_{\omega_k}[r] = \begin{cases} \Phi_{\omega_k}[r], & \text{if } r = 0, \\ \Psi_{\omega_k}[r-1] + d_r, & \text{if } r > 0, \end{cases} \tag{6.14}$$

where

$$d_r = \Big(\big(\Phi_{\omega_k}[r] - \Phi_{\omega_k}[r-1] + \pi \big) \mod 2\pi \Big) - \pi. \tag{6.15}$$

Estimating the vertical radial frequency ω_r can then be done by applying linear regression to the *unwrapped* flicker phase $\Psi_{\omega_k}[r]$ and recording the slope [1]. In summary,

[1] One should keep in mind that ω_k can be positive or negative as it is dependent on f_a as defined in Equation (6.10). Due to the symmetry of the DFT practice, it is common practice to focus on the positive side of the frequency axis. As a result, the slope of the phase $\Phi_{|\omega_k|}[r]$ can be reversed as observed in Figure 6.5. In that case, one could recover the true radial frequency simply by taking the absolute value of the estimated slope.

one could estimate ω_k by recording the radial frequency that maximizes $\mathbf{S}[r^*, \omega]$, for an arbitrarily chosen row r^*, and then recover the vertical radial frequency ω_r by estimating the slope of the vertical phase $\Phi_{\omega_k}[r]$ [105].

6.3.2 Exploiting the Harmonics of the Flicker

In practice, the strongest peak of the spectrum $\mathbf{S}[r^*, \omega]$ may not always correspond to ω_k. If some prior knowledge about the back-light frequency f_{bl} is available, one can compute candidate aliased frequency values using Equation (6.10) and restrict the search for a peak in $\mathbf{S}[r^*, \omega]$ to the corresponding radial frequency ranges [94]. However, such a priori information may not be available depending on the application scenario and alternate solutions are therefore desirable.

Many of the peaks present in the spectrum $\mathbf{S}[r^*, \omega]$ actually correspond to aliased versions of the harmonics of the back-light frequency f_{bl}. When a peak is detected at some frequency $\omega^\dagger$, a tell-tale cue that it is related to the back-light is that its corresponding unwrapped phase $\Psi_{\omega^\dagger}[r]$ defined in Equation (6.14) should be linear, with a slope that is a multiple of ω_r in line with its harmonic index. Such linear phase characteristic could be exploited to pinpoint frequencies in the spectrum that are related to the back-light and thus collect several observations of ω_r or its multiples, and thereby opening avenues for consolidating the estimate of ω_r.

To begin with, the temporal DFT $\mathbf{S}[r, \omega]$ is computed for all the rows of the considered video sequence. For each frequency ω, the phase $\Phi_\omega[r]$ is unwrapped

using Equation (6.14) and its linearity is evaluated based on the root mean square error ξ of its linear regression. A frequency ω_i is conserved for further analysis if (i) its unwrapped phase does not deviate from linearity more than a specified threshold (i.e., $\xi_i < \tau_\xi$), and (ii) its linear slope exceeds another threshold (i.e., $\alpha_i > \tau_\alpha$). The latter test is intended to discard frequencies whose slopes are too small to correspond to read-out times and back-light frequencies typically used in the industry. Our empirical observations led us to use a fixed linearity threshold $\tau_\xi = 0.04$ and to adjust the other threshold according to $\tau_\alpha = 70\pi/Rf_c$. Keeping in mind the piracy identity given by Equation (6.12), this setting gives a strong lower bound on the product Π_{flicker} that is far from what we have observed for the devices that we have tested. When no frequency satisfies these two constraints, we switch to a fall-back strategy to be described in Section 6.3.3.

At this stage, the different slopes $\mathcal{A} = \{\alpha_i\}_{1 \leq i \leq A}$ that we have retained are expected to be multiples of the vertical radial frequency ω_r that we are trying to estimate. However, these observations may be noisy and contaminated by parasite samples. To identify the best common denominator for all α_i's, one could look for the frequency α^* that maximizes the number of inlying harmonics, i.e.,

$$\alpha^* = \arg \max_{\alpha \in [0, +\infty[} \sum_{i \in \mathcal{I}_\alpha} \kappa_i, \tag{6.16}$$

where $\kappa_i > 0$ are some weights and $\mathcal{I}_\alpha$ is the set of inlying harmonics for a given slope α defined as follows:

$$\mathcal{I}_\alpha = \left\{ i, \left| \frac{\alpha_i}{\alpha} - \left\lfloor \frac{\alpha_i}{\alpha} \right\rceil \right| < \delta_{\text{int}} \right\}, \tag{6.17}$$

where $\lfloor . \rceil$ is the nearest integer rounding operator and δ_{int} some threshold that we

arbitrarily set to 0.1 in our experiments. In other words, those observed α_i whose ratios with a hypothesized fundamental frequency α deviate by more than δ_{int} from an integer value are considered to be outliers and are discarded. When $\kappa_i = 1$, Equation (6.16) reduces to the count of inlying harmonics. Alternately, one could use the weights κ_i to force more reliable observations into the inlier set. For example, the observations corresponding to frequencies that have a more linear phase may be privileged, i.e.,

$$\kappa_i = 1 - \frac{\xi_i}{\tau_\xi}. \tag{6.18}$$

The objective function optimized in Equation (6.16) is highly non-linear and may not have a unique solution. This is the reason why, in practice, we simply enumerate the slopes $\alpha \in \mathcal{A}$ and compute the weighted average of the ones that maximize the objective function.

This best denominator α^* should be roughly equal to the desired vertical radial frequency ω_r. However, this approximate value may be rather rough, especially when the optimization process of Equation (6.16) locks on a single observation $\alpha_i \in \mathcal{A}$. Now that we have found a fundamental frequency α^*, we can define a harmonic index $h_i = \lfloor \alpha_i / \alpha^* \rceil$ for each observation and compute α_i / h_i as multiples estimates of the vertical radial frequency. A refined estimate $\hat{\omega}_r$ can then be obtained using a least squares formulation, taking into account the weights κ_i's:

$$\hat{\omega}_r = \arg \min_{\alpha \in [0, +\infty[} \sum_{i \in \mathcal{I}_{\alpha^*}} \kappa_i \left(\frac{\alpha_i}{h_i} - \alpha \right)^2, \tag{6.19}$$

whose solution is given by the usual weighted average:

$$\hat{\omega}_r = \frac{\sum_{i \in \mathcal{I}_{\alpha^*}} \kappa_i \frac{\alpha_i}{h_i}}{\sum_{i \in \mathcal{I}_{\alpha^*}} \kappa_i}. \tag{6.20}$$

Table 6.2: Benefit of using harmonics with a toy example (strongest peak in italic)

$\omega/\mathbf{2}\pi$ **(Hz)**	$\alpha_\mathbf{i}$ **(rad/s)**	$\kappa_\mathbf{i}$	$\mathbf{h_i}$	$\alpha_\mathbf{i}/\mathbf{h_i}$ **(rad/s)**
-9.957	0.0224	0.20	2	0.0112
-9.961	0.0225	0.78	2	0.0113
-9.984	*0.0222*	*0.84*	*2*	*0.0111*
-9.994	0.0223	0.78	2	0.0111
-10.010	0.0219	0.49	2	0.0110
10.027	0.0334	0.08	3	0.0111
-19.924	0.0450	0.44	4	0.0112
-19.958	0.0442	0.86	4	0.0111
-19.968	0.0445	0.85	4	0.0111
-19.978	0.0448	0.64	4	0.0112
19.984	0.0114	0.25	1	0.0114
20.005	0.0116	0.20	1	0.0116
20.015	0.0117	0.52	1	0.0117
20.038	0.0114	0.43	1	0.0114

Table 6.2 illustrates the benefit of considering harmonics to estimate ω_r by showing sample results from applying the proposed approach on the case of a pirated video made using an LCD screen with a back-light frequency of 120Hz and a camcorder that captures 1080p video at 50 fps. As a result, the fundamental frequency and its fourth harmonic alias to $\pm$20Hz, and the second and third harmonics alias to $\pm$10Hz. Moreover, the ground truth for the vertical radial frequency is $\omega_r = 0.0112$. When using the method described in Section 6.3.1, the algorithm locks on the largest peak of the spectrum at -9.984Hz which happens to be related to the second harmonic yielding an erroneous estimate of $\hat{\omega}_r = 0.0222$. Even if we consider a priori information as in [94], another peak can be found in the spectrum at -19.968Hz, which is actually associated with the fourth harmonic and so, the fundamental frequency cannot be recovered. In contrast, the exhaustive search of frequency bins with linear phase and the aggregation of these harmonic estimates first yields a coarse estimate $\alpha^* = 0.0113$, which is later refined to $\hat{\omega}_r = 0.0112$.

6.3.3 Content Cancellation Method

The estimation techniques described in the previous sections inherently assume that it is possible to lock on the frequency bin in the spectrum that corresponds to the back-light. However, as discussed in Section 6.2.4, when the frame rate f_c of the camcorder is a multiple of the back-light frequency f_{bl}, the observed temporal flicker is close to zero, which precludes accurate estimation due to strong interference from content in the low-frequency band. A fall-back strategy is therefore needed to cope with such situations [94].

The row luminance signatures $\mathbf{s}[r,k]$'s are mostly dominated by the video content, thereby making the subtle changes underpinning the luminance flicker difficult to analyze. Nevertheless, the content interference is expected to vary slowly along the rows. It can therefore be canceled to some extent by applying a high-pass filter over time or by removing the trend of the signal using some fitting tool, i.e.,

$$\acute{\mathbf{s}}[r,k] = \mathrm{h}\left(\mathbf{s}[r,k]\right), \tag{6.21}$$

where h(.) is a generic signal processing primitive used to remove the low frequency components of a signal. Our empirical observations indicate that the effect of this cleanup process is higher for video frames having more predictable row luminance signatures (e.g., frames with more uniform content).

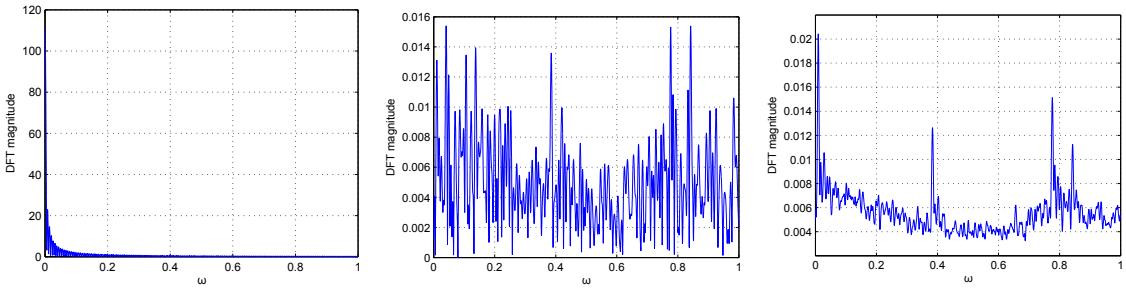
Individually cleaned row luminance signatures $\acute{\mathbf{s}}[r,k]$ still present significant content energy, and estimating the vertical radial frequency ω_r based on the spectrum analysis of a single signature is likely to be unsuccessful. To improve the

SNR, a common technique in multimedia security consists of aggregating several observations to reduce the interference introduced by uncorrelated noise components [106, 107]. This aggregation can be performed directly in the DFT domain:

$$|\bar{\mathbf{S}}[\omega]| = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} |\acute{\mathbf{S}}[\omega, k_u]|, \tag{6.22}$$

where $|\acute{\mathbf{S}}[\omega, k]|$ is the magnitude of the DFT of the k-th cleaned row luminance signature $\acute{\mathbf{s}}[r, k]$, $|\bar{\mathbf{S}}[\omega]|$ the magnitude of the vertical flicker aggregate spectrum, and the set $\mathcal{U}$ indicates which frames of the video have been considered for aggregation. In practice, we used the $|\mathcal{U}| = 40$ most uniform video frames (i.e., the frames of the video with the lowest variance). Aggregating directly in the DFT domain helps cope with the slight phase offsets of the individual row luminance signatures $\acute{\mathbf{s}}[r, k_u]$.

Eventually, the vertical radial frequency ω_r is given by the frequency whose magnitude is maximal in the spectrum $|\bar{\mathbf{S}}[\omega]|$. To avoid false estimations, we discard the frequencies that exceed a threshold τ_{ω_r} when they correspond to back-light frequencies that are never used in practice. Our empirical observations showed that using $\tau_{\omega_r} = 1$ radians/row provides good performance in general. For reference, Figure 6.6 depicts the benefit of the cleaning and aggregation processes in a particularly difficult case.



(a) $|\mathbf{S}[\omega, k]|$ (b) $|\acute{\mathbf{S}}[\omega, k]|$ (c) $|\bar{\mathbf{S}}[\omega]|$

Figure 6.6: Spectrum of the signal of interest at various stages of the content cancellation process. The spectrum $|\mathbf{S}[\omega, k]|$ of the row luminance signature (a) is dominated by the visual content and the flicker signal is not visible. In contrast, the cleaning process (b) reveals the peak corresponding to the flicker at 0.009 rad/row although it lies hidden amongst other noise components. After aggregation (c), the flicker frequency peak clearly appears thanks to SNR improvement.

6.4 Pirate Devices Identification

As mentioned earlier, the LCD-camcorder piracy identity given by Equation (6.12) opens avenues to linking a pirated video sample to the devices that have been used to produce it. For example, police investigators, who receive traitor-tracing evidence recovered from pirated movies, may raid the home of a suspect pirate and seize a collection of devices. In that case, flicker forensics could corroborate whether or not the flicker signal observed in the pirated movies could be produced using these devices. Additionally, when forensic watermarking detection fails, flicker forensics could provide a fall-back mechanism to link together pirated samples that originate from the same group of pirates.

Identifying the pirate devices among a collection of suspect LCD screens and camcorders can be done by isolating the pair of devices that yields the closest-to-zero difference between the two sides of the LCD-camcorder piracy identity of

Equation (6.12). While Section 6.3 provides means to evaluate the right-hand side of the identity, Section 6.4.1 details how to recover the read-out time T_{ro} and the back-light frequency f_{bl} from suspected devices. This is indeed required to evaluate the product on the left-hand side of the identity, and thus perform flicker-based pirate devices identification. Experimental results are reported in Section 6.4.2 and clearly demonstrate the ability of our proposed approach to perform the forensic task concerned here.

6.4.1 Extracting Internal Parameter Values from Devices

In this study, we have used seven LCD screens and four camcorders that are listed in Tables 6.3 and 6.4 and that allowed us to produce camcorded video samples of twenty eight possible screen–camcorder combinations. To compute the product Π_{flicker} for these devices, it is necessary to have access to their inner characteristics, namely, the back-light frequency f_{bl} of the LCD screens and the read-out time T_{ro} of the camcorders. Such low-level characteristics are usually not indicated in the datasheets or manuals of consumer electronics products. Nevertheless, when the suspect devices are available, it is possible to conduct some semi-non-intrusive forensic analysis where devices can be fed with specific stimuli to facilitate measurements [108, 109]. The only constraint is to avoid tampering with the integrity of the device (i.e., breaking it apart to examine its individual components), which is either impossible or undesirable to safeguard the chain of custody.

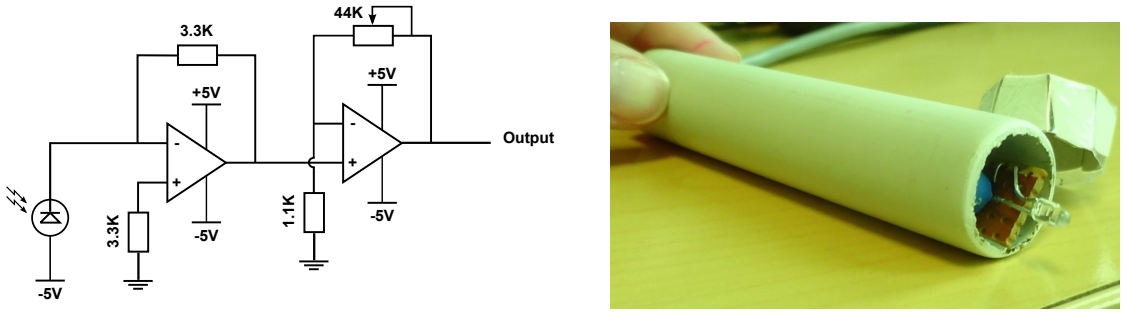


Figure 6.7: Custom-made light sensing probe. The photo-diode converts light into electric current, which is amplified by a first amplifier on the left-hand side. Namely, 0.1 mW/cm^2 yields a current of 0.8 μA and 2.64 mV. The adjustable gain amplifier on the right-hand side is then useful for accommodating to various light intensities of different screens. The gain can vary between 1 and 44.

6.4.1.1 LCD Screen Back-light Frequency

To recover the back-light frequency f_{bl} of an LCD screen, we custom-made a sensing circuit that converts captured light into an electrical signal and the reversed-current of a photo-diode is amplified. The whole circuit is embedded within a pen-like casing that has a pin hole to let incoming light in as illustrated in Figure 6.7. The output of the sensing circuit can be connected to a PC or an oscilloscope for live analysis, or to a recording device for off-line analysis. Placing this apparatus on the surface of a screen displaying a static uniformly gray frame provides direct access to the back-light signal without interference from other light sources or from the temporal dynamic of a motion picture. The recorded signal is typically a periodic signal whose fundamental frequency is equal to the back-light frequency f_{bl} of the LCD screen. The ground truth back-light frequency of the screen is then retrieved through spectrum analysis. Table 6.3 reports the measurements obtained with the seven LCD screens used in our experiments. The reverse-engineered back-light fre-

quencies are within a 120-250Hz frequency band, which is in line with the known practices of the industry.

Table 6.3: LCD screens used in our experiments

ID	**Brand**	**Model**	$\mathbf{f_{bl}}$ **(Hz)**
1	Dell	2209WA	240.06
2	Dell	U2410	180.43
3	Samsung	LE37B652T4WXXC	159.98
4	Samsung	UE32C6000RWXZF	120.00
5	Sony	KDL-32P3000	146.61
6	Sony	KDL-37P3000	226.70
7	Sony	KDL-32W5710	172.80

6.4.1.2 Camcorder Read-out Time

Accessing the read-out time T_{ro} of a camcorder is less direct than measuring the back-light frequency using a probe. The trick is to use the suspect camcorder to record a reference LCD screen displaying uniform grayscale content, as depicted in Figure 6.1, obtaining a short video sequence (e.g., 30 seconds long) where the flicker is apparent. The benefit of this neutral stimulus is that it prevents visual content interference in the estimation methods described in Section 6.3. Using the ground truth back-light frequency of the reference screen, it is thus possible to first obtain an accurate estimate of the flicker vertical radial frequency ω_r and then compute the read-out time using the piracy identity of Equation (6.12). For improved accuracy, one can combine measurements obtained from several reference screens. Table 6.4 lists the ground truth read-out time values obtained for the four camcorders used in our experiments[2].

[2]All camcorders are progressive, except the Sony camera that is interlaced. For convenience, we kept a single field for this camera, thereby resulting in a vertical resolution of 540 rows although

Table 6.4: Camcorders used in our experiments

Brand	**Model**	**R**	$\mathbf{T_c = 1/f_c}$ **(ms)**	$\mathbf{T_{ro}}$ **(ms)**
JVC	GC-PX100BE	1080	20	13.5
Panasonic	HDC-SDT750	1080	20	16
Sony	HDR-CX200E	540	40	15
Toshiba	PA5081E-1C0K	1080	33.33	32.65

6.4.2 Experimental Results

To validate whether or not the proposed flicker-based forensic protocol can accurately identify pirate devices from a video, we first recorded 1 minute long video sequences encompassing all possible screen–camcorder combinations of the devices listed in Tables 6.3 and 6.4. No specific care (e.g., distance to screen, zoom, focus, ambient lighting, etc) has been taken to calibrate these captures in order to reflect real-world piracy conditions. Samples of screenshots from the camcorded videos can be seen in Figure 6.8. The flicker present in the resulting 28 pirated videos is more or less apparent depending on the device combination. To evaluate the accuracy of the different techniques proposed to estimate ω_r described in Section 6.3, we consider the relative estimation error defined as follows:

$$\epsilon = 100 \times \left| 1 - \frac{R\,\hat{\omega}_r}{2\pi\, T_{\text{ro}}\, f_{\text{bl}}} \right|. \tag{6.23}$$

In order to better appreciate the added value of considering harmonics during the estimation process, the relative estimation error is reported in Table 6.5 both with and without harmonics, with the latter case being reported in [94]. Entries in italic highlight screen–camcorder combinations that eventually yield a classification

the camcorder has the ability to capture 1080 rows.

(a) `Screen 2` and `JVC` (b) `Screen 3` and `Panasonic` (c) `Screen 7` and `Sony` (d) `Screen 1` and `Toshiba`

Figure 6.8: Screenshots of camcorded video sequences using various screen–camcorder pairs. Besides very different color changes, the flicker signal is more or less apparent depending on the pair of pirate devices.

Table 6.5: Relative estimation error with (without) harmonics (Italic entries highlight identification errors)

	`JVC`	`Panasonic`	`Sony`	`Toshiba`
`Screen 1`	0.51 (1.32)	0.94 (0.78)	0.27 (0.17)	0.19 (0.50)
`Screen 2`	1.10 (*1.49*)	0.55 (1.09)	0.00 (0.36)	0.18 (0.41)
`Screen 3`	0.13 (0.38)	0.37 (1.03)	0.01 (*1.93*)	0.17 (0.02)
`Screen 4`	0.02 (0.72)	0.13 (*98.87*)	0.05 (*6.09*)	0.02 (0.74)
`Screen 5`	0.14 (0.02)	*1.05* (*6.74*)	0.00 (0.01)	0.03 (0.65)
`Screen 6`	0.10 (0.39)	*1.31* (0.43)	0.64 (0.17)	0.02 (0.14)
`Screen 7`	0.24 (*0.39*)	0.24 (0.24)	0.28 (0.12)	0.58 (0.56)

error. A first observation from these results is that incorporating harmonics notably decreases the overall relative estimation error, although it does not *always* improve the estimation, especially in cases where the quality of the harmonics is low. The average relative estimation error is 0.33% with harmonics, compared to 4.49% without harmonics. The improvement is particularly evident for the `Panasonic-Screen4` combination (which is the corner case discussed at the end of Section 6.3.2). In that case, only considering the peaks of the spectrum locks on the second harmonic of the flicker and thus eventually yields an identification error. This improved estimation accuracy naturally translates in fewer identification errors as we observe an error rate of 6/28 without harmonics vs. 2/28 with harmonics.

The likelihood of having an identification error is not strictly related to the

Table 6.6: Rank of the correct screen–camcorder pair using flicker-based pirate device identification with (without) harmonics

	JVC	Panasonic	Sony	Toshiba
Screen 1	1 (1)	1 (1)	1 (1)	1 (1)
Screen 2	1 (2)	1 (1)	1 (1)	1 (1)
Screen 3	1 (1)	1 (1)	1 (3)	1 (1)
Screen 4	1 (1)	1 (23)	1 (2)	1 (1)
Screen 5	1 (1)	2 (4)	1 (1)	1 (1)
Screen 6	1 (1)	2 (1)	1 (1)	1 (1)
Screen 7	1 (2)	1 (1)	1 (1)	1 (1)

relative estimation error ϵ. It is rather highly dependent on the distribution of the values $\Pi_{\text{flicker}} = T_{\text{ro}}\, f_{\text{bl}}$ for the considered suspect devices. For instance, the pairs `Panasonic-Screen5` and `JVC-Screen7` respectively have Π_{flicker} values equal to 2346 and 2333, which can make them highly interchangeable during the identification process. The different screen–camcorder combinations can be sorted according to how much they deviate from the piracy identity given in Equation (6.12). Table 6.6 indicates what is the rank of the ground truth pirate screen–camcorder pair for the different videos in our dataset. Most of the times when there is an identification mistake, the correct pair of suspect devices is high in the sorted candidate list. In summary, although flicker-based forensics may not achieve yet the high accuracy to hold in court, it is capable of providing additional intelligence to direct the investigations or infer relationships in a large database of pirated video samples.

6.5 Extension to Flicker Profile Recovery

In Section 6.4, we showed how to link suspect devices to pirated video samples based on their inner parameters that yield a very specific flicker characteristic in the

pirated video. It is a definite step forward compared to simply detecting camcorder piracy, and it may be possible to pursue the forensic analysis even further. The mathematical model derived in Section 6.2 relies on the simplifying assumption that the back-light signal is a perfect periodical rectangular signal. In real life, the back-light deviates from this ideal signal and the resulting profile of the flicker signal (the coefficients A_i in Equation (6.5)) is affected accordingly. Nevertheless, the frequency analysis detailed throughout Section 6.3 remains valid.

In some application scenarios, it may be useful to estimate this so-called *profile* of the flicker. For example, it can be used to cancel the flicker component either to improve the quality of a recorded video [110] or to clean a pirated sample prior to applying a watermark detection algorithm [105]. Conversely, a finer estimation of the flicker profile could be exploited to better mimic the distortion occurring along the camcorder path and thus provide a convenient tool for benchmarking [111,112]. In this section, we explore an estimation technique to recover the flicker profile from a pirated sample, and examine its applicability in a sample application scenario: inferring the back-light technology that was implemented by the LCD screen used for piracy.

6.5.1 Recovering the Flicker Profile

Our baseline approach to estimate the flicker profile is to collect several noisy estimates of the flicker and to pool them together to improve the SNR. In contrast with the content cancellation method of Section 6.3.3, this pooling process cannot

be performed in the frequency domain. Indeed, as we intend to recover the profile of the signal, it is necessary to account for the mismatching phases of the different estimates. As a result, the overall estimation methodology reduces to a three-step process:

1. collect O noisy estimates $\mathbf{p}_o[r]$ of the flicker profile, e.g., from different frames of the camcorded video;

2. compute a collection of lags $\mathcal{D} = \{\hat{\delta}_o\}$ to re-align all $\mathbf{p}_o$'s, one with respect to the other;

3. pool the realigned estimates $\mathbf{p}_o[r + \hat{\delta}_o]$ to obtain a refined estimate $\hat{\mathbf{p}}$ of the flicker profile.

A straightforward way to collect estimates $\mathbf{p}_o$ of the flicker profile is to cancel the contribution from the original video content in the row luminance signature of selected frames as in Equation (6.21). This process is most efficient when applied to frames with uniform content. These individual estimates are then post-processed to have zero mean and unit variance. While a tempting strategy to obtain such estimates would be to only keep flicker harmonics components, our early attempts suggest that this approach does not manage to accurately capture the profile of the flicker. This may be worth further investigation in follow-up studies.

The individual estimates $\mathbf{p}_o$ are misaligned in general because the frame rate of the camcorder is rarely a multiple of the screen back-light frequency. To re-align two estimates $\mathbf{p}_{o_1}$ and $\mathbf{p}_{o_2}$, one strategy is to identify the lag δ that maximizes some

paired correlation score. That is,

$$\mathrm{corr}(\mathbf{p}_{o_1}, 0; \mathbf{p}_{o_2}, \delta) = \mathbf{p}_{o_1}[r] \cdot \mathbf{p}_{o_2}[r+\delta], \tag{6.24}$$

where $\cdot$ is the linear correlation and $\mathbf{p}_{o_1}$ serves as the reference signal. When several estimates $\{\mathbf{p}_o\}$ are available, the re-alignment process then amounts to jointly optimizing the alignment between all pairs of estimates and can be written as:

$$\mathcal{D} = \arg\max_{\{\delta_o\}} \mathrm{Obj}\Big(\big\{\mathrm{corr}(\mathbf{p}_{o_1}, \delta_{o_1}; \mathbf{p}_{o_2}, \delta_{o_2})\big\}_{o_1<o_2}\Big), \tag{6.25}$$

where one of the lags, e.g., δ_1, is arbitrarily set to zero to serve as a reference, and Obj(.) is an objective function to optimize. For instance, one may want to maximize the average of all paired correlation scores. Alternately, one may want to maximize the lowest paired correlation scores.

Eventually, all re-aligned noisy estimates are pooled together to improve the SNR of the flicker profile. The multiplicative nature of the flicker reported in [105] calls for utilizing similar optimized pooling functions originally derived for the Photo Response Non Uniformity (PRNU) pattern estimation [113]. However, our empirical observations showed only marginal improvement of the convergence speed compared to conventional averaging. So, we adopted the latter for the sake of simplicity:

$$\hat{\mathbf{p}}[r] = \frac{1}{O_r} \sum_{1\le o\le O} \mathbf{p}_o[r+\hat{\delta}_o], \quad 1 \le r \le R, \tag{6.26}$$

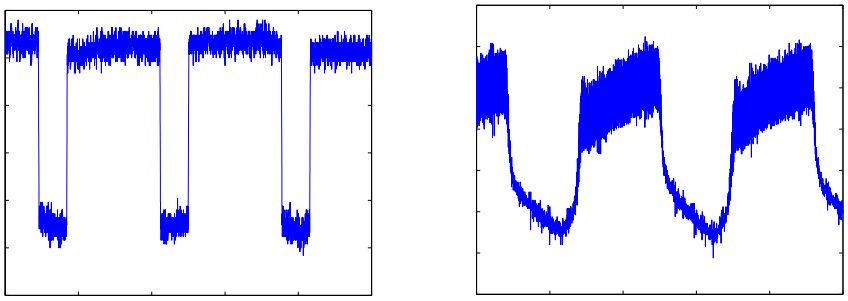
where O_r is a normalization constant that counts how many individual estimates, for each row, contributed to the final row luminance signature $\hat{\mathbf{p}}$.

In practice, the optimization problem defined by Equation (6.25) is difficult to solve. A suboptimal strategy consists of iteratively incorporating individual esti-

mates $\mathbf{p}_o$ one at a time to the average, where the lag $\hat{\delta}_o$ is given by the offset that maximizes the correlation with the current estimation of $\hat{\mathbf{p}}$. This straightforward approach is inherently dependent on the scanning order, which can create a bias. A better approach is to adopt a *divide and conquer* paradigm: the individual estimates $\mathbf{p}_o$ are virtually placed on the leaves of a binary trees and are pooled according to a bottom-up traversal of the tree. On the way up, two estimates of the flicker profile are re-aligned using Equation (6.24) and averaged at each node of the tree.

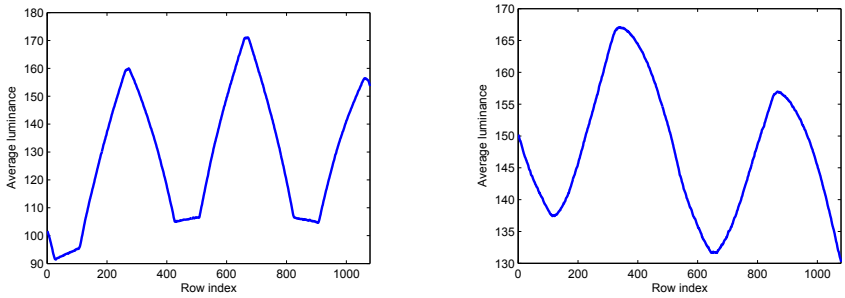
6.5.2 Revealing LCD Back-light Technology

Back-light is generated from two main types of sources in practice: Light-Emitting Diodes (LED) and Cold Cathode Fluorescent Lamps (CCFL). While LED screens are becoming dominant, there is still a large number of legacy CCFL screens that could be used for piracy. Figure 6.9 shows samples of raw back-light signals of LCD screens of different back-light technologies made using the light sensing probe described in Figure 6.7. Observing such signals, we empirically found CCFL and LED signals to be notably different. CCFL features a transitory regime between the on/off states of the back-light in contrast with LED that has sharp transitions.



(a) LED (b) CCFL

Figure 6.9: Temporal back-light signal captured at an arbitrary point of an LCD screen displaying a uniformly gray image using a probe equipped with a photodetector connected to an oscilloscope.

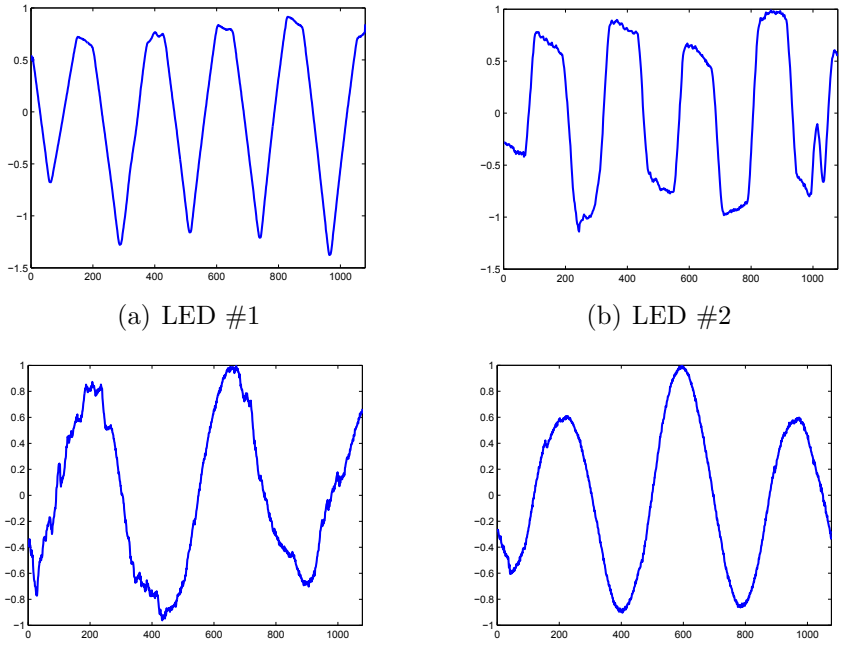


(a) LED (b) CCFL

Figure 6.10: Illustration of the row luminance signature in a toy example where the LCD screen is fed with a uniform grey video.

According to Equation (6.5), the flicker profile is related to the convolution between the back-light signal of the LCD screen and the exposure mechanism of the camcorder. The intrinsic difference between LED and CCFL back-lights should therefore be reflected in the flicker signal present in camcorded videos. As a quick verification, one can see differences in the extracted row luminance signatures, shown in Figure 6.10 extracted from camcorded recordings of a uniformly grey image shown on LED and CCFL screens.

The remaining question is whether or not the flicker profile estimated from a camcorded video exhibits characteristic properties that could serve as a forensic



(a) LED #1 (b) LED #2

(c) CCFL #1 (d) CCFL #2

Figure 6.11: Illustration of the flicker profiles extracted from *Wall-E* camcorded videos using various LED and CCFL screens. For each of these figures, the x-axis shows the index of the row in a video frame, and the y-axis shows the normalized magnitude of the extracted flicker profile.

cue to identify the back-light technology of the pirate screen. To investigate this problem, we recorded 1-minute long *Wall-E* video samples using three camcorders and six LCD screens, half of which has an LED back-lights and the other half has CCFL back-lights. We then extracted the flicker shape present using the approach described in Section 6.5.2 with $O = 30$ frames.

Figure 6.11 depicts the flicker profile $\hat{\mathbf{p}}$ extracted from real camcorded videos using two LED and two CCFL screens. The extracted flicker profile extracted with CCFL screens is much smoother than with LED ones. Numerous features have been proposed in the literature to characterize the smoothness of a signal and, after several tests and trials, we opted for polynomial fitting goodness of fit. Since LED

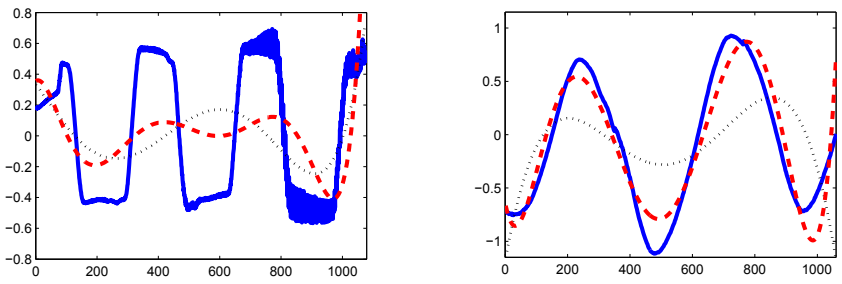
flicker profiles feature gradient discontinuities, they should be more difficult to fit using a polynomial function than CCFL flicker profiles.

To capture this difference, we considered the following feature:

$$f = C_{z+1} - C_{z-1}, \tag{6.27}$$

where C_n denotes the correlation coefficient between the flicker profile $\hat{p}$ and its corresponding polynomial fit of degree n, and z is the number of zero crossings of the flicker profile. In principle, as a correlation coefficient lies within $[-1, 1]$, the feature f can span the interval $[-2, 2]$. However, given that we are carrying out the polynomial fitting at an order close to the number of zero crossings and that the higher order fitting generally yields a higher correlation value than the lower order one, the feature f is expected to be within $[0, 1]$. Both polynomials of degrees $z-1$ and $z+1$ should be rather poor fits of LED flicker profiles, thereby producing a low feature value. On the other hand, smooth CCFL profiles should be nearly perfectly fitted with a polynomial of degree $z+1$, whereas the other polynomial fit should deviate from the flicker profile since the polynomial degree is lower than the number of zero crossings[3]. This drastic fitting improvement is expected to yield much larger feature values of f. Figure 6.12 shows examples of these polynomial fittings for row luminance signatures extracted from camcorded LED and CCFL screens, showing visibly more fitting improvement in the CCFL case.

[3]Due to the normalization process, the number of zero crossings relates to the number of gradient sign changes of the (low-pass) flicker profile.



(a) LED (b) CCFL

Figure 6.12: Examples of polynomial fitting for row luminance signatures extracted from camcorded copies of *Wall-E*. Legend: the solid blue line is the row luminance signature; the dotted black line is the polynomial fit of degree $z - 1$; the dashed red line is the polynomial fit of degree $z + 1$.

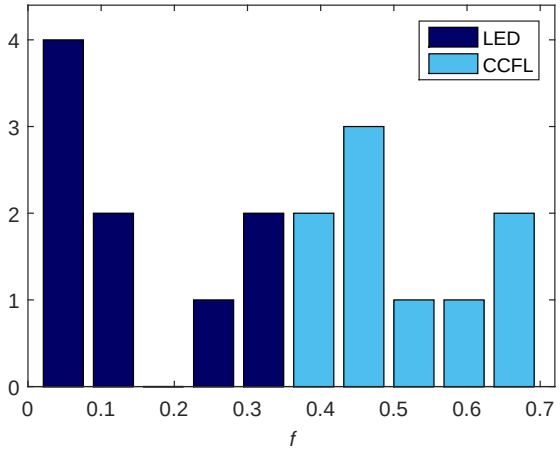


Figure 6.13: Histogram of the feature values f computed according to Equation (6.27) for several pirated copies using various combinations of LCD screens and camcorders.

Figure 6.13 shows a histogram of the feature values f computed using the eighteen pirated video samples in our dataset. As predicted, pirated copies involving CCFL screens produce higher feature values than the ones involving LED screens. For the devices and data in our study, the two sets of values were perfectly separable with a threshold of around 0.35. These proof-of-concept results should be confirmed with a larger diversity of pirated samples. Nevertheless, they already suggest that the estimation of the flicker profile is accurate enough to capture characteristics

inherited from the back-light signal, which can be exploited to reveal the back-light technology implemented by the pirate screen in case of camcorder piracy.

6.6 Chapter Summary

Camcorder piracy is an increasing threat nowadays as high-valued premium video content is delivered to homes earlier after theatrical releases. Previous works have focused on detecting that camcorder piracy occurred by several techniques that can include feeding relevant features to a supervised machine learning engine [55–58, 97–100]. In this chapter, we have gone a step further and performed an in-depth analysis of one of these discriminative features, namely, the luminance flicker that originates from the interplay between the back-light of the LCD screen and the shutter of the camcorder. As such, flicker forensics can serve as a complementary analysis once the pirated sample has been detected to be an LCD camcord (using, for example, one of the previously published learning based techniques.

After deriving a theoretical model for the flicker signal, we have established a mathematical identity between the flicker frequency and some inner parameters of the pirate devices, namely, the back-light frequency of the screen and the read-out time of the camcorder. We then showed that components of the flicker signal could be pinpointed due to their linear phase in some frequency representation. This characteristic can be used to isolate the flicker component and then recover its fundamental frequency. Our experimental results demonstrated that it enables piracy attribution by identifying which screen–camcorder pair is the most likely

to have been used to produce the analyzed pirated video with high accuracy. We also conducted a preliminary study to demonstrate that the flicker profile could be accurately estimated. We then exemplified that it could be used to tell-tale which back-light technology is implemented by the pirate screen.

In future work, a first task will be to better understand the applicability of flicker forensics. It will involve large scale validation experiments with a wide diversity of devices and also benchmarking evaluation against subsequent video processing. Another line of research will be to look for other statistical footprints in pirated movies that reveal camcorder piracy. Flicker forensics indeed involves only a handful of parameters and is thus a rather moderate piece of intelligence on its own. Complementary cues would serve as corroborating intelligence that strengthen the forensics process. Further, this refined understanding of the luminance flicker may be exploited to improve detection of watermarking systems, possibly thanks to enhanced flicker removal techniques. It would be interesting to explore other possibilities in forensics that can utilize the framework investigated in this chapter.

Chapter 7

Conclusions and Future Perspectives

In this dissertation, we have explored two intrinsic near-invisible signatures that can be unintentionally captured in a media recording due to influences from the environment in which it was made and the recording device that was used to make it. We considered two main signatures, the Electric Network Frequency (ENF) signal and the flicker signal, and examined their use to address problems in multimedia forensics.

We have discussed that the ENF signal is an imprint of the activity of the power grid in which a recording is made. It is a signature that can be used to address questions about a recording's integrity and origins. The validity of the final results of most ENF-based applications will highly depend on the accuracy in which the ENF signal is extracted. For that purpose, we developed our proposed spectrum combining approach that exploits the presence of the ENF traces at different harmonics to achieve more robust and accurate ENF signal estimations.

After observing that ENF signals from different grids show different statistical characteristics, we explored the use of the ENF signal as a signature that can be used within a learning framework to identify the grid-of-origin of a media recording. This can pave the way to identify the origins of such videos as those of terrorist attacks, ransom demands, and child exploitation.

In our work on the flicker signal, we focused on problems in the entertainment industry addressing issues in movie piracy related investigations. The flicker signal is an artifact created when a camcorder is used to record media content shown on a Liquid Crystal Display (LCD) screen. We built an analytic model of the flicker, relating it to inner parameters of the camcorder and the screen used to produce the flicker-containing video, and showed that it can be used to identify and/or characterize the camcorder and the screen used to produce the flicker-containing video.

Our work on the flicker signal also inspired an ENF-based application, discussed in this dissertation, where we use the ENF traces captured in a video to characterize the video camera that had produced it. This was an example of how the study of one forensic signature can inspire new work on another forensic signature.

Several interesting lines of research along the directions of the work discussed in this dissertation can be explored. We have discussed that, despite the large amount of work in the research community targeted at estimating the ENF signal and using it for applications, we still do not have a solid understanding on the specific recording situations that would confidently lead to the capture of ENF traces in recordings.

In our study on the factors affecting the capture of ENF traces in audio recordings, we have sensed the need for a larger scale comprehensive study that would examine various factors that can affect the capture of ENF traces in both audio and video recordings, related to the environment in which a recording is made and the devices used to make it. Such a study would be important to gain a better understanding on the true real-world applicability of the ENF signal.

In addition to examining the factors affecting the capture of ENF traces, modeling the statistical properties of the ENF signal, especially across different seasons of a year and times of a day, can help researchers understand better how this signature behaves. This could also help inform the design of ENF-based applications and improve their performance.

In this context, we believe that there are still more real-world applications that the ENF signal can be used for, both in multimedia forensics and elsewhere, and it would be interesting to explore such applications as well as improve on the performance of the ones proposed already. For instance, our work on using the ENF for grid-of-origin identification is a first work of its type, so there is room for further improvement. This can possibly be achieved through the collection and use of more ENF data and the exploration of additional features and different learning frameworks.

Similar lines of research can be explored for the flicker signal. Carrying out large scale validation experiments with a wide diversity of devices and benchmarking evaluations against subsequent video processing can help in gaining a better understanding on the applicability of flicker forensics. Also, our improved under-

standing of the flicker signal through the work for this dissertation may be exploited to enhance techniques for removing flicker from a video, which can help improve the performance of watermark-detecting systems. Another potential line of research pertains to looking for other statistical footprints in pirated movies that may reveal camcorder piracy.

Overall, it would be interesting to explore other signatures that can be intrinsically embedded in media, which can utilize and benefit from the analyses and frameworks investigated in this dissertation.

Bibliography

[1] Math H. J. Bollen and Irene Y. H. Gu. *Signal Processing of Power Quality Disturbances*. Wiley-IEEE Press, 2006.

[2] Catalin Grigoras. Digital audio recordings analysis: The electric network frequency (ENF) criterion. *International Journal of Speech, Language and the Law*, 12(1):63–76, 2005.

[3] Niklas Fechner and Matthias Kirchner. The humming hum: Background noise as a carrier of ENF artifacts in mobile device audio recordings. In *Proceedings of the International Conference on IT Security, Incident Management, and IT Forensics*, pages 3–13, May 2014.

[4] Ravi Garg, Avinash L. Varna, and Min Wu. 'Seeing' ENF: Natural time stamp for digital video via optical sensing and signal processing. In *Proceedings of the ACM International Conference on Multimedia*, pages 23–32, 2011.

[5] Doug W. Oard, Min Wu, Kari Kraus, Adi Hajj-Ahmad, Hui Su, and Ravi Garg. It's about time: Projecting temporal metadata for historically significant recordings. In *Proceedings of iConference*, pages 4–7, March 2014.

[6] Hui Su, Ravi Garg, Adi Hajj-Ahmad, and Min Wu. ENF analysis on recaptured audio recordings. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3018–3022, May 2013.

[7] Hui Su, Adi Hajj-Ahmad, Min Wu, and Doug W. Oard. Exploring the use of ENF for multimedia synchronization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4613–4617, May 2014.

[8] Hui Su, Adi Hajj-Ahmad, Chau-Wai Wong, Ravi Garg, and Min Wu. ENF signal induced by power grid: A new modality for video synchronization. In *Proceedings of the ACM International Workshop on Immersive Media Experiences*, pages 13–18, 2014.

[9] Catalin Grigoras. Applications of ENF criterion in forensic audio, video, computer and telecommunication analysis. *Forensic Science International*, 167(2-3):136–145, April 2007.

[10] Catalin Grigoras. Applications of ENF analysis in forensic authentication of digital audio and video recordings. *Journal of the Audio Engineering Society*, 57(9):643–661, 2009.

[11] Catalin Grigoras, Jeffrey Smith, and Christopher Jenkins. Advances in ENF database configuration for forensic authentication of digital media. In *Audio Engineering Society Convention 131*, October 2011.

[12] Catalin Grigoras and Jeff M. Smith. Advances in ENF analysis for digital media authentication. In *Proceedings of the AES International Conference on Audio Forensics*, June 2012.

[13] Mateusz Kajstura, Agata Trawinska, and Jacek Hebenstreit. Application of the electric network frequency (ENF) criterion: A case of a digital recording. *Forensic Science International*, 155(2-3):165–171, December 2005.

[14] Eddy B. Brixen. Techniques for the authentication of digital audio recordings. In *Audio Engineering Society Convention 122*, May 2007.

[15] Eddy B. Brixen. Further investigation into the ENF criterion for forensic authentication. In *Audio Engineering Society Convention 123*, October 2007.

[16] Eddy B. Brixen. ENF; quantification of the magnetic field. In *Proceedings of the AES International Conference on Audio Forensics*, June 2008.

[17] Alan J. Cooper. The electric network frequency (ENF) as an aid to authenticating forensic digital audio recordings – an automated approach. In *Proceedings of the AES International Conference on Audio Forensics*, June 2008.

[18] Alan J. Cooper. An automated approach to the electric network frequency (ENF) criterion: theory and practice. *International Journal of Speech, Language, and the Law*, 16(2):193–218, 2009.

[19] Alan J. Cooper. Further considerations for the analysis of ENF data for forensic audio and video applications. *International Journal of Speech, Language, and the Law*, 18(1):99–120, 2011.

[20] Richard Sanders and Peter S. Popolo. Extraction of electric network frequency signals from recordings made in a controlled magnetic field. In *Audio Engineering Society Convention 125*, October 2008.

[21] Richard W. Sanders. Digital audio authenticity using the electric network frequency. In *Proceedings of the AES International Conference on Audio Forensics*, June 2008.

[22] Yuming Liu, Zhiyong Yuan, Penn N. Markham, Richard W. Conners, and Yilu Liu. Wide-area frequency as a criterion for digital audio recording authentication. In *IEEE Power and Energy Society General Meeting*, pages 1–7, July 2011.

[23] Sean Coetzee. Phase and amplitude analysis of the ENF for digital audio authentication. In *Proceedings of the AES International Conference on Audio Forensics*, June 2012.

[24] Ode Ojowu, Johan Karlsson, Jian Li, and Yilu Liu. ENF extraction from digital recordings using adaptive techniques and frequency tracking. *IEEE Transactions on Information Forensics and Security*, 7(4):1330–1338, August 2012.

[25] Yuming Liu, Zhiyong Yuan, Penn N. Markham, Richard W. Conners, and Yilu Liu. Application of power system frequency for digital audio authentication. *IEEE Transactions on Power Delivery*, 27(4):1820–1828, October 2012.

[26] Ravi Garg, Avinash L. Varna, and Min Wu. Modeling and analysis of electric network frequency signal for timestamp verification. In *Proceedings of the IEEE International Workshop on Information Forensics and Security*, pages 67–72, December 2012.

[27] Adi Hajj-Ahmad, Ravi Garg, and Min Wu. Instantaneous frequency estimation and localization for ENF signals. In *Proceedings of the Asia-Pacific Signal Information Processing Association Annual Summit and Conference*, pages 1–10, December 2012.

[28] Wei-Hong Chuang, Ravi Garg, and Min Wu. How secure are power network signature based time stamps? In *Proceedings of the ACM Conference on Computer and Communications Security*, pages 428–438, 2012.

[29] Ravi Garg, Adi Hajj-Ahmad, and Min Wu. Geo-location estimation from electrical network frequency signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2862–2866, May 2013.

[30] Ling Fu, Penn N. Markham, Richard W. Conners, and Yilu Liu. An improved discrete fourier transform-based algorithm for electric network frequency extraction. *IEEE Transactions on Information Forensics and Security*, 8(7):1173–1181, July 2013.

[31] Ravi Garg, Avinash L. Varna, Adi Hajj-Ahmad, and Min Wu. 'Seeing' ENF: Power-signature-based timestamp for digital multimedia via optical sensing

and signal processing. *IEEE Transactions on Information Forensics and Security*, 8(9):1417–1432, September 2013.

[32] Adi Hajj-Ahmad, Ravi Garg, and Min Wu. Spectrum combining for ENF signal estimation. *IEEE Signal Processing Letters*, 20(9):885–888, September 2013.

[33] Jidong Chai, Fan Liu, Zhiyong Yuan, Richard W. Conners, and Yilu Liu. Source of ENF in battery-powered digital recordings. In *Audio Engineering Society Convention 135*, October 2013.

[34] Adi Hajj-Ahmad, Ravi Garg, and Min Wu. ENF based location classification of sensor recordings. In *Proceedings of the IEEE International Workshop on Information Forensics and Security*, pages 138–143, November 2013.

[35] Wei-Hong Chuang, Ravi Garg, and Min Wu. Anti-forensics and countermeasures of electrical network frequency analysis. *IEEE Transactions on Information Forensics and Security*, 8(12):2073–2088, December 2013.

[36] Hui Su, Adi Hajj-Ahmad, Ravi Garg, and Min Wu. Exploiting rolling shutter for ENF signal extraction from video. In *Proceedings of the IEEE International Conference on Image Processing*, pages 5367–5371, October 2014.

[37] Adi Hajj-Ahmad, Ravi Garg, and Min Wu. ENF-based region-of-recording identification for media signals. *IEEE Transactions on Information Forensics and Security*, 10(6):1125–1136, June 2015.

[38] Luke Dosiek. Extracting electrical network frequency from digital recordings using frequency demodulation. *IEEE Signal Processing Letters*, 22(6):691–695, June 2015.

[39] Maarten Huijbregtse and Zeno Geradts. Using the ENF criterion for determining the time of recording of short digital audio recordings. In *Proceedings of the International Workshop on Computational Forensics*, pages 116–124, August 2009.

[40] Daniel Patricio Nicolalde and Jose Antonio Apolinario. Evaluating digital audio authenticity with spectral distances and ENF phase change. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 1417–1420, April 2009.

[41] Daniel Patricio Nicolalde Rodríguez, José Antonio Apolinário, and Luiz Wagner Pereira Biscainho. Audio authenticity: Detecting ENF discontinuity with high precision phase analysis. *IEEE Transactions on Information Forensics and Security*, 5(3):534–543, September 2010.

[42] Daniel P. Nicolalde-Rodríguez, José A. Apolinário, and Luiz W. P. Biscainho. Audio authenticity based on the discontinuity of ENF higher harmonics. In

Proceedings of the European Signal Processing Conference, pages 1–5, September 2013.

[43] Paulo Antonio Andrade Esquef, José Antonio Apolinário, and Luiz W. P. Biscainho. Edit detection in speech recordings via instantaneous electric network frequency variations. *IEEE Transactions on Information Forensics and Security*, 9(12):2314–2326, December 2014.

[44] Feng-Cheng Chang and Hsiang-Cheh Huang. Electrical network frequency as a tool for audio concealment process. In *Proceedings of the International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 175–178, 2010.

[45] Feng-Cheng Chang and Hsiang-Cheh Huang. A study on ENF discontinuity detection techniques. In *Proceedings of the International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 9–12, 2011.

[46] Moustafa M. Eissa, Mahmoud M. Elmesalawy, Yilu Liu, and Hossam Gabbar. Wide area synchronized frequency measurement system architecture with secure communication for 500kv/220kv egyptian grid. In *Proceedings of the IEEE International Conference on Smart Grid Engineering*, pages 1–1, August 2012.

[47] Mahmoud M. Elmesalawy and Moustafa M. Eissa. New forensic ENF reference database for media recording authentication based on harmony search technique using gis and wide area frequency measurements. *IEEE Transactions on Information Forensics and Security*, 9(4):633–644, April 2014.

[48] Dima Bykhovsky and Asaf Cohen. Electrical network frequency (ENF) maximum-likelihood estimation via a multitone harmonic model. *IEEE Transactions on Information Forensics and Security*, 8(5):744–753, May 2013.

[49] Guang Hua, Jonathan Goh, and Vrizlynn L. L. Thing. A dynamic matching algorithm for audio timestamp identification using the ENF criterion. *IEEE Transactions on Information Forensics and Security*, 9(7):1045–1055, July 2014.

[50] G. Hua, Y. Zhang, J. Goh, and V. L. L. Thing. Audio authentication by exploring the absolute-error-map of ENF signals. *IEEE Transactions on Information Forensics and Security*, 11(5):1003–1016, May 2016.

[51] Adi Hajj-Ahmad, Andrew Berkovich, and Min Wu. Exploiting power signatures for camera forensics. *IEEE Signal Processing Letters*, 23(5):713–717, May 2016.

[52] William Rosenblatt, Stephen Mooney, and William Trippe. *Digital Rights Management: Business and Technology.* John Wiley & Sons, Inc., 2001.

[53] Ingemar Cox, Matthew Miller, Jeffrey Bloom, Jessica Fridrich, and Ton Kalker. *Digital Watermarking and Steganography*. Morgan Kaufmann Publishers Inc., 2nd edition, 2008.

[54] Teddy Furon and Gwenaël Doërr. Tracing pirated content on the internet: Unwinding Ariadne's thread. *IEEE Security & Privacy*, 8(5):69–71, September-October 2010.

[55] Ji-Won Lee, Min-Jeong Lee, Hae-Yeoun Lee, and Heung-Kyu Lee. Screenshot identification by analysis of directional inequality of interlaced video. *EURASIP Journal on Image and Video Processing*, 2012(7):1–15, May 2012.

[56] Marco Visentini-Scarzanella and Pier Luigi Dragotti. Video jitter analysis for automatic bootleg detection. In *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, pages 101–106, September 2012.

[57] Paolo Bestagini, Marco Visentini-Scarzanella, Marco Tagliasacchi, Pier Luigi Dragotti, and Stefano Tubaro. Video recapture detection based on ghosting artifact analysis. In *Proceedings of the IEEE International Conference on Image Processing*, pages 4457–4461, September 2013.

[58] Xavier Rolland-Nevière, Bertrand Chupeau, Gwenaël Doërr, and Laurent Blondé. Forensic characterization of camcorded movies: Digital cinema vs. celluloid film prints. In *Media Watermarking, Security, and Forensics*, volume 8303 of *Proceedings of SPIE*, January 2012.

[59] Yingchen Zhang, Penn Markham, Tao Xia, Lang Chen, Yanzhu Ye, Zhongyu Wu, Zhiyong Yuan, Lei Wang, Jason Bank, Jon Burgett, Richard W. Conners, and Yilu Liu. Wide-area frequency monitoring network (FNET) architecture and applications. *IEEE Transactions on Smart Grid*, 1(2):159–167, September 2010.

[60] Arun G. Phadke, James S. Thorp, and Mark G. Adamiak. A new measurement technique for tracking voltage phasors, local system frequency, and rate of change of frequency. *IEEE Transactions on Power Apparatus and Systems*, PAS-102(5):1025–1038, May 1983.

[61] FNET Server Web Display. [Online]. Available:http://fnetpublic.utk.edu/. Accessed: December 2015.

[62] Zhian Zhong, Chunchun Xu, Billian J. Billian, Li Zhang, Shu-Jen Steven Tsai, Richard W. Conners, Virgilio A. Centeno, Arun G. Phadke, and Yilu Liu. Power system frequency monitoring network (FNET) implementation. *IEEE Transactions on Power Systems*, 20(4):1914–1921, November 2005.

[63] Shu-Jen Tsai, Li Zhang, Arun G. Phadke, Yilu Liu, Michael R. Ingram, Sandra C. Bell, Ian S. Grant, Dale T. Bradshaw, David Lubkeman, and Le Tang. Frequency sensitivity and electromechanical propagation simulation study in

large power systems. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 54(8):1819–1828, August 2007.

[64] Philip Top, Mark R. Bell, Ed Coyle, and Oleg Wasynczuk. Observing the power grid: Working toward a more intelligent, efficient, and reliable smart grid with increasing user visibility. *IEEE Signal Processing Magazine*, 29(5):24–32, September 2012.

[65] Julius O. Smith and Xavier Serra. PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation. *International Computer Music Conference*, 2004.

[66] George-Othon Glentis and Andreas Jokobssen. Time-recursive IAA spectral estimation. *IEEE Signal Processing Letters*, 18(2):111–114, February 2011.

[67] Dimitris G. Manolakis, Stephen M. Kogon, and Vinay K. Ingle. *Statistical and Adaptive Signal Processing*. MaGraw-Hill, Inc., 2000.

[68] Alex C. Kot, S. Parthasarathy, Donald W. Tufts, and Richard J. Vaccaro. Statistical performance of single frequency estimation in white noise using state variable balancing and linear prediction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35:1639–1642, November 1987.

[69] Yinong Ding and Richard J. Vaccar. Determination of data matrix dimensions for subspace-based parameter estimation algorithms. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 5:2547 – 2550, May 1996.

[70] Peter Stoica and Randolph Moses. *Spectral Analysis of Signals*. Prentice Hall, Inc., 2005.

[71] http://www.sal.ufl.edu/newcomers/enf_data.rar. Accessed: July 2012.

[72] Andrea Goldsmith. *Wireless Communications*. Cambridge Univ. Press, 2005.

[73] Mads Græsbøll Christensen, Søren Holdt Jensen, Søren Vang Andersen, and Andreas Jakobsson. Subspace-based fundamental frequency estimation. In *Proceedings of the European Signal Processing Conference*, pages 637–640, September 2004.

[74] Monson H. Hayes. *Statistical Digital Signal Processing and Modeling*. Wiley, 1996.

[75] Olympus Digital Voice Recorder Detailed Instructions. [Online]. Available:http://www.olympusamerica.com/files/oima_cckb/ws-710m_ws-700m_ws-600s_instructions_en.pdf. Accessed: August 2016.

[76] Sound Waves and Music. [Online]. Available: http://www.physicsclassroom.com/class/sound/. Accessed: June 2016.

[77] The Doppler Effect and Sound Interference – Scientific Sofia. [Online]. Available: http://dropdeadsofia.weebly.com/the-doppler-effect-and-sound-interference.html. Accessed: August 2016.

[78] UMCP / Kim Engineering Building Floor Plan – Floor 2. [Online]. Available: https://apra.umd.edu/publicAccess/kim/floors.jsp?floor=2. Accessed: July 2016.

[79] TYPE 4191 – Brüel & Kjær sound & vibration. [Online]. Avaliable: https://www.bksv.com/en/products/transducers/acoustic/microphones/microphone-cartridges/4191. Accessed: August 2016.

[80] Electret Microphone Amplifier - MAX4466. [Online]. Available: https://www.adafruit.com/product/1063. Accessed: August 2016.

[81] UNICEF: Child protection from violence, exploitation and abuse child trafficking. URL: http://www.unicef.org/protection/57929_58005.html.

[82] Palghat P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice Hall, 1993.

[83] Martin Vetterli and Jelena Kovačević. *Wavelets and Subband Coding*. Originally published by Prentice Hall PTR, Englewood Cliffs, New Jersey, 1995.

[84] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/$\sim$cjlin/libsvm`.

[85] Yi-Min Huang and Shu-Xin Du. Weighted support vector machine for classification with uneven training class sizes. In *Proceedings of the International Conference on Machine Learning and Cybernetics*, volume 7, pages 4365–4369 Vol. 7, 2005.

[86] Ji Ming, Peter Jancovic, Phillip Hanna, and Darryl Stewart. Modeling the mixtures of known noise and unknown unexpected noise for robust speech recognition. In *Interspeech*, pages 1111–1114, 2001.

[87] Ji Ming, Timothy J. Hazen, James R. Glass, and Douglas A. Reynolds. Robust speaker recognition in noisy conditions. *IEEE Transactions on Audio, Speech and Language Processing*, 15(5):1711–1723, 2007.

[88] Douglas A. Reynolds. Gaussian mixture models. In *Encyclopedia Biometric Recognition*. New York, NY, USA: Springer-Verlag, 2009, pages 659-663.

[89] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley and Sons, 2001.

[90] M. Wu, A. Hajj-Ahmad, M. Kirchner, Y. Ren, C. Zhang, and P. Campisi. Location signature that you don't see – highlights from the IEEE signal processing cup 2016 student competition. *IEEE Signal Processing Magazine*, 33(5):149–156, September 2016.

[91] Omar Ait-Aider, Adrien Bartoli, and Nicolas Andreff. Kinematics from lines in a single rolling shutter image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6, June 2007.

[92] Jinwei Gu, Yasunobu Hitomi, Tomoo Mitsunaga, and Shree Nayar. Coded rolling shutter photography: Flexible space-time sampling. In *Proceedings of the IEEE International Conference on Computational Photography*, pages 1–8, March 2010.

[93] Adi Hajj-Ahmad, Séverine Baudry, Bertrand Chupeau, Gwenaël Doërr, and Min Wu. Flicker forensics for camcorder piracy. *IEEE Transactions on Information Forensics and Security*, 12(1):89–100, January 2017.

[94] Adi Hajj-Ahmad, Séverine Baudry, Bertrand Chupeau, and Gwenaël Doërr. Flicker forensics for pirate device identification. In *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, IH&MMSec, pages 75–84, Portland, Oregon, USA, 2015.

[95] Digital Cinema Initiatives, LLC. *Digital Cinema System Specification*, 1.2 edition, March 2008.

[96] MovieLabs. MovieLabs specifications for next generation of video and enhanced content protection. Technical report, 2013.

[97] Juan José Moreira-Pérez, Bertrand Chupeau, Séverine Baudry, and Gwenaël Doërr. Exploring color information to characterize camcorder piracy. In *Proceedings of the IEEE International Workshop on Information Forensics and Security*, pages 132–137, November 2013.

[98] Bertrand Chupeau, Séverine Baudry, and Gwenaël Doërr. Forensic characterization of pirated movies: Digital cinema cam vs. optical disc rip. In *Proceedings of the IEEE International Workshop on Information Forensics and Security*, pages 155–160, December 2014.

[99] Thirapiroon Thongkamwitoon, Hani Muammar, and Pier-Luigi Dragotti. An image recapture detection algorithm based on learning dictionaries of edge profiles. *IEEE Transactions on Information Forensics and Security*, 10(5):953–968, May 2015.

[100] Badak Mahdian, Adam Novozámský, and Stanislav Saic. Identification of aliasing-based patterns in re-captured LCD screens. In *Proceedings of the IEEE International Conference on Image Processing*, September 2015.

[101] Dwight Poplin. An automatic flicker detection method for embedded camera systems. *IEEE Transactions on Consumer Electronics*, 52(2):308–311, May 2006.

[102] Naehyuck Chang, Inseok Choi, and Hojun Shim. DLS: Dynamic backlight luminance scaling of liquid crystal display. *IEEE Transactions on Very Large Scale Integration*, 12(8):837–846, August 2004.

[103] Abbas El Gamal and Helmy Eltoukhy. Cmos image sensors. *IEEE Circuits and Devices Magazine*, 21(3):6–20, May 2005.

[104] Chia-Kai Liang, Li-Wen Chang, and Homer H. Chen. Analysis and compensation of rolling shutter effect. *IEEE Transactions on Image Processing*, 17(8):1323–1330, August 2008.

[105] Séverine Baudry, Bertrand Chupeau, Mario de Vito, and Gwenaël Doërr. Modeling the flicker effect in camcorded videos to improve watermark robustness. In *Proceedings of the IEEE International Workshop on Information Forensics and Security*, pages 42–47, Atlanta, GA, USA, December 2014.

[106] Gwenaël Doërr and Jean-Luc Dugelay. Security pitfalls of frame-by-frame approaches to video watermarking. *IEEE Transactions on Signal Processing*, 52(10):2955–2964, October 2004.

[107] Jan Lukas, Jessica Fridrich, and Miroslav Goljan. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2):205–214, June 2006.

[108] Ashwin Swaminathan, Min Wu, and K. J. Ray Liu. Nonintrusive component forensics of visual sensors using output images. *IEEE Transactions on Information Forensics and Security*, 2(1):91–106, March 2007.

[109] Ashwin Swaminathan, Min Wu, and K. J. Ray Liu. Component forensics. *IEEE Signal Processing Magazine*, 26(2):38–48, March 2009.

[110] Yoonjong Yoo, Jaehyun Im, and Joonki Paik. Flicker removal for CMOS wide dynamic range imaging based on alternating current component analysis. *IEEE Transactions on Consumer Electronics*, 60(3):294–301, August 2014.

[111] Philipp Schaber, Stephan Kopf, Sina Wetzel, Tyler Ballast, Christoph Wesch, and Wolfgang Effelsberg. CamMark: Analyzing, modeling, and simulating artifacts in camcorder copies. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 11(2s):42:1–42:23, February 2015.

[112] Philipp Schaber, Sally Dong, Benjamin Guthier, Stephan Kopf, and Wolfgang Effelsberg. Modeling temporal effects in re-captured video. In *Proceedings of the ACM Conference on Multimedia*, pages 1279–1282, October 2015.

[113] Mo Chen, Jessica Fridrich, Miroslav Goljan, and Jan Lukáš. Determining image origin and integrity using sensor noise. *IEEE Transactions on Information Forensics and Security*, 3(1):74–90, March 2008.