ENVIRONMENTAL SOUNDS: ACQUISITION, ANALYSIS, AND REPRESENTATION

A Dissertation
Presented to
The Academic Faculty

By

Muhammad Umair Bin Altaf

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
in
Electrical and Computer Engineering

School of Electrical and Computer Engineering
Georgia Institute of Technology
August 2015

ENVIRONMENTAL SOUNDS: ACQUISITION, ANALYSIS, AND REPRESENTATION

Approved by:

Dr. Mark A. Clements, Committee Chair
Professor, School of ECE
Georgia Institute of Technology

Dr. Pamela T. Bhatti
Assoc. Professor, School of ECE
Georgia Institute of Technology

Dr. Biing-Hwang (Fred) Juang, Advisor
Professor, School of ECE
Georgia Institute of Technology

Dr. Yao Xie
Asst. Professor, School of ISYE
Georgia Institute of Technology

Dr. David V. Anderson
Professor, School of ECE
Georgia Institute of Technology

Date Approved: 21 July 2015

To My Muses...

ACKNOWLEDGMENTS

I wish to express my sincere appreciation to my adviser, Professor Fred Juang, for his endless hours of patience, guidance, encouragement, and support. I am extremely fortunate to have been able to work with a researcher and intellect of his caliber. I am also grateful to Professors David V. Anderson, Mark A. Clements, Pamela T. Bhatti, and Yao Xie for their efforts on my PhD defense committee.

During my stay at Georgia Tech, I have met many dedicated students. In particular, I am grateful to my fellow colleagues and graduate students, past and present, in Fred's group: Enrique Robledo-Arnuncio, Antonio Moreno-Daniel, Roy Munkong, Soohyun Bae, Ted Wada, Dwi S. Mansjur, Yong Zhao, Jason Wung, Sunghwan Shin, Migyu Chen, Chao Weng, and Zhong Meng. I thank Taras Butko for serving as my intellectual touchstone. This work has improved significantly as a result of his feedback. Mehrez Souden must also be mentioned for providing guidance in designing the localization algorithms. I would also like to thank the staff at CSIP and the ECE graduate Office: Pat Dixon, Daniela Staiculescu, Tasha Torrence, Jennifer Bomar, and Raqeul Plasket for administrative support.

My friends Salman Asif, Muhammad Omer, and Farsat Munir made my life outside of the lab a pleasure. The sleepless nights engaged in philosophical debates were the highlights of my life at Tech. I should also mention Vivek Kaul, to whom I owe an immense debt of gratitude for introducing me to "the life of mind".

I cannot overstate my gratitude to my parents, my sister and brothers, who encouraged and supported me at every step of the way. My deepest appreciation goes out to you all. Finally, I wish to thank my wife, Mariam. We married in the middle of my PhD program. Her emotional support and sacrifice during the ups and downs of made it possible for me to keep going.

I would also like to acknowledge the partial support of the Fulbright Program.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The dissertation presents the design and development of a systematic signal analysis and representation framework beyond short-time Fourier power spectrum for sounds, in particular environmental sounds. This framework is consistent with the underlying assumptions of the analysis method and its elements are correlated with human perception. The sound signal has to conform to certain conditions for its power spectrum to have a physical and perceptual meaning. We contend that very few environmental sounds readily meet these criteria and argue that the quantities that are traditionally used to describe sounds need to be repurposed and, if necessary, redefined to represent sounds by non-Fourier means. We propose a perceptuo-analytic organization of sounds so that any environmental sound can be analyzed based on its signal characteristics and perception.

We present environmental sound acquisition in the context of collection and annotation of a database for the footstep sounds, a common environmental sound, and show that it can be represented by these unconventional means and further analyzed to produce descriptions which are obscured with the traditional analysis. We present a novel application of extracting gait characteristics from the footstep sounds which is enabled by the proposed framework.

CHAPTER 1

INTRODUCTION

1.1 Background and Motivations

Audible sounds provide a rich source of information about the ambient environment. For humans, detecting and recognizing such sounds is the first step towards developing an acoustic awareness of the environment which surrounds them followed by formulating a response. It is no surprise, then, that humans, and other animals, have after a long evolutionary process developed a particular sense of hearing; they devote considerable area within the cortical regions of the brain to this task.

As the term suggests, audible sounds encompass the full hearing ability of the human ear and range from rudimentary impact sounds and sophisticated animal calls to speech and musical sounds. Speech and music have been widely studied within the signal processing community, which is indicated in commercially available automatic speech recognition (ASR) and music information retrieval systems. It is a reflection of the complex interplay of preferences assigned by researchers, funders, and consumers to these type of sounds. However, the importance that humans assign to each type is clearly not equal and can vary with the technological frame of reference within which the human may be operating. For example, historically, on the evolutionary time scale, music has not been as important as the naturally occurring sounds in the environment since these sounds provide crucial information which had survival value during our evolutionary development over hundreds of millions of years.

Sounds which cannot be categorized as either music or speech have been widely termed *environmental sounds*. These sounds are instrumental in generating an awareness of the surrounding environment as the various active and passive sources that constitute the acoustic environment generate, reflect, and modify the sound in ways which leaves their imprints within the sound signal. This information becomes parts of our understanding in subtle

ways and we are consciously unaware of this process unless we acoustically isolate ourselves from the environment, e.g., in anechoic chambers where people may feel disoriented because the information derived from room reverberation cues is missing or when we try to drive a car or cross a busy road while listening to music with our headphones. It is at these times that we realize that these sounds generate a subconscious awareness of the surroundings, e.g., indicates the presence of invisible objects, provides location, focuses our attention if its important, orients us with respect to the surroundings—and all of this in a timely manner. The awareness of the surroundings generated through this first-order environmental sound analysis is an indispensable tool in the repertoire of an intelligent self-aware being. If one wants to augment this awareness for disabled humans under sensory deprived conditions or replicate it in autonomous systems, such as robots, an appreciation of the role played by environmental sounds is necessary.

Despite the information embedded in the environmental sounds, the characterization of environmental sounds as noise is quite prevalent. In fact, the term environmental sound first appeared in an abstract on noise control [1]. This view reflects an attitude on part of researchers in many fields who find these sounds uninteresting and thus worthy of little independent attention. Part of that can be traced to an underlying bias in favor of the visual modality; after all, we view ourselves as visual species, but that does not explain the extent of the observed oversight for such sounds.

Arguably, this oversight can be due to the analysis framework used to study environmental sounds in many disciplines. In particular, within the signal processing literature the environmental sounds are generally studied under the same theoretical framework as speech and music—the short-time Fourier analysis (STFA). This is surprising given that these sounds differ from speech and music, and among themselves, in their production mechanism, human perception, and signal structures. The choice of analysis framework inevitably obscures the potential information that is available within these sounds. The prevalence of environmental sounds combined with the ubiquity of devices able to capture

them in the environment only serves to highlight this gap in the analysis framework.

Almost by definition, environmental sounds also do not carry much lexical information or a semantic organizing structure, which means that the basic sound characteristics themselves take on a more prominent role in the information that the sound carries. In the absence of the lexical information it becomes important to look at all dimensions of signal characteristics, in particular non-spectral characteristics. We can then incorporate this information to improve its representation and analysis.

From the conventional standpoint, the spectral structure of many of these sounds seems arbitrary and thus uninformative. This is a correct assessment but only if one believes that the statistical construct of power spectrum—and its physical counterpart namely the frequency—estimated through STFA is the only structure that a sound can utilize to embody and convey information. It is not that power spectrum is a bad choice. The problem is that it seems the *only* choice for sound analysis when it is just one of the many guiding structures that organize and convey the information within the sounds. The diversity of such structures in environmental sounds serves to bring this anomaly into sharp focus.

This insight is motivated by the human auditory system. Beginning at the ear and extending up to and beyond the auditory cortex, this auditory system extracts, maintains, and utilizes multiple representations for sounds and as a result, it has been argued, exhibits capabilities in perceiving and extracting relevant information from the environmental sounds. The exact nature of the particular representations and the mechanisms employed by the biological auditory system are perhaps not as important as the principle of maintaining multiple representations relevant to the sound at hand.

1.2 Scientific Goals

The objective of the thesis is to design and develop a systematic sound analysis and representation framework beyond STFA for sounds, in particular environmental sounds, which is consistent with the underlying assumptions of the analysis method and whose elements

are correlated with human perception. We propose an organization of the signals for which the conventional STFA information is neither suitable nor useful because of these discrepancies. Once we understand the nature of the signal, we then propose analysis methods to automatically bring out the significance of a particular representation, ready for perceptual interpretations. We are guided by the following scientific goals:

a) An investigation of the discrepancies between the STFA model, with its mathematical foundations, and the human auditory perception, and relationship of the two with the temporal nature of the audio information.

b) A systematization of the analysis of environmental sounds by proposing a *signal profile* for the sound which creates a foundation for their study as signals that carry information. This provides an extensible and general structure to organize the sounds for further analysis.

c) A multi-dimensional sound representation framework. Once we understand the nature of the signal, we then propose analysis methods to automatically bring out the significance of a particular representation, ready for perceptual interpretations. We build a system which brings together the signal profile, the analysis methods, and the resulting representation.

d) Compilation of database of environmental sounds. When compared to speech and music, there are very few well-organized databases of diverse environmental sounds which enable interesting scientific research. We record a database of footstep sounds which allows for such research insights.

This has the potential to improve the acoustic event detection, classification, and recognition tasks and, as we will describe it may point to new applications which may not be apparent at present. It also has implications for speech and music analysis as it may lead to a better understanding for designing novel features and improving the current ones.

1.3 Dissertation Outline

We begin with an introduction to the environmental sounds and their treatment in scientific and engineering disciplines, in particular signal processing, in Chapter 2. We argue that as a distinct category environmental sounds are not well-understood or studied. We trace the history of the analysis framework, the STFA, which dominates sound analysis. As a sound category, environmental sounds inherit this legacy; however, very little attempts have been made to justify this legacy beyond the obvious analogy with other sound types.

In Chapter 3, we review the STFA representation and point out the violations of its basic assumptions which occur when this framework is used for environmental sounds. Using these discrepancies as our branching point, we propose an analysis hierarchy for these sounds. The first component of this hierarchy is the sound profile and the second component is set of non-Fourier analysis methods. The sound profile is a multi-dimensional characterization of a sound which provides an objective measure of the sound's distance from the conventional STFA framework and points towards a suitable analysis method. We also develop and validate the sound profile for sounds from multiple databases.

We demonstrate the application of the framework outlined in the analysis hierarchy on the regular sound of human footstep in Chapter 4. We estimate the sound profile for these sounds, choose suitable methods for the analysis from the hierarchy, design features for a conventional signal processing task of footstep sound detection, and show substantial improvement over published results. We record a database of footsteps sounds and attempt a rather unconventional task of biometrics extraction from footstep sounds. The proposed framework also allows us to design and develop a novel task of estimating clinical gait parameters from footstep sounds. We also demonstrate the statistical consistency and reliability of the acoustic gait analysis in Appendix A.

The muti-channel sound recordings of the new database are used to localize and track the footstep sound source in Chapter 5. We use a statistical particle filtering approach supported by a model of the sound source movement and evaluate the performance of this

algorithm using the manual annotations of the location.

We conclude with a summary of our contributions in this dissertation and offer perspectives on the limitations and interesting extensions of this work in Chapter 6.

CHAPTER 2

ENVIRONMENTAL SOUNDS

2.1 Introduction

Vanderveer [2] defined the term environmental sound as an "audible acoustic event which is caused by motions in the ordinary human environment ... [these are] vibrational fields which exist in the air, and are potential source of information for the auditory system ... [are usually] more complex than laboratory sinusoids ... [and the sounds are] taken in their literal rather than signal or symbolic interpretation".

This definition results in an expansive category containing a diverse array of impact sounds, water sounds, ambient sounds, wind sounds, fire crackling sounds, and body sounds, to name a few. This is in contrast to an easily recognizable category of sounds—the speech sound, which can be defined as a restricted subset of the sounds producible by the human vocal tract and recognizable within the confines of the spoken language. As a working definition, the concept of an environmental sound excludes sounds which contain semantic linguistic structures, such as the sounds carrying speech or music [3, 4, 5].

Within the space of all audible sounds, calling sounds other than speech and music as environmental can be inaccurate as both speech and musical sounds also occur in the same environment as the environmental sounds. This categorization can also be ambiguous as some sounds can be heard as part of a musical performance, and hence music, yet be understood as environmental sounds when heard in isolation. The same can be said about speech. Nevertheless, a symbolic linguistic structure is embedded in all sounds carrying speech and a harmonic structure supported by a musical language is embedded in musical sounds. The same cannot be said about environmental sounds, at least at this stage, though they do carry *meaning*, in the sense of being informative.

As might be expected for a relatively nascent area of research spread across multiple disciplines, the nomenclature for the environmental sound category is not settled. It has

been variously referred to as natural sounds [6], ambient sounds [7], or even just a type of noise [8, 3].

The issues with nomenclature reflect the disparate nature of disciplines under which these sounds have been studied. However, a lack of an overarching framework have made these sounds curiosities of individual researchers within the respective disciplines as opposed to the subject of a broad interdisciplinary scientific effort focused at understanding them. As a result, environmental sounds as an independent category have not been studied as systematically or as extensively as music and speech [4]. This deficiency is acutely felt as these sounds already play an important role, and modern technological developments are creating applications which place an increasing emphasis on them. The innovation in this area will only increase, since the devices which collect audio have become inexpensive and ubiquitous. For example, people using hearing aids face problems when using them in noisy environments. It is suggested in [9] that this may be due to difficulties in identifying sound sources. If the essential clues for sound identification were better understood, hearing aids might be designed that could make those cues available. Furthermore, understanding recognition of sources would also be useful in cases when paying attention to multiple information-bearing channels is essential, such as in an airplane cockpit or a battlefield. A cockpit can have as many as 100 different warning signals, but the maximum number operators can respond to appropriately at one time is much smaller, about four to six [10]. It would be useful to determine the acoustic features that enable identification of a large number of sound sources, as a guide to the design of effective warning signals and augmentation systems.

As we have mentioned, the environmental sounds have been studied under a broad range of disciplines, e.g., psychologists have used them to study hearing, ecologists use them to develop diversity maps of urban and wilderness landscapes [11], and Foley engineers have used them to create convincing immersive acoustic environments within movies, video games, and virtual reality applications [12]. Acoustic sensing uses them to detect

flaws in materials, mechanical diagnostics, fatigue, machine monitoring [13, 14, 15, 16] and even predict electrical faults [17]. Researchers in the multimedia, biomedical, and signal processing community have been engaged in developing varied applications of these sounds such as multi-modal activity tracking [18], scene identification [19], and sleep-apnoea detection [20].

Psychoacoustic researchers form the most prominent cohort of researchers interested in these sounds and they are interested in the hearing mechanism for these sounds. Their initial studies on environmental sounds focused on certain qualitative aspects such as pleasantness [21] or subjective preference [22]. Broader studies [23] added as many as 50 different aspects such as loudness, heaviness, wetness, etc. A semantic differential method was applied in [24] to 45 environmental sounds such as bird songs, waves crashing and people shouting. The aim was to associate the semantic attributes of sounds with three emotion-eliciting qualities: evaluation, activity and potency. In [25] 145 environmental sounds were used to obtain ratings representing various categories of naturally occurring environmental sounds on 20 binary scales. The researchers calculated values for 13 acoustic properties of the sounds, some based on statistics of the waveforms (centroid, skewness, number of peaks) and some based on listener judgments (pitch, energy variation) but the acoustic properties together accounted for no more than 60% of the variance for any factor.

While the above is a representative of the research effort in psychoacoustics on environmental sounds, in general, psychoacousticians have not used environmental sounds in their research, in part because of the difficulty of characterizing them acoustically, and in part because it was thought that studying environmental sounds would not add anything to the basic understanding of the auditory system. This assumption seems to be inaccurate. Cognitive psychologists working with environmental sounds have shown that presenting a label for a sound did not facilitate identification of the sounds, whereas prior exposure to the sound did [26]. Other cognitive studies seem to indicate, in general, that environmental sounds are remembered differently and less well than speech [27]. Whereas speech,

and to a lesser extent, music, can be abstracted away from the auditory stimulus, it seems that memory for environmental sounds is more explicitly bound to the details of the waveform [4].

2.2 Conventional Representation of Environmental Sounds—the Short-time Fourier Analysis

Short-time Fourier analysis (STFA) is the general framework that underlies modern modeling and analysis of speech and musical sounds. Briefly, STFA uses the short-time power spectral features from the Fourier transform. The mel-frequency cepstral coefficients (MFCC) [28] and perceptual linear prediction features (PLP) [29], used in speech and a multitude of short-time spectral, perceptual, and semantic features (summarized in [30]) used primarily with music, are based on STFA. Environmental sound analysis inherits this framework.

STFA in audio can be traced to the work of Ohm [31] and Helmholtz [32] on musical tones. They established the rough equivalence of the percept of *pitch* with the physical quantity *frequency* that is grounded in the mathematical description provided by Fourier analysis [33]. Further work by Nyquist [34] on the power spectral bandwidth of speech and by Fletcher [35] on cochlea—the seat of the frequency analysis in the hearing system—and the articulatory index [36], firmly established this framework for analysis of speech.

The institution of STFA as a framework for study of speech and music did provide an impetus for their subsequent study but, even in those early years, they were many critiques of a frequency-only characterization of hearing. It was pointed out that the percepts of the various combination tones [37] and the residue [38] cannot be interpreted by a Fourier power spectrum-only explanation. However, the sounds considered interesting at the time seemed not to be effected by these phenomena and these studies were deemed to be of limited interest.

Later researchers showed the importance of temporal and other factors for general audio [39, 40] and speech [41]. For example, the classical frequency response is not invariant to sound pressure levels [42], periodic sounds give rise to static, rather than oscillating perceptions and the sensation of impulsive sounds is that of sudden onset rather than white noise, as they should if the ear were a power spectrum analyzer. Moreover, in many cases amplitude modulations (AM), such as beats [43], are perceived as temporal fluctuations rather than side bands of a tone.

More recently, brain imaging studies have not only confirmed that the *tonotopic* or frequency-specific information is preserved in the highest regions of the auditory cortex, but that these regions also employ temporal encoding [44] to store auditory information. In addition to the spectro-temporal dichotomy, phase-locked response [45] and AM/frequency modulations are also preserved [46]. All of these dimensions are sourced from the inner ear structures.

2.3 Environmental Sounds and Signal Processing Applications

Researchers with engineering and computing backgrounds became interested in environmental sounds in the late 80s and early 90s. Initial work was focused on sound synthesis for creating immersive virtual environments and designing real world augmentation equipment [47], which required building physical models of the sounds and their sources. Later studies demonstrate physically motivated models of interaction between solid objects for a 3D virtual environment [12]. These interactions can produce impact, rolling, or scrapping sounds. The sounds are modeled as a combination of damped sinusoid oscillators. The perceptually motivated models of sound textures in [48] for classification task and the Sound Understanding Testbed grounded in the artificial intelligence framework [49] are other examples where certain types of environmental sounds were treated as an independent category.

Traditional signal processing researchers were influenced by computational models of

music in recognition, detection, classification, and retrieval tasks when they began work on environmental sounds. For example, a sound retrieval task was attempted in [50] using short-time Fourier features, an audio classification task using linear prediction features along with hidden Markov models (HMM) was described in [8], and short-time Fourier spectrum was used to model and detect sound transients in [51]. All of these systems, and many others [52, 53, 54, 30, 55, 56], rely on short-time Fourier analysis to model environmental sounds. The use of the then recently developed time-frequency (TF) distributions [57] were initially limited to specific environmental sounds which were transient or non-stationary. In [58] the Wigner-Ville distribution, a type of TF distribution, was used to classify under water acoustic signals and in [15] multiple TF distributions were used in an engine knock detection task using the pressure wave signals. A system for general sound recognition was built using wavelets in [59]. In [60] non-negative matrix factorization (NMF), a non-Cohen class time-frequency transform, was used to model non-stationary audio signals. However, such scant attempts tend to validate the particular technique, rather than the suitability of that technique for sound signal of interest.

Unlike speech or music, environmental sounds are not limited in their realization due to limitations of the source. The constraints of language, as in speech, or considerations of aesthetics, as in music, are also absent. As a result, the restrictions on the sound are only the limitations of the listener. This has led to the realization that the listener must have a way to learn some organization of these sounds. Many possible organizations or taxonomies have been proposed based on the source of sound [52], the context of those sources [53], or the similarity of the sound's perception [61].

In the context of sound organization, computational auditory scene analysis (CASA) framework has influenced how signal processing researchers view the problem of environmental sounds analysis [62]. CASA is an implementation of theories of ecological acoustics [2] and auditory streaming. Briefly, ecological acoustics theory claims that the listener

does not attend to sound per se but the event that produced it and this information is embedded in the characteristics of the sound. The auditory scene analysis [63] contextualizes multiple audio events (AEs) as part of an *auditory scene*. The audio events are differentiated from each other on the basis of their onset-times, source, location, or similar time and frequency dynamics. Audio events with similar characteristics constitute an *audio stream*. CASA systems mostly rely on TF transforms for feature-extraction (e.g., the systems in [19] or [64]). The audio stream theories are applicable to all sounds and not necessarily limited to the environmental sounds.

The National Institute of Standards and Technology (NIST), as part of their Classification of Events, Activities and Relationships (CLEAR) initiative, organized two evaluations for the acoustic event detection and classification tasks within a typical room as the acoustic scene [55, 56]. The influence of CASA is clear in the task names. The majority of the systems used MFCCs, even though speech was only one of the more than a dozen event-classes. The results from these evaluations were not encouraging, with the best system reporting an average detection accuracy at less than 40%. One of the major problems of such a low recognition rate is the presence of signal overlaps [65]. Another source of mistakes comes from mismatch conditions in training and testing. We will elaborate on this source of error in the next chapter. The CLEAR evaluations were also notable in that none of the nine systems used a time-frequency front end.

2.4 Summary

Environmental sounds have garnered some interest from multiple, and often disparate, scientific disciplines. As a result they have taken on the character of isolated curiosities of individual researchers and have failed to develop enough interest to merit a comprehensive, overarching framework for their study. From the initial beginnings, as a type of noise to be ignored, the potential of applications that can be realized with their analysis have brought

them into the spotlight, but only under the dominant paradigm of speech and music analysis—the STFA.

Of course, the sound analysis framework has been updated since the time of Ohm and Fletcher—specifically with the MFCCs, the non-linearities are incorporated with mel-frequency scales, and the cepstrum incorporates the peculiarities of the speech production. Moreover, a strict reliance on non-Fourier dimensions of analysis is just as narrow as any other isolated approach to environmental sound analysis.

Nevertheless, the diversity of environmental sounds exposes the problems that are inherent within the STFA framework. We aim to organize and analyze the sounds based on meaningful categories such as sounds with similar source, sounds which create a certain ambience or sounds which constitute a particular soundscape. For such goals, the sounds need to be properly analyzed. After proper analysis we can connect the analysis with sound perception and use the information to extract meaning relevant to humans.

In the following chapter, we review the STFA and point out the discrepancies in its basic assumptions with environmental sounds. We organize the analysis of these sounds into a hierarchy composed of the sounds profile and the accompanying analysis methods. In the absence of large, diverse, and appropriately labeled databases for environmental sounds, we evaluate the sound profile and validate its utility with an ad-hoc collection of such sounds and a limited number of sound source classes from the RWCP database.

CHAPTER 3

ANALYSIS AND REPRESENTATION OF ENVIRONMENTAL SOUNDS

We use the term *representation* to mean a description of the signal, in our case the sound, in a particular domain of interest in terms of the elements from that domain. A representation emphasizes some aspects of the signal, which are important for a particular task. The emphasis is typically the result of prior models and assumptions that are thought to be satisfied by the sounds *and* the systems that generate and analyze those sounds. The elements of the representation then provide the necessary information which broadly simplifies the presentation and understanding by developing and honing intuition for further analysis.

Physically, sounds are compression waves propagating through the transmission media in three dimensional space with time. The 2D amplitude-time plot, generally referred to as the waveform, is the most basic representation of sound that we notice in signal processing literature. Fourier transform provides another representation in the power spectral domain in terms of energy and frequency. The Fourier representation completely specifies a stationary sound if the system, that generated the sound or will analyze it, is linear. Thus, for further analysis, if the conditions of linearity and stationarity are satisfied, it does not matter which representation is relied upon. Some researchers also use the term representation or parametric representation to indicate the features that are extracted from a sound, such as the MFCCs, though we use the term representation in a much broader sense. Within our terminology, the MFCCs are derived from both the short-time spectral and cepstral representations of the sound and are more appropriately called parameters or features.

In the following, we will discuss the representations for environmental sounds. As we have stated, the representation provided by STFA in terms of energy and frequency does not completely model the full range of sounds that the human auditory system is capable of perceiving. In addition, the environmental sound need not conform to the conditions

for its power spectrum to be interpreted in physical and perceptual terms. To that end, we revisit STFA to elaborate on these criteria in Section 3.1. In Sections 3.2 and 3.3, we argue that very few environmental sounds readily meet these criteria and we show that the quantities that are traditionally used to describe sounds—the elements of the power spectral representation—need to be repurposed and, if necessary, redefined to represent sounds by non-Fourier means; in broad terms by incorporating their temporal variation. We proceed towards generalizing this process with the analysis hierarchy in Section 3.4 so that any environmental sound can be analyzed based on its signal characteristics and perception. Section 3.5 defines the sound profile. Finally, in Section 3.6 we evaluate the sound profile on different environmental sounds.

3.1 Short-Time Fourier Analysis (STFA)

For a deterministic signal, it is well known that the existence of its Fourier transform is governed by the Dirichlet conditions (see, for example, [66, p. 290]). The physical description of a deterministic signal is known completely. However, most signals encountered in practice, including sounds, are not modeled as deterministic. Instead, sounds are modeled as realizations of an underlying stochastic process, as it is not possible to specify the sound in advance of its observations. The stochastic process associates the uncertainty and randomness with the occurrence of a particular realization which, once realized, becomes an observation from the stochastic process.

Strictly speaking, the Fourier transform does not exist for all stochastic processes [67, p. 248]. In particular, the realizations of a wide-sense stationary (WSS) stochastic processes [68] are not absolutely integrable, which is one of the Dirichlet conditions. This is because a WSS process assumes, among other assumptions, that the value of its statistical expectation is constant for all time which implies that the integral will not converge and hence undefined.

This technical difficulty is resolved by the Wiener-Khinchin-Einstein (WKE) theorem

[69, 70]. WKE proves the existence of the Fourier transform of the autocorrelation function of WSS process—known as the power spectral density (PSD). Let $\mathbf{x}(t)$ be a WSS process at time t. Then, by WKE, the PSD of $\mathbf{x}(t)$ exits and is defined as[1]

$$S(\omega) = \int_{-\infty}^{\infty} R(\tau) e^{-j\omega\tau} d\tau, \tag{1}$$

where $R(\tau) = E\{\mathbf{x}(t+\tau)\mathbf{x}(t)\}$ is the autocorrelation. (1) is not generally used to calculate the PSD of the process. For a continuous signal $x(t)$, assumed to be a realization of a WSS process $\mathbf{x}(t)$, its Fourier transform (FT) is defined as

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt. \tag{2}$$

$|X(\omega)|^2$ is equivalent to the $S(\omega)$ of the signal. For a signal, this equivalence hinges on the WSS of the process, which is reflected in time invariance of first and second order statistics of $x(t)$. Intuitively, WSS ensures that the power indicated at particular frequency is physically part of the signal. For a non-WSS signal, the $\omega = 2\pi f$ in $|X(\omega)|^2$ has no physical correlate; in fact, the autocorrelation function is no longer a function of only the time difference but the time itself and thus the notion of frequency does not correspond to the power of a physical oscillation. We will elaborate on this point when we discuss frequency (cf. Section 3.2.3).

Information in a sound is encoded in the temporal variation of the various attributes of the sound signal. To estimate temporal variation in PSD, we modify (2) to introduce the short-time Fourier transform (STFT)

$$X_T(\omega, t) = \int_{-\infty}^{\infty} x(\tau) h(\tau - t) e^{-j\omega\tau} d\tau, \tag{3}$$

where, $h(t)$ is a time window with finite support, T, roughly given as

$$h(t) \approx \begin{cases} 1, & \text{for } t \in [0, T], \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

[1]We will state most of the definitions and results in this and subsequent chapters in continuous time, though some relations will be stated in continuous time and discrete time forms or discrete time forms only. All the results are equally valid for either forms, where applicable.

In general, window functions are not limited to the template in (4), though (4) does convey the idea behind the use of a time window function—observe a part of signal in isolation from the rest of signal. The resulting FT is a signal's frequency behavior during the time interval, T, covered by the data block within the window. We will subsequently drop the T in $X_T(\omega, t)$, when specifying the window interval is peripheral to our discussion.

$|X(\omega, t)|^2$ is frequently used as an estimate of the short-time PSD and its temporal variation, and has traditionally been the measurement of choice in extracting features used in speech, music, and as explained before, for environmental sounds. We term the analysis of sounds using $|X_T(\omega, t)|^2$ and it various transformations as STFA. As we have mentioned before, these features include the MFCCs, PLPs, for speech and almost all short-time spectral features for music which are used in sound analysis [30]. STFA derives its applicability from the assumption that the signal is WSS within T. This is referred to as *quasi*-WSS. Speech is commonly assumed to satisfy this assumption which justifies the uses of STFA and its derivatives in speech. Music has comparatively larger time windows as the signal is generally harmonic and, more often than not, changes more slowly than speech. We use the term harmonic for a sound to indicate presence of tones, not necessarily in musical harmony.

Broadly speaking, the discretization of the speech sound signal into perceptually and physically meaningful temporal intervals represents the attempt to minimize the potential discrepancies in the *quasi*-WSS assumption. The selection of the time window duration is based on the temporal structure of the signal and its perception. It reflects a compromise between capturing the temporal PSD variation and the interval of *quasi*-WSS. For speech, the window duration is generally set between 20-40 ms. The time derivatives of the PSD over multiple windows capture the temporal variation at the scale of 100 ms, which is at the higher end of the duration of a phoneme in spoken English [71]. The existence of perceptually detectable units of speech—a phoneme for standard ASRs—and higher lexical structures such as the type of language, words, and grammar, also help in organizing the

sound characteristics that occur in speech. The written musical language plays a similar role for musical compositions. Furthermore, speech is solely generated by a biological mechanism which imposes physiological limitations on the rates of movements that the articulators can sustain. This constraints the characteristics of the resulting sound, e.g., its limited spectral bandwidth.

With environmental sounds, the diversity of sounds and the sources that can produce such sounds precludes *a priori* assumptions on the source, though it may be possible to infer the nature of the source from certain sounds. *Quasi*-WSS is no longer uniformly applicable. Moreover, the sounds may not possess any harmonic regularity which could appear in the PSD estimates. This makes it dubious to apply the STFA framework on environmental sounds, as it is done conventionally, e.g., in the CLEAR evaluations [55].

3.1.1 CLEAR experiments

In the following we further illustrate this anomaly with experiments inspired from the CLEAR evaluations [55, 56] for sounds heard in a meeting room environment. We use the controlled databases for these evaluations that were recorded in the meeting rooms at the Universitat Politècnica de Catalunya (UPC) [72] in Spain and at Fondazione Bruno Kessler (FBK) [73] in Italy. The recorded sounds are semantically similar but they are recorded under different environmental conditions. The cross-environmental effect is in the form of different room impulse responses, variation in the sound sources, and the background noise. The evaluations did not study the effect of the different environments in the meeting room databases, which could have shown whether the systems are modeling the sound or the peculiarities of the environment. The sound classes are tabulated in Table 1.

We run audio event detection experiments with the UPC and FBK databases. The event detection task, as defined in CLEAR evaluations, includes the segmentation of the target sound in addition to detection. Instead, we only detect the class label of the sound for pre-segmented data under the two different conditions: matched and mismatched. The results in Table 1 are in terms of the equal error rate (EER) performance measure corresponding

Table 1: All numbers are percentage EERs. Adapted from [74].

Class Labels	Models trained using UPC train data		Models trained using FBK train data	
	UPC test data	FBK test data	FBK test data	UPC test data
Applause	0.0	11.1	0.0	13.3
Chair moving	2.6	68.6	3.0	42.1
Cough	1.5	13.9	0.0	23.1
Cup clink	0.0	69.4	0.0	4.7
Door close	1.6	61.5	0.0	13.1
Door knock	0.0	28.6	0.0	16.3
Door open	0.0	63.9	0.0	96.7
Footsteps	1.2	76.3	0.0	80.8
Key jingle	1.6	41.7	2.8	21.5
Keyboard typing	1.7	60.0	0.0	15.2
Laughter	4.7	19.4	0.0	48.4
Phone ringing	0.8	19.7	0.0	25.0
Paper wrapping	2.2	41.7	2.8	28.6
Average	**1.37**	**44.3**	**0.66**	**33.0**

to four different experimental scenarios, depending on the database (FBK or UPC) that is used to train and test the models. For matched conditions, we use 2/3 of the database to train and use the remaining part the same database to test. The results are averaged from a 3-fold cross-validation procedure. For mismatched conditions, we use the entire data from one database to train and use the entire data from the other database to test. We used a conventional MFCC/HMM based system. The features consist of 12 Mel-Frequency Cepstral Coefficients (MFCCs) including energy coefficient, extracted every 10 msec with a Hamming window of 25 msec The resulting parameters together with their first and second order time derivatives are arranged into a single observation vector of 39 components. Cepstral mean normalization is applied. Each sound is modeled by a 2-state full-connected HMM and each state is represented by a Gaussian mixture model (GMM) of 64 mixtures with diagonal covariance matrix. The training was accomplished using the standard Baum-Welch training procedure. For evaluation the sound events were cut from the continuous audio according to the ground-truth labels. Then the isolated audio segments were fed to each HMM corresponding to a set of acoustic classes to perform Viterbi decoding.

In Table 1 we notice that the EER is very low (1%) for matched conditions. However, in

mismatched conditions the EER increases drastically, indicating that the standard system with MFCCs based on STFA struggles to model the variations introduced by the cross-environmental effect. Among the problematic classes are 'footsteps', 'door open', 'paper wrapping', 'keyboard typing', and 'door knock'. These sounds belong to the general group of impact sounds that combine a sudden short burst of sound energy with long term temporal structure. Others such as 'phone ringing', 'cup clink', and 'key jingle' have spectral structure that can be estimated with PSD, but the use of uniform duration short-time windows ignores the different intervals of *quasi*-WSS that the sound classes may have. We will revisit these experiments in Section 4.4 where we analyze an impact sound, the footstep sound, and show that insights that follow from the proposed framework improve the detection performance for that particular sound.

We believe that the nuances that characterize such signals are not captured by the STFA-derived quantities as those quantities are bound by the mathematical constraints, which necessarily limit their scope to well-structured stationary signals. In the following, we discuss the effect of these bounds on sound analysis and perception as a first step towards the proposed framework.

3.2 Physical Descriptors of Sound

In the following, we elaborate on the basic descriptions of sound that are suggested by the STFT—frequency, energy, and phase—and point out the issues which arise when these are used to represent an environmental sound signal, and discuss the changes that can be made to the quantity in light of the non-Fourier perspective.

3.2.1 Energy

For a signal $x(t)$, its square, $|x(t)|^2$, is conventionally defined as the energy or intensity per unit time at time t.[2] This quantity is also known as energy density. The total energy, E is,

[2] Admittedly, energy as a function of time can always be described in terms of another quantity, power. Thus it has also been called instantaneous power.

the integration of density over all time, i.e.,

$$E = \int_{-\infty}^{\infty} |x(t)|^2 dt < \infty. \tag{5}$$

This *square law* estimate of the energy density is also used in the Fourier spectral representation of $x(t)$, the PSD. The PSD is usually defined on a per Hz basis.

The reasons for using the square law as a measure of energy are not as rigorous as its ubiquity might suggest. The basic result was derived for the energy of point sources emitting electromagnetic (EM) waves using Maxwell's equations in the early 20$^{\text{th}}$ century. It was developed when the study of EM resonances led to the discovery of radiation, a form of energy transfer in the space [75]. And since sound waves and EM waves are both *waves*, hence, by analogy, it has been extended to sound waves. Since then, the mathematical simplicity and utility of the estimate have made the squared law and energy identical with each other in signal processing literature. The analogy has weakened considerably since the days of classical physics and intensity of an EM radiation is no longer considered a measure of its energy. This is a result of Planck's energy-frequency relation [76]. In fact, the energy of an EM radiation is now taken as proportional to the frequency of the EM quanta as EM signals are no more modeled exclusively as waves.

Of course, we are not suggesting a purely frequency based measure for energy in sound signal processing. The quantum effects only become significant at very small scales and sound waves or the ear are certainly not operating at that scale. Rather, our motivations are intuitive: the energy required to produce a high frequency compressive vibration in a given time is more than that required to produce a comparatively low frequency vibration at the same amplitude. This is not reflected in the squared law density estimate. This observation motivates the Teager-Kaiser energy operator (TKEO) [77], discussed later, which estimates the kinetic and potential energy of a signal in terms of its amplitude and frequency.

The square law estimate may also not be ideal for impulsive signals. Let us have two sinusoids at frequency ω, where one, $p(t)$, is a pure tone and the other, $p'(t)$, is the same with an additional time-varying phase factor, $\phi(t)$, i.e., $p(t) = \sin(\omega t)$ and $p'(t) = \sin(\omega t + \phi(t))$.

We fix $\phi(t)$ to be the following piecewise continuous function,

$$\phi(t) = \begin{cases} 0, & t \in (2n, 2n+1], n \in \mathbb{Z}, \\ \pi, & t \in (2n+1, 2n+2], n \in \mathbb{Z}. \end{cases} \tag{6}$$

We note that $|p(t)|^2 = |p'(t)|^2$ at every time instant, t. This is inconsistent with the intuition of square law being an energy density, i.e., it does not have a one to one correspondence with the signal at every t. Any feature based on this measure will surely miss these transitions. It is also inconsistent with auditory perception, i.e., listening to the two waveforms, we can easily identify the instant in time where the phase transition occurs, as each transition in phase produces a sensation similar to that of an impact sound. Clearly, the ear is not acting as a squared law detector, at least in the short term.

Energy estimates can also be derived from the *instantaneous envelope* of the signal which can be calculated by, e.g., Hilbert transform (cf. Section 3.3.2). For signals that can be modeled with AM (cf. Section 3.3.3), the message signal provides a smoothed envelope or an *envelope* [78], for short.

3.2.2 Phase

Phase is an oft-ignored quantity in STFT. In (3), we can see that $X(\omega, t) \in \mathbb{C}$. This implies that it can be written in the polar form, i.e.,

$$X(\omega, t) = |X(\omega, t)| e^{j\theta_f(\omega,t)}. \tag{7}$$

Thus, the Fourier phase, $\theta_f(\omega, t) \triangleq \arg(X(\omega, t))$, constitutes the Fourier phase spectrum. The physical meaning of θ_f is clear for a class of deterministic signals which are a sum of sinusoids, i.e., if $x(t) = \sum_{k=1}^{N} a_k \cos(\omega_k t + \phi_k)$, then $\theta_f(\omega)_{|\omega=\omega_k} = \phi_k$ is the phase shift of each sinusoid. For less structured signals, it is widely believed that θ_f encodes temporal evolution of the frequency components in terms of phase delay and group delay [79]. θ_f is also more sensitive to temporal variations than $|X(\omega)|$ and the problems associated with *unwrapping* lead to limited utility in signal processing applications [80].

The phase information is ignored in the traditional STFA for speech. Among other things, this is reasonable because the short-time frame interval is chosen to be 20-40 ms. Kazama *et al.* [81] have shown that sentence intelligibility is not affected by changing short-time Fourier phase, if the frame interval is in the 10-70 ms range. If the frame interval is less than 4 ms or greater than 128 ms, changing the short-time Fourier phase significantly reduces sentence intelligibility. In our informal listening tests, we confirmed this perceptual effect with a number of audio events. The reconstructed signals and their perception changed significantly once their short-time Fourier phase was changed for very short (<10 ms) and long (>100 ms) frame intervals. This seems to indicate that Fourier phase can be ignored for the case of perceptually meaningful sounds only if the short-time interval is chosen to be in the range of 20-60 ms. As we discuss later, this frame interval may not be appropriate for the general acoustic event. Psychoacoustic studies also indicate that there are at least two time scales on which perception seems to operate–around 30 ms and 250 ms [39]. The latter overlaps with the intervals for which the Fourier phase is significant.

Let $x_{imag}(t)$ be an imaginary part of $x(t)$. Then, its complex extension is

$$z(t) = x(t) + jx_{imag}(t) = A(t)e^{j\theta_i(t)}, \tag{8}$$

where $\theta_i(t) \triangleq \arg(z(t))$ is the *instantaneous phase* and $A(t)$ is the instantaneous amplitude envelope of the signal. There are potentially infinite ways to calculate $x_{imag}(t)$ but two methods, the analytical signal method and the quadrature method, are mathematically tractable and provide physically meaningful estimates [57].

3.2.3 Frequency

The frequency f of a vibratory signal is classically defined as the number of oscillations per unit time, where vibratory signal is any signal that has temporal fluctuations and an oscillation is a complete repetition of the temporal fluctuation [82]. Fourier transform incorporates this definition of frequency to estimate $\omega = 2\pi f$ in (2). Dirichlet conditions

guarantee that all deterministic integrable functions can be written as a linear sum of sinusoids which allows separation of the classical frequency components.

To estimate the frequency of a signal, a sufficient amount of data is necessary in ensuring a meaningful result. If the fluctuations are not complete or are irregular then the estimates will be meaningless by definition. Clearly, this puts limits on the time resolution of any analysis method using the classical frequency measurements. Additionally, the restrictions on time resolution due to the time-frequency uncertainty principle [83] apply to the classical frequency. Moreover, the auditory system needs to be acting in a more-or-less linear manner for frequency as estimated in the PSD to be an effective correlate of perception. Nevertheless, when addressed in perceptual terms, the notion of frequency may be rather illusive. The perceptual phenomena of beats [43] and missing fundamentals [42] highlight this issue. We briefly discuss the former.

Beats are the slow and periodic fluctuations in peak amplitude which occur when two sinusoids with slightly different frequencies, f_1 and f_2, are linearly superimposed with $|f_1 - f_2|$ around 2-10 Hz as given in (9a)

$$b(t) = \sin(2\pi f_1 t) + \sin(2\pi f_2 t), \qquad |f_1 - f_2| < 10, \tag{9a}$$

$$= 2\sin\left(\frac{2\pi(f_1 + f_2)t}{2}\right)\cos\left(\frac{2\pi(f_1 - f_2)t}{2}\right). \tag{9b}$$

The signal shown in Figure 1, with $f_1 = 200$ Hz and $f_2 = 205$ Hz, audibly beats at 5 Hz even though the beating frequency is far below the normal hearing frequency range. It is clear from this description that the signal is WSS and thus it follows from (2) that the PSD will have two peaks for $b(t)$ at f_1 and f_2. This predicts that we should hear f_1 or f_2. Instead, when the difference in the two frequencies is small (within a few tens of cycles per second) we hear a tone around $(f_1 + f_2)/2$ beating at $|f_1 - f_2|$ Hz, not the two tones. A mathematically equivalent product form in (9b) provides an interpretation closer to the perception that frequencies $\frac{(f_1 \pm f_2)}{2}$ may also be present, though this is not clear from FT. The ambiguity in the perception model between an additive and multiplicative frequency

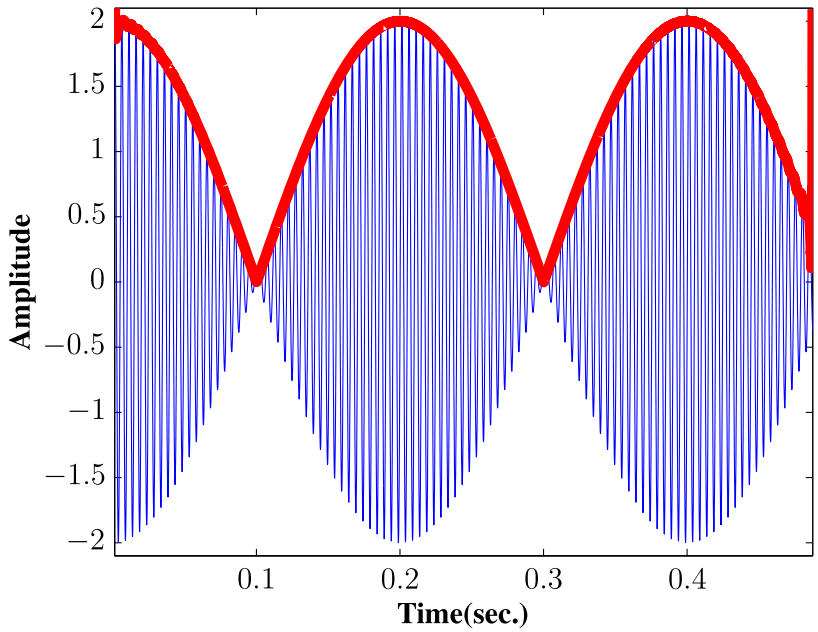


Figure 1: The signal which produces the auditory sensation of beats, in blue, and its Hilbert envelope, described in Section 3.3.2, in red.

indicates that a non-linear process is involved.

When we take $|X(\omega)|^2$ to be a spectral density function in the statistical sense[3], we can define a mean frequency $\bar{\omega}$ as the expectation of the density

$$\bar{\omega} \triangleq \int \omega |X(\omega)|^2 d\omega. \tag{10}$$

For a complex signal of the form of $z(t)$ in (8), it can be shown that $\bar{\omega}$ is given by [57]

$$\bar{\omega} = \int \dot{\theta}_i(t) A^2(t) dt. \tag{11}$$

(11) is the expectation of energy density, $A^2(t)$, over the derivative of instantaneous phase, $\dot{\theta}_i(t)$. Comparing (10) and (11), $\omega_i(t) \triangleq \dot{\theta}_i(t)$ is the *instantaneous frequency* (IF). $\omega_i(t)$ is a useful measure when the signal is not WSS—as signal characteristics need not be maintained over a time interval. If the average frequency from STFA is given by

$$\bar{\omega}_T(t) = \frac{\int \omega |X_T(\omega, t)|^2 d\omega}{\int |X_T(\omega, t)|^2 d\omega}, \tag{12}$$

then it can be shown that $\lim_{T \to 0} \bar{\omega}_T(t) \to \dot{\theta}_i(t)$, where convergence is in the mean [57].

[3]This is valid since $|X(\omega)|^2$ is non-negative, piecewise continuous function with a finite integral.

3.3 Estimation of Sound Descriptors

The previous section makes it clear that PSD, as estimated with STFA, is one of the many ways of representing a sound. In the following, we describe the methods which estimate the physical descriptors of sound without recourse to Fourier analysis.

3.3.1 Teager-Kaiser Energy Operator (TKEO)

TKEO [84] provides an unconventional perspective on the instantaneous energy of a signal. Building on the physical notion that the energy of a signal is the sum of its potential and kinetic energy. The TKEO is given as

$$\psi[x(t)] = \left(\frac{dx(t)}{dt}\right)^2 - x(t)\frac{d^2x(t)}{dt^2}, \tag{13}$$

where the time derivatives incorporate the rate of change of the signal and provide a measure of the potential energy. It can be shown that the quantity estimated in (13) relates signal energy to square of the signal amplitude *and* the square of its frequency [84]. We notice from (13) that for the signals $p(t)$ and $p'(t)$ from Section 3.2.1, $\psi[p(t)] \neq \psi[p'(t)]$. Moreover, TKEO can provide better localization of the impulses because the decay that follows an impulse is similar to an exponential decay and $\psi(e^{-bt}) = 0$, $\forall b \geq 0$.

The discrete instantaneous energy, $\psi_d(x[n])$ of a discrete signal $x[n]$ given by TKEO is

$$\psi_d(x[n]) = x^2[n] - x[n+1]x[n-1]. \tag{14}$$

3.3.2 Hilbert Transform (HT)

HT provides the Hilbert envelope—a measure of signal energy, instantaneous phase, and frequency by extending a signal into the complex domain. It is defined as

$$\mathcal{H}\{x(t)\} \triangleq \frac{1}{\pi}\mathrm{PV}\int_{-\infty}^{\infty}\frac{x(s)}{s-t}ds, \tag{15}$$

where PV indicates that the principal value of the integral is used. The complex signal formed by $x_h(t) = x(t) + j\mathcal{H}\{x(t)\}$ is the *analytic* representation of $x(t)$. The instantaneous Hilbert envelope is given by $\mathrm{env}_h(t) = |x_h(t)|^2$ and the derivative of the phase of $x_h(t)$ is the

IF associated with $x(t)$ i.e., $\omega_i(t) \triangleq \frac{d}{dt}\left(\arctan\left(\frac{\mathcal{H}\{x(t)\}}{x(t)}\right)\right)$. The Hilbert envelope of the beating signal in Figure 1 correctly depicts the perceptual sensation of that signal.

For the discrete signal $x[n]$, its discrete HT is

$$\mathcal{H}\{x[n]\} = x[n] * h[n], \tag{16}$$

where $*$ is the convolution operator and

$$h[n] \triangleq \begin{cases} 0, & n \in \text{even}, \\ \frac{2}{n\pi}, & n \in \text{odd}. \end{cases} \tag{17}$$

After obtaining $\mathcal{H}\{x[n]\}$, we construct a complex signal $x_a[n]$ as follows

$$x_a[n] \triangleq x[n] + i\mathcal{H}\{x[n]\}, \tag{18}$$

where $x_a[n]$ is the discrete analytic signal obtained from $x[n]$. The absolute value, $|x_a[n]|^2$, is the discrete *Hilbert envelope* of $x[n]$.

3.3.3 Amplitude and Frequency Modulations (AM-FM)

The description in terms of IF and a slowly varying envelope can be generalized to a multiplicative AM-FM model of the signal. AM-FM model provides a distinctly non-Fourier description of a sound signal. Mathematically, the model can be specified as the real part of $z(t)$ in (8), i.e.,

$$x(t) = \mathcal{R}[z(t)] = A(t)\cos\underbrace{\left(\omega_c t + \omega_m \int_0^t q(\tau) + \phi\right)}_{\theta_i(t)}, \tag{19}$$

where $A(t)$ is the AM component, ω_c is the carrier/center frequency of the sinusoid, ϕ is the initial phase shift while ω_m is the maximum frequency deviation from ω_c. $|q(t)| \leq 1$ is the normalized frequency modulating signal.

The Discrete Energy Separator Algorithm (DESA) estimates the discrete IF and amplitude envelope using the TKEO. In [85], we parameterize this discrete IF and amplitude

envelope using an ARMA model and report results on classifying eight relatively similar environmental sounds from the RWCP database [86]. Compared to the MFCC based baseline system, the proposed system showed improved performance on this task.

3.3.4 TF distributions

Equation (12) enables the interpretation of frequency as the mean of a TF distribution. The ad-hoc manner with which STFT treats time dimension is in contrast to TF distributions, where time is an integral part of the spectrum. A broad category of such distributions is defined by the bilinear distributions of the *Cohen class* [87]. The WV distribution is one of the earliest and most used TF distributions. For an analytic signal, $x_h(t)$, it is defined by the following mathematical relation

$$P(\omega, t) = \frac{1}{2\pi} \int e^{-j\tau\omega} x_h^*(t - \frac{1}{2}\tau) x_h(t + \frac{1}{2}\tau) d\tau. \tag{20}$$

From the definition it is clear that, unlike the analysis techniques we have discussed so far, it is a bilinear. Moreover it is the Fourier transform of the central covariance function of the signal. WV has an advantage over the STFT in that it does not require any windows and produces superior representations in a time-frequency plot at the cost of linearity.

3.4 Environmental Sounds Analysis Hierarchy

As we have noted before, the diversity of environmental sounds does not allow a uniform application of STFA, if one is interested in a perceptually relevant, signal-centric analysis. While the non-STFA descriptors, with the help of respective estimators, bring out the non-Fourier facets of the signals, they are not immune from a similar criticism as they also make certain assumptions and bring their own models to bear upon the sounds.

Thus, in order to systematize the process of matching a signal to its analysis method in a perceptually meaningful manner, we propose to organize the sounds into a *perceptuo-analytic* hierarchy guided by two principles: a) The analysis method should be suited to the physical properties of the signal; b) The analysis results should be correlated with the

perceptual sensation of the sound. Our use of perception is rather imprecise at this point but its meaning will become clear in due course. We hypothesize that when the properties of the sound as inferred from its temporal and Fourier spectral analyses, in the sense that these properties are discernible by the ear, are in conflict with each other or if the temporal or Fourier spectral analysis is ambiguous due to non-linearities within the auditory system, then the monopoly of either model becomes tenuous. Based on the hypothesis, we outline the partly overlapping broad classes of signals for which this mismatch would occur:

Non-WSS Signals (NSS): We first identify a broad class of non-WSS signals, which range from the transients to slowly-varying signals. As explained before, for signals of this type, STFA does not provide physically meaningful estimates of the PSD and it becomes useful to concentrate on non-Fourier attributes of the signal. In such signals, the *unit* of perception is correlated with the interval of non-WSS. The *quasi*-WSS signals with non-WSS intervals between 20-60 ms are not included in this category as STFA is applicable to them (cf. Section 3.2.2). A notable subclass of NSS is the *obviously oscillatory* signal which may or may not be strictly periodic, but nevertheless can be best viewed in terms of temporal variation of their spectral attributes, e.g., linear chirp and almost periodic signals [88]. The spectrum in such signals is better viewed as a temporal property rather than the orthogonality of time and frequency as implied by FT.

Non-linear Sounds (NLS): The perception of sounds in this class is better explained if the auditory systems is modeled as a non-linear system. These can be WSS sounds but the PSD as estimated from STFA does not conform to the perceptual sensation. The Auditory system exhibits various non-linearities for a range of sounds [42]. A few of them, such as the logarithmic loudness response and masking, are already part of the traditional power spectrum model but most are not. Sometimes the percepts due to non-linearities are mischaracterized as illusions, e.g., the missing fundamental, beats, and many others [71]. To fully characterize the signals which exploit the non-linear behavior, we need a sophisticated model of the Ear. However, such a model does not exist. Hence, we adopt a time-domain

approach for analysis because it allows a general modeling of non-linearities in the absence of particular information about nature of the non-linearity.

Note that the classes that we define above are not mutually exclusive and there can be considerable overlap, e.g., non-stationary signals may also evoke a non-linear response from the auditory system. In fact, if a sound belongs to more than one class then that brings out multiple facets of the signal. Collectively, these classes form the *Non-Fourier Group* for sound analysis. We begin with the estimation of non-WSS and its duration, as both are integral to the applicability of STFA to a sound signal.

3.4.1 Estimation of Non-WSS and its Duration

There are many ways to test for the WSS (or non-WSS) of a time-varying signal [89, 90, 91, 92]. In the following, we discuss the *surrogate* method by Borgnat *et al.* [93]. An advantage of this method is that it allows the time interval of non-WSS to be inferred, in addition to a measure of the non-WSS.

Let T_o be the global observation interval of a signal $x(t)$. Its *surrogate*, $x_s(t)$, is defined to have the same Fourier spectrum as the signal, but the Fourier phase is replaced by a random variable uniformly distributed over $[-\pi, \pi]$. Fourier phase governs the temporal non-WSS structure of the signal and replacing it with a random variable generates a WSS version of $x(t)$ [93]. $\{x_s^j(t)\}_{j=1}^J$, where J is the number of surrogates in the ensemble generated by multiple realizations of the uniform random variable. Let $S_x(\omega, t)$ be TF distribution for $x(t)$. For a WSS signal, the marginal or global spectrum over T_O, $S_x(\omega) = \int_{T_O} S_x(\omega, t)dt$, should be close to $S_x(\omega, t)$, the local spectrum, under some distance measure. We can define the distance between two spectra as $\{c_n^{(x)} = D(S_x(\omega, t_n), S_x(\omega))\}_{n=1}^N$, where $D(.,.)$ is the distance function and $c_n^{(x)}$ is calculated at discrete time instances t_n. We define the effective test statistic as $\Theta_1 = \text{var}(c_n^{(x)})$ and the null hypothesis of WSS can be stated with the statistic $\{\Theta_0(j) = \text{var}(c_n^{x_s^j})\}_{j=1}^J$, where the variance is estimated over n. The

empirical distribution of $\Theta_0(j)$ is best modeled as a gamma distribution, whose parameters can be estimated by fitting $\Theta_0(j)$. From that, it is straightforward to estimate α for a particular statistical significance level for the following one-sided test

$$\begin{cases} \Theta_1 > \alpha, & \text{Non-WSS}, \\ \Theta_1 < \alpha, & \text{WSS}. \end{cases} \tag{21}$$

In our implementation, we aim for a 95% confidence interval with $J = 50$ and estimate α accordingly for the above test. We use the Wigner-Ville (WV) distribution with Hermite functions as short-time windows of duration, T, to estimate the local and global spectrum. We chose $D(,.,)$ as a scaled product of Kullback-Leibler divergence and log-spectral distance, as suggested in [93]. The degree of non-WSS, DOS, is then a ratio between the test statistic Θ_1 and the mean value of its surrogate counterparts, $\Theta_0(j)$, i.e., $\text{DOS} \triangleq \sqrt{\frac{\Theta_1}{\frac{1}{J}\sum_{j=1}^{J}\Theta_0(j)}}$. For a WSS signal, DOS is expected to take a value close to unity. Given T_O, the DOS is a function of T and repeating the test with different values for T, we introduce a typical time-scale of non-WSS, SNS, defined as

$$\text{SNS} \triangleq \underset{0<T<T_O}{\arg\max}\{\text{DOS}(T)\}. \tag{22}$$

We can also introduce an index of non-WSS (INS) as

$$\text{INS} \triangleq \max_{0<T<T_O}\{\text{DOS}(T)\}. \tag{23}$$

Figure 2 shows the DOS as a function of T for three different sounds. As expected, DOS is close to 1 for the beats signal, which is WSS by construction, while the walking and running sounds show a high INS with the former having a high SNS than the latter. The high INS indicates a non-WSS sound which should be least amenable to STFA. The large rate of repetition of the impulses in running than walking points towards an analytically grounded perceptual unit of signal analysis for environmental sounds.

3.4.2 Regularity of Envelope

If the signal exhibits a high degree of WSS, indicated with a high INS value, we calculate the dissimilarity between smoothed envelope and compare its TKEO energy with the

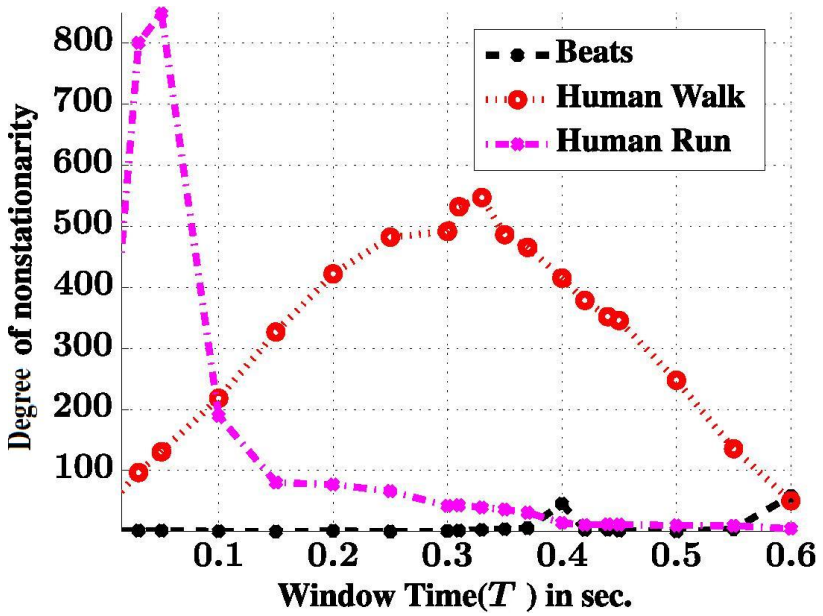


Figure 2: DOS curve for three sounds. 'Human Walk' and 'Human Run' are the sounds of the footsteps as the person walks and runs, respectively. The 'Beats' signal is shown in Figure 1

TKEO energy of the original signal as

$$\Delta E = D_{JS}\left(\left\{\psi[\mathcal{L}_c[\text{env}_h(t)]]\right\}_P \middle\| \left\{\psi[\text{env}_h(t)]\right\}_P\right), \tag{24}$$

where $\text{env}_h(t)$ is instantaneous Hilbert envelope (cf. Section 3.3.2), $\{.\}_P$ indicates that a histogram of the time series was normalized to generate a probability density function (PDF), $\mathcal{L}_c$ is the low-pass filter at c Hz with $c = 20$ Hz. $D_{JS}(p_1\|p_2)$ is the Jensen-Shannon divergence (JSD) [94] between the two PDFs, $p_1(x)$ and $p_2(x)$, and measures the statistical similarity between them. $D_{JS}(p_1\|p_2) = 0$ when $p_1 = p_2$. It is bounded, i.e., $D_{JS}(p_1\|p_2) \leq \ln(2)\,\forall p_1, p_2$, where ln(.) is the natural logarithm, and symmetric, i.e., $D_{JS}(p_1\|p_2) = D_{JS}(p_2\|p_1)$. It is not a metric as it does not satisfy the triangle inequality. If we define a divergence $D(.\|.)$ between two PDFs $p_1(x)$ and $p_2(x)$ as,

$$D(p_1\|p_2) = \int_X p_1(x) \ln \frac{p_1(x)}{\frac{1}{2}p_1(x) + \frac{1}{2}p_2(x)} dx, \tag{25}$$

then JSD is defined as,

$$D_{JS}(p_1\|p_2) = \frac{1}{2}D(p_1\|p_2) + \frac{1}{2}D(p_2\|p_1). \tag{26}$$

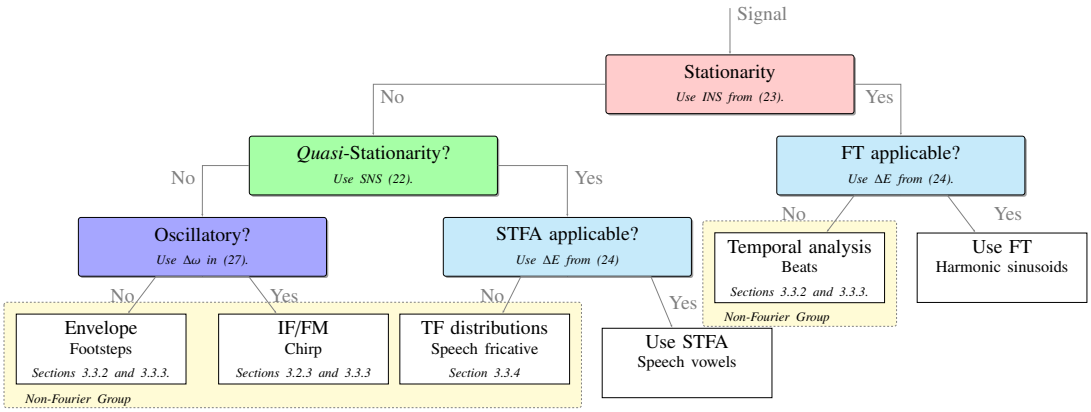


Figure 3: The perceptuo-analytic analysis hierarchy for environmental sounds.

For signals with below hearing-frequency beating, with missing fundamentals perceptions, and similar so-called illusions, ΔE will have a high value. The presence of a strong energy at low (<20 Hz) frequencies within the envelope indicates a sound that is not characterized by the Fourier analysis and thus a temporal envelope analysis, using the Hilbert or DESA envelope, is more suitable. Similarly, for *quasi*-WSS sounds with a dominant envelope, as determined by ΔE, and intervals of non-WSS greater than 60-80 ms, an analysis with a TF distribution (cf. Section 3.3.4) is appropriate as it does not rely on short-time frames to calculate the signal spectrum.

3.4.3 Divergence between Fourier and Instantaneous Frequency

Linear chirp is an example of a non-WSS oscillatory signal that can be detected by observing the difference between IF and the Fourier frequency. $\Delta\omega_T$ is a measurement of this divergence and calculated as the JSD

$$\Delta\omega_T = D_{JS}\left(\left\{\mathcal{L}_{c'}[\omega_i(t)]\right\}_P \middle\| \left\{\bar{\omega}_T(t)\right\}_P\right) \tag{27}$$

where $\bar{\omega}_T(t)$ (cf. (12)), also known as the spectral centroid, is a measure of Fourier frequency. $\bar{\omega}_T(t)$ is assumed to be constant within the short-time interval, T. $\omega_i(t)$ is the IF defined in Section 3.3.2 and $\mathcal{L}_{c'}$ is the low-pass filter at c' Hz that is necessary to smooth out the abrupt changes in the phase function which are accentuated with the time-derivate used in the definition of IF. We set $c' = 40$ Hz.

In case of presence of strong spectral signature the IF can be estimated from the DESA algorithm or the HT to analyze the signal. A non-WSS and non-oscillatory signal can only be analyzed with its temporal envelope. Our inclusion of perception should not be taken to mean that we advocate a purely perceptual analysis of the environmental sounds. Instead, we use physically meaningful time scales of perceptual significance—correlated with environmental sound analogues of speech 'phonemes' and 'syllables'—which can be used to build higher order perceptual units to a first-order approximation.

3.4.4 Discussion

In Figure 3, we show a stylized preliminary organization of the sounds based on the above classes. It consists of decision and analysis units. Figure 3 also outlines the analysis method blocks that are used to analyze the sounds that exhibit the relevant characteristics along with the prototypical signals for the specific block. However, these analysis methods are not exhaustive and we have only compared the results of one with the other for the case of footsteps in acoustic gaits in Chapter 4. The decision units do not make binary decisions; rather they indicate the degree of presence of a particular characteristic in the sound signal.

3.5 The Sound Profile

We have outlined ways in which one can measure the amenability of a signal to STFA and estimate the degree to which the sounds diverge from the STFA framework. Together, the four dimensions—INS, SNS, $\Delta\omega_T$, and ΔE— quantify this divergence and constitute the *sound profile*. The sound profile provides a new structure for a systematic study of these sounds in the absence of lexical structures. For example, the SNS can provide the unit of analysis, much like the phonemes do for speech, or the $\Delta\omega_T$ can be helpful in determining the type of sound analogous to the tonality of certain languages.

The dimensions of the sound profile, which is the first part of the hierarchy in Figure 3, provides guidance towards a suitable analysis method, which is the second part of hierarchy in Figure 3. The signal profile thus provides a pre-analysis of the sound, i.e., before we

spectrally analyze the sound, the profile of the sound will tell us whether a spectral STFA will even be useful for the given sound and further suggest an analysis method which will be relevant and useful.

Essentially, the sound profile organizes the sounds with similar signal characteristics into soft analysis categories. Then, the sounds grouped together in a particular analysis category to be processed by similar and suitable analysis method, e.g., non-stationary sounds with high values of INS and large durations of non-stationarity are better analyzed with the HT, a non-STFA method.

The sound profile dimensions are global and as such are not equivalent to the *features* such as MFCCs or PLPs. In the proposed framework, the features which enable the various experimental classification or detection tasks, are extracted from the representation generated by the proper analysis method. This analysis forms the second component of the hierarchy in Figure 3. We demonstrate the process of analysis of a particular type of regular impact sounds and feature design in Chapter 4.

3.6 Experiments on Sound Profile

In this section, we build sound profile of the sounds from two sets of environmental sounds. The first is a relatively diverse collection of single instances of environmental sounds available from a different data sources; the second are a set of impact sounds generated when objects from different materials collide. The experimental results serve as a preliminary validation of the concept of the *sound profile* as a general structure which should guide further sound analysis. This forms the basis of our argument that the characteristics of the sound, reflected in the sound profile, should govern the sound's analysis.

3.6.1 Environmental Sounds, their Sound Profile, and the Analysis Hierarchy

We now show path of environmental sounds through the hierarchy of Figure 3. The description of the audio events which produce the environmental sounds, their labels, and the results are shown in Table 2. The sounds are collected from the FBK [73], UPC [72],

RWCP [86], TIMIT, and BBC sound effects [95] databases. In addition to the prototypical sounds, some of which we have already mentioned, we include more than 30 widely understood environmental sounds with different sources such as animal and human vocalizations, different solid objects, etc. The sounds also vary in their temporal and spectral characteristics such as transience and harmony.

From Table 2, we first describe the prototypical signals—artificial and natural—that served as design primitives for the analysis hierarchy and now validate the initial hypotheses. The 'beats' sound is WSS and has a high relative value for ΔE, which would suggest a non-Fourier temporal envelope analysis. The 'harmonic sound', which is summation of a 200 Hz sinusoid and its 4 even harmonics, is also WSS with a relatively low value for ΔE which indicates primarily FT mode of analysis. The 'chirp' is highly non-WSS but oscillatory, as indicated by the relatively low $\Delta\omega$ value for a non-WSS sound, and thus more suitable to be represented in terms of IF rather than the classical Fourier frequency, while the 'footstep' is non-WSS but with a very low Fourier spectral content, as indicated by a very high $\Delta\omega$ value. The speech phonemes /sh/ and /ao/ are *quasi*-WSS, the former less so than the latter, with low INS scores. The ΔE is much less for /ao/ than /sh/, indicating that a non-Fourier spectral analysis is more suited for /sh/ than /ao/.

Figures 4 and 5 shows these basic and real-world environmental sounds along the sound profile plane. In general, the sounds close to the origin are most amenable to STFA. The regular harmonic sounds, 'phone ringing', 'cars honking', and 'bird calling' are such sounds. In Figure 4, these have low values of $\Delta\omega$, as expected, with similar SNS but increasing INS from 'phone ringing' to 'bird calling', indicating a regular harmonic structure with different scales of repetition of that structure. The low value of ΔE for 'phone ringing' makes it ideal for STFA. 'Door close', 'key jingle', and 'cymbals hitting' have similar $\Delta\omega$ and INS but large SNS, which indicates an irregular harmonic structure of these sounds which is consistent with their perceptual sensation.

The impact sounds such as 'clock ticking', 'clapping', 'keyboard typing', 'drumming',

Table 2: The table of environmental sounds with the values of INS, SNS, $\Delta\omega_T$, and ΔE. $T = 10, 20, 30, \ldots, 450$ ms.

Class Descriptions	Class Label	INS	SNS(s)	$\Delta\omega_{T\|T=[10:10:450]ms}$	ΔE
Airplane Flying	AIR	3553.6	0.30	0.62	0.02
Baby Crying	BAB	1060.8	0.25	0.36	0.03
Beats	BET	45.1	0.79	0.18	0.48
Bells chiming	BLS	266.9	0.12	0.24	0.01
Bowling	BOW	188.2	0.62	0.50	0.00
Bird calling	BRD	973.5	0.27	0.23	0.04
Bubbling	BUB	26.9	0.07	0.35	0.02
Cat meowing	CAT	555.8	0.38	0.18	0.01
Chirp	CHI	1467.4	0.33	0.42	0.00
Chair Moving	CHM	117.5	0.40	0.52	0.02
Chimp calling	CHP	154.6	0.38	0.40	0.02
Clock Ticking	CLK	225.6	0.15	0.32	0.04
Clapping	CLP	15.9	0.58	0.57	0.01
Cough	COU	102.5	0.24	0.45	0.01
Cup clink	CPC	75.7	0.03	0.37	0.01
Cars honking	CRH	450.9	0.24	0.12	0.15
Car Starting	CRS	393.8	0.54	0.43	0.01
Cymbals hitting	CYM	199.5	0.24	0.42	0.00
Dog Barking	DOG	59.9	0.15	0.35	0.22
Door close	DRC	42.8	0.23	0.33	0.03
Door knock	DRK	65.9	0.17	0.41	0.01
Drumming	DRM	1566.5	0.48	0.14	0.41
Door open	DRO	138.1	0.27	0.42	0.00
Electric saw cutting	ESW	40.6	0.30	0.49	0.05
Footsteps	FST	493.9	0.42	0.51	0.00
Glass breaking	GLS	385.3	0.18	0.18	0.01
Gun Shot	GUN	1029.7	0.22	0.34	0.03
Harmonic sound	HAS	1.6	0.08	0.08	0.15
Keyboard typing	KBT	91.4	0.04	0.50	0.01
Key jingle	KYJ	77.9	0.29	0.23	0.03
Laugh	LAU	130.3	0.50	0.41	0.00
Phone ringing	PHR	63.3	0.06	0.15	0.03
Paper wrapping	PPW	1152.1	0.13	0.11	0.02
Speech /ao/	SAO	16.2	0.02	0.33	0.02
Speech /sh/	SSH	68.4	0.14	0.44	0.21

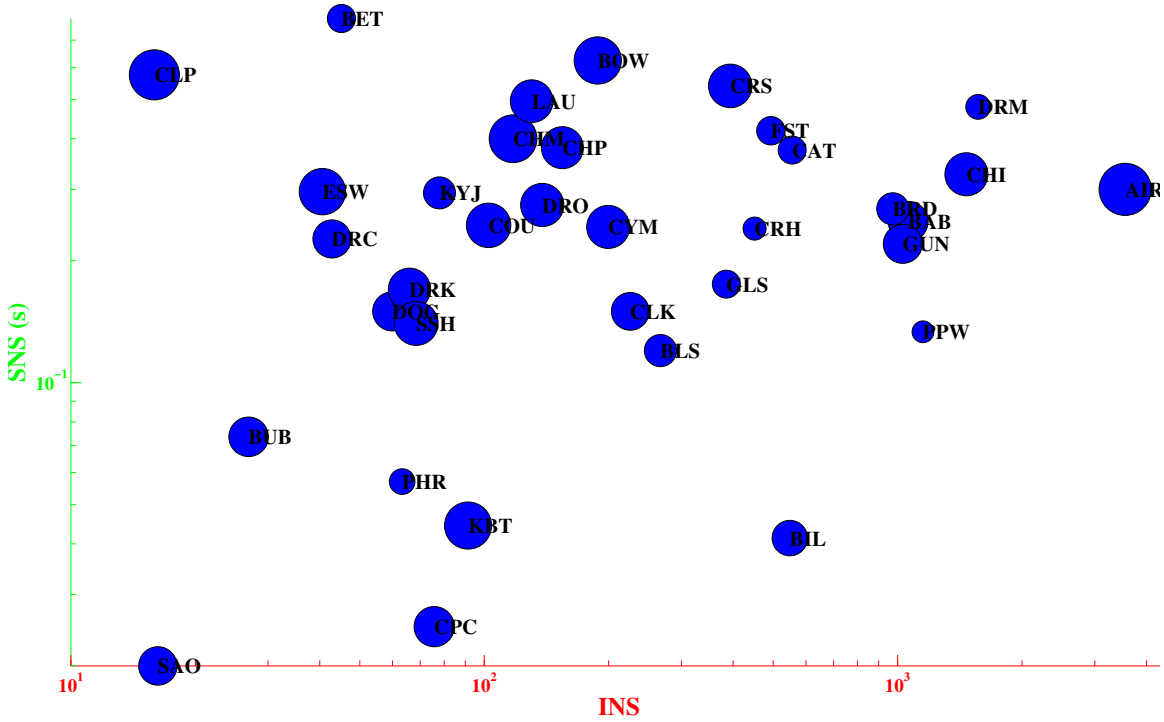


Figure 4: The environmental sounds from Table 2 in the non-Fourier dimensions. The axis colors correspond to the similarly colored blocks in Figure 3. The diameter of the bubble is proportional to $\Delta\omega_T$.

'cymbals hitting', 'bells chiming' and 'door knock' show SNS which is consistent with their repetition interval. Among them, 'cymbals hitting', 'bells chiming', and 'drumming', being sounds from musical instruments, give rise to a harmonic sound after the impact which is evident in their low values for $\Delta\omega$. 'Drumming' is notable as it has one of the lowest value for $\Delta\omega$ but highest value for ΔE which indicates the presence of a longterm envelope in addition to the short-term spectrum. Irregular and non-harmonic impact sounds—'gun shot', 'bowling', and 'glass breaking'—give a high value for SNS,INS, and $\Delta\omega$. The non-speech sounds from the human vocal tract: 'laugh', 'cough' and 'baby crying', have pronounced non-STFA structure which is indicated with comparatively large values of $\Delta\omega$ (>0.35) and INS (>100). Sounds from non-human vocal tracts: 'bat meowing', 'dog barking', 'chimp calling' have comparatively similar of $\Delta\omega$ indicating largely vowel-like vocalization from these animals. The large ΔE value for 'dog barking' indicates a long-term temporal structure. 'Bubbling' and 'airplane flying' and large values of

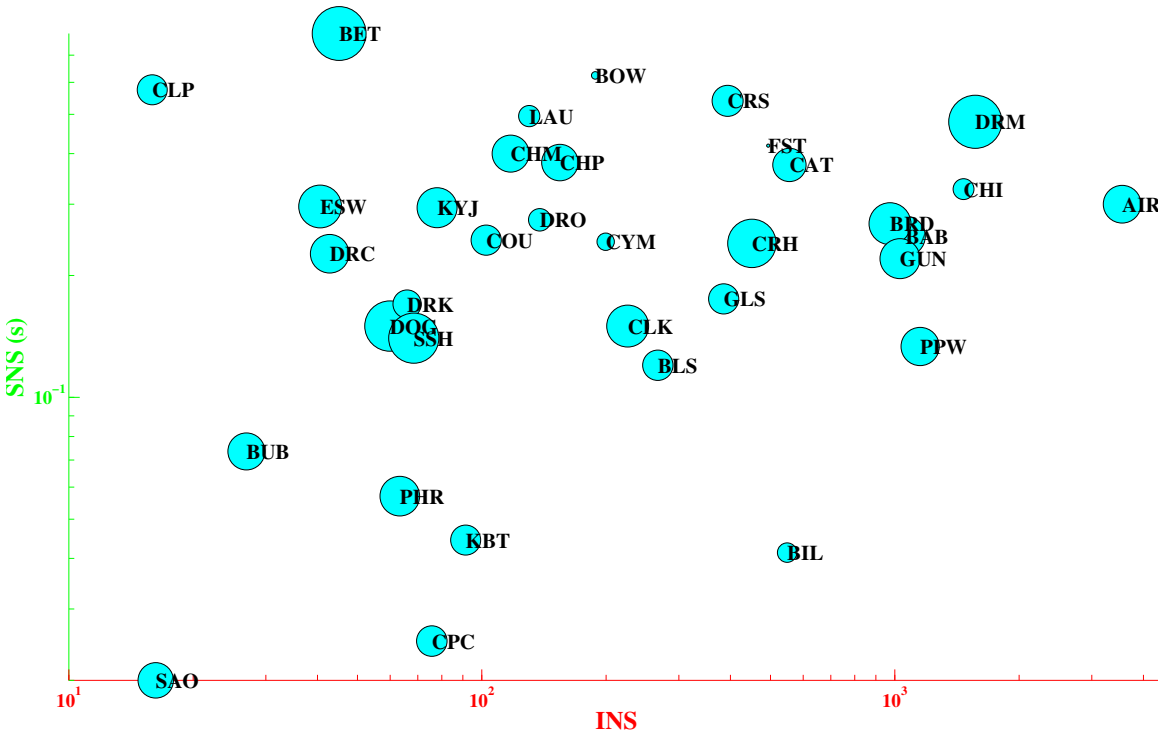


Figure 5: The environmental sounds from Table 2 in the non-Fourier dimensions. The axis colors correspond to the similarly colored blocks in Figure 3. The diameter of the bubble is proportional to ΔE.

$\Delta\omega$, indicating sounds with frequency content that is changing rapidly. In general, SNS is correlated with the interval of repetition of a temporal or spectral structure, while a high INS indicates rapid changes within that structure. However, the very low value of INS for 'bubbling' suggests a temporally stable spectral structure, which is not the case especially when compared to 'phone ringing'.

3.6.2 Sound Profile of the Sounds from the RWCP Database

At the most basic level, the profile helps organize the sounds with similar signal characteristics. This is useful in providing an initial representation but that may not be the end goal. We may wish to organize the sounds based on meaningful categories such as sounds with similar source, sounds which create a certain ambience or sounds which constitute a particular soundscape. For such goals, the profile points towards an analysis method which is tailored to a particular signal characteristic of sounds. After proper analysis we can connect the analysis with sound perception and use the information to extract meaning relevant to humans.

We now introduce and describe our results from the RWCP database [86] which show that the source characteristics embedded in sound signal can be understood in terms of the new perceptuo-analytic hierarchy. Unlike the ad-hoc collection of sounds in the previous section, this database provides a large number of instances of sounds from a sound category. We use four sound source classes from this database: Wood, Ceramic, Metal, and Plastic. The sound in each class from the database is generated by striking together two objects constructed from the respective source material. Table 3 provides the number of sounds from each class and the mean duration of the sound from each source class. These sounds were recorded in an anechoic room. The table also provides the type of objects, such as sticks, cups, bowls, etc., and the actual material of objects that were used to generate the sound for each category, e.g., the category Ceramic has 800 sounds created by collision of bottles, cups, and bowls made from glass and china ceramic (porcelain).

Figures 6 to 11 are the discrete unnormalized histograms for the parameters of the sound

Table 3: An overview of the sounds from the four sound sources available in the RWCP database.

Sound Source	Number of Sounds	Mean Duration (s)	Type
Wood	1187	0.453	Wooden sticks and boards from Teak, Magnolia, Cherry.
Ceramic	800	0.690	Bottles, cups, and bowls made from glass and China.
Metal	1000	0.652	Metal cans, bowls, boards, and boxes.
Plastic	550	0.752	Plastic cases and dice.

profile—INS, SNS, $\Delta\omega$, and ΔE—for the four sound source classes. E.g., Figure 6(a) is the histogram of the 1187 values of INS for the sounds in the Wood sound source class or Figure 8(c) is the histogram of the 1000 values of ΔE for the sounds in the Metal sound source class.

All the sounds are impact sounds from the four materials and as a result we expect them to have a high value of non-stationarity, which is borne out by Figure 6. The INS values place all the sounds from these classes in the highly non-stationary category. The average is approximately 2000 for sounds from all the classes, except for Wood which, at 3000, is higher than the others. The distribution of the values indicates the variation of the sounds within and across the database source classes. For Wood and Metal in Figures 6(a) and 6(c), the values are clustered around the mean while Ceramic and Plastic can be seen to have at least two clusters. For Plastic in Figure 6(d), the higher values clustered around 3000 are from the sounds of collision of the plastic dice while the lower clustered around 1500 are for the sounds from various plastic cases. For Ceramic in Figure 6(b), the lower clustering around 500-700 is porcelain cup clinks while a somewhat higher cluster around 1500-2000 is for glass bottles. Thus, as indicated by INS, the sounds are highly non-stationary while inter-class differences in the value of INS reflect inter-class variation in the database sounds.

In Figure 7, the scale of non-stationarity is shown. It is clustered at approximately 0.25 s for Wood and Plastic in Figure 7(a) and Figure 7(b), respectively. Ceramic again shows a bimodal behavior for SNS similar to INS in Figure 7(b), with the Ceramic porcelain cup

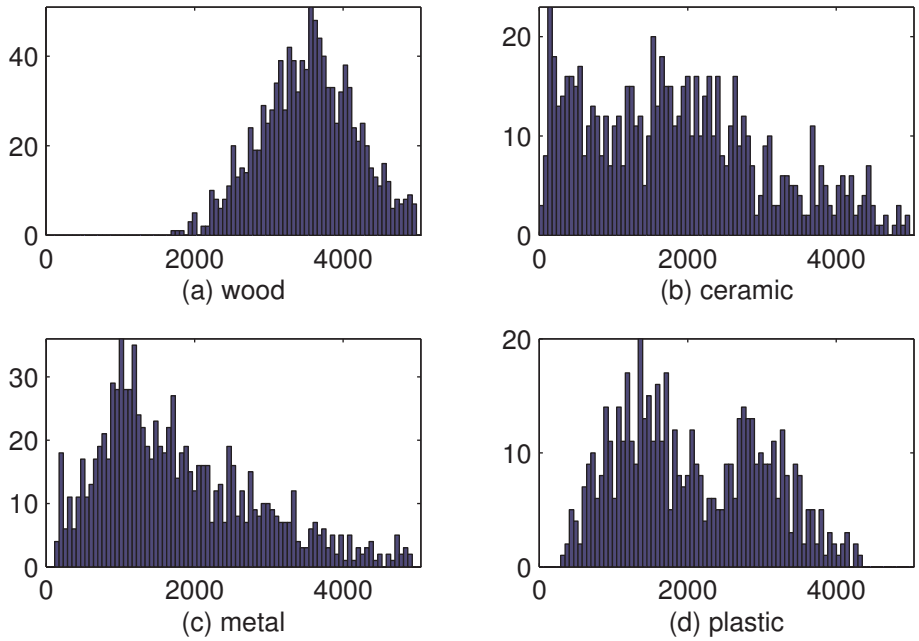


Figure 6: Histogram of Index of Nonstationarity

sound more similar to Wood and plastic at 0.2 s, while the Ceramic glass bottle sound more closer to the metal sound at 0.4 s. The high value of SNS is expected for a metallic sound as its harmonics tends to resonate longer instead of quickly dying down like the Wood sound. The variance of Metal class is also larger for SNS, reflecting the inter-class variation in the database for the sounds in this class, though it is still less than the magnitude variation for Ceramic, where it is quite large and obvious. When compared with INS, the variation in SNS is much less. In particular, the bi-modality of Plastic due to presence of sounds from two distinct subclasses is absent for SNS in Figure 7(d).

Figure 8 provides the histograms for ΔE for the four sound source classes. The values are more tightly clustered and bi-modality which we observed for SNS and INS in some classes is absent in the histograms for ΔE. Being impulsive sounds, the values are generally limited to small values as there are no slow, long term variations. The ΔE for Metal has the smallest mean which indicates it has mostly quick variations. Wood has the highest value of ΔE, which would indicate a relatively strong envelope.

In Figures 9 to 11, we plot the histograms of $\Delta\omega_T$ for three different window durations T

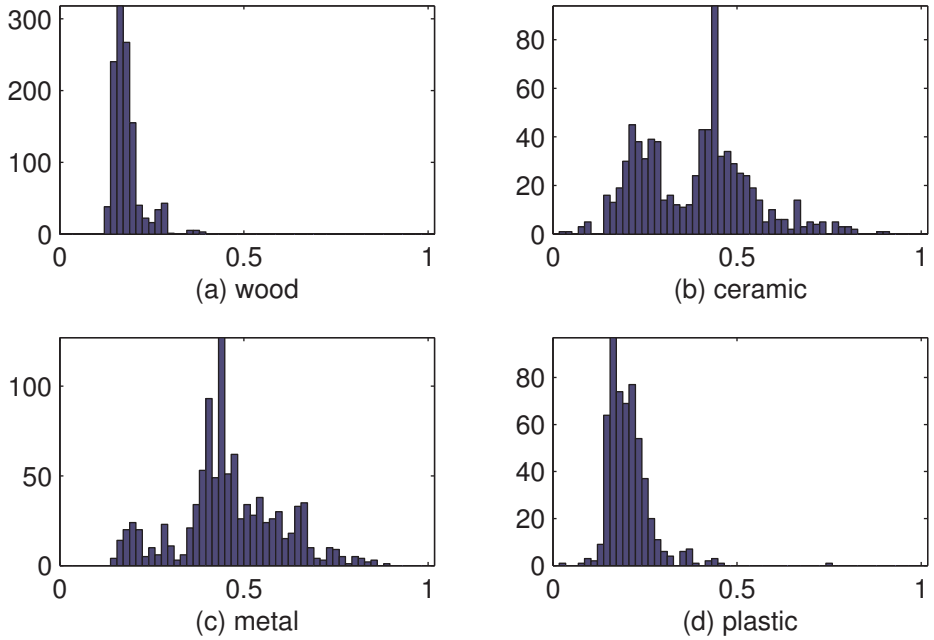


Figure 7: Histogram of Scale of Non-stationarity in seconds.

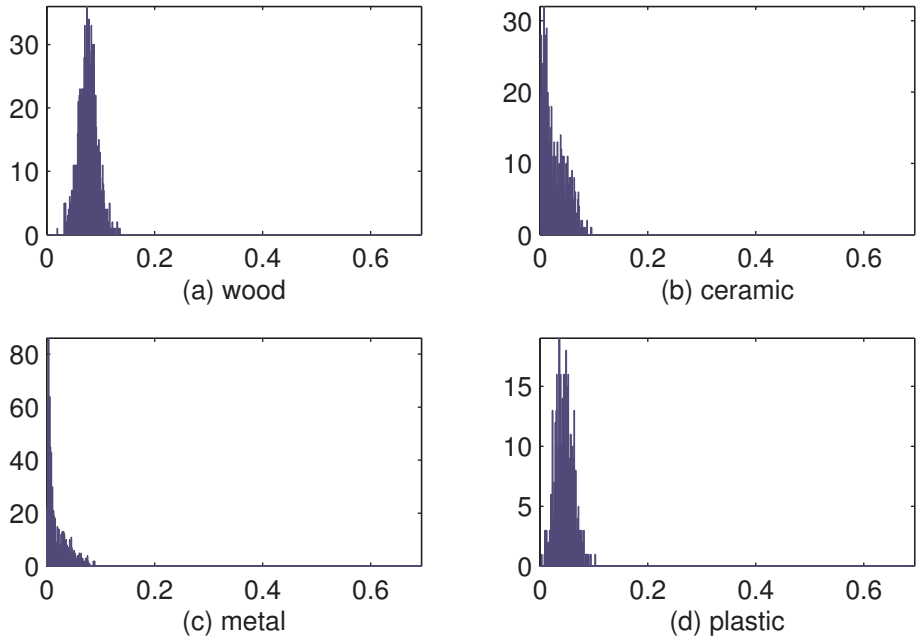


Figure 8: Histogram of ΔE for RWCP.

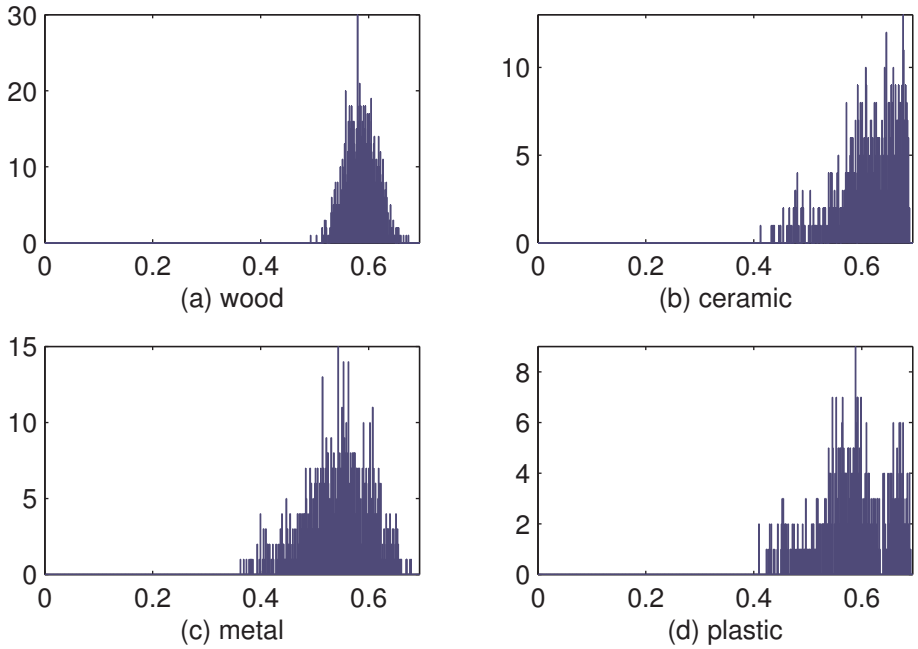


Figure 9: $\Delta\omega_T$ for T=20 ms

for calculating the spectral centroid: 20 ms, 150 ms, and 400 ms, respectively. These three values give us three different time scales to observe the behavior of Fourier frequency; with 20 ms much smaller than the SNS, 150 ms slightly smaller than the SNS, and 400 ms somewhat larger than the average SNS for the sounds. For T = 20 ms in Figure 9, the impulsive character of the sounds means that the stationary spectral content is limited and thus we observe that the difference between the Fourier frequency and instantaneous frequency is large. As expected, the class for which the difference is relatively least is Metal in Figure 9(c). Wood shows the least variation in Figure 9(a), while we can again see the bi-modality for Plastic and Ceramic in Figure 9(d) and Figure 9(b), respectively.

With relatively larger values of T in Figures 10 and 11, we notice that all classes are affected and move close to the upper bound for JSD. The least effect seems to be Metal; it retains considerable number of values less than the upper bound. This behavior is reasonable as this is the only sound class which, while being impulsive, has relatively substantial spectral content at these time-scales. Thus, in addition to the absolute value, the variation with window duration also forms an element of the sound profile.

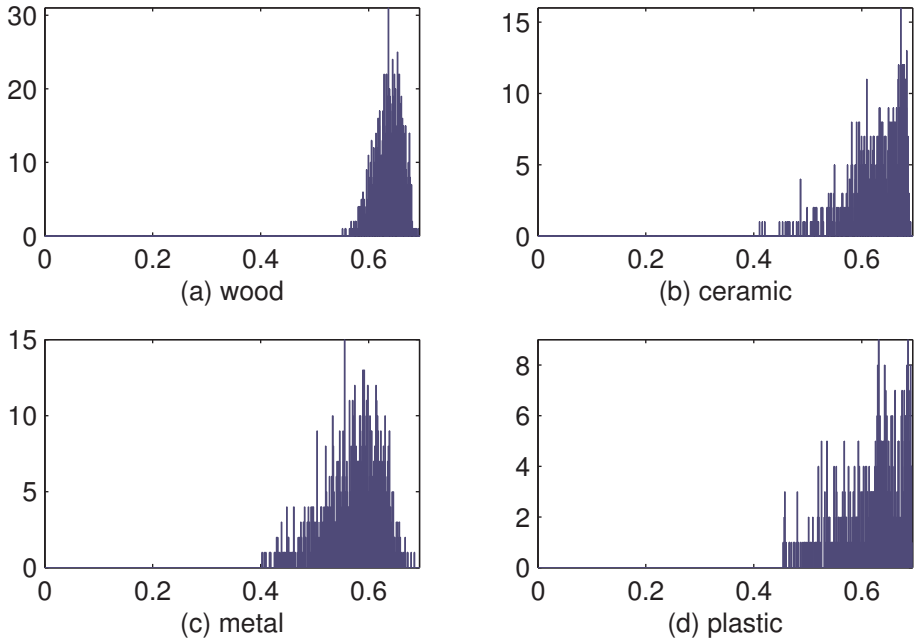


Figure 10: $\Delta\omega_T$ for T=150 ms

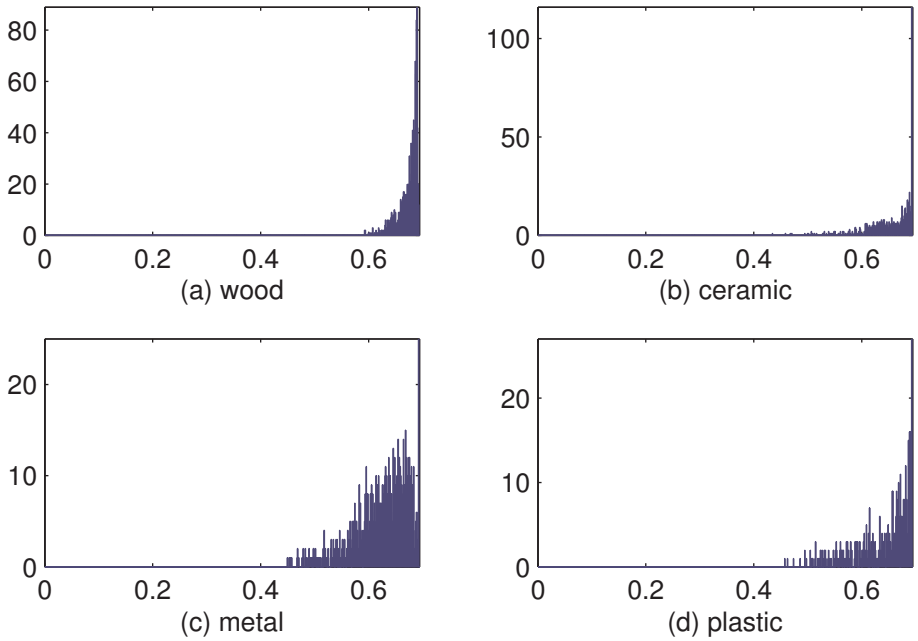


Figure 11: $\Delta\omega_T$ for T=400 ms

Table 4: The parameters are defined in Section 3.4. The values in column SNS are in seconds.

Source	ΔE	$\Delta\omega_T\|_{T=20\text{ ms}}$	$\Delta\omega_T\|_{T=150\text{ ms}}$	$\Delta\omega_T\|_{T=400\text{ ms}}$	SNS	INS
Wood	0.076	0.586	0.636	0.674	0.181	3.62e+03
Ceramic	0.027	0.613	0.625	0.656	0.387	2.07e+03
Metal	0.018	0.544	0.571	0.625	0.460	2.00e+03
Plastic	0.045	0.583	0.616	0.643	0.205	2.06e+03

Table 4 summarizes the results from Figures 6 to 11 and provides the sound profile for each source class. The results in this table are averaged over the number of sounds from each class. Metal and Ceramic have similar values of ΔE but $\Delta\omega_T$ is larger for Ceramic, indicating that Metal has more Fourier specific harmonic content. Ceramic and Plastic have similar values for $\Delta\omega_T$ but ΔE is much higher for Plastic, which indicates that the sound produced by the plastic objects has a more prominent envelope. The SNS is longest for metal as the sound from metal objects lingers the most due to metallic resonances while the wooden sound dies out relatively quickly. This makes SNS consistent with the perceived duration of these sounds.

3.6.3 Discussion of Results

We report the results on the four dimensions of the sound profile in Sections 3.6.1 and 3.6.2. The results support our basic assumptions—the sounds are diverse and need a rich enough representation which matches their diversity in terms of signal structures and information content. We observed this when different types of sounds correspond to different points on the sound profile plane, i.e., in Figures 4 and 5. We also note that interesting sounds need not be quasi-WSS, e.g., the INS of the various impact sounds that we discussed or the ambient sounds which contain information about their sources and environments. Furthermore, the sounds that are quasi-WSS do not necessarily have the stationarity spans of less than a 100 ms, as evidenced by the wide-range of SNS for different sounds in Table 2.

The results from $\Delta\omega_T$ and ΔE in Tables 2 and 4 meet our design motivations; the former a measure of WSS spectral content, while the latter a measure of presence of a dominant

slowly varying envelope—which indirectly indicates the non-linearity as the human auditory perceptual system tends to switch to non-linear modulation mechanisms to process such signals.

The sound profile dimensions are not meant to be exhaustive. Rather, they are descriptive of the sounds characteristics along certain dimensions that we identified. At the very least, they provide a much richer interpretation suited to the diversity of environmental sounds, as opposed to relying on just STFA. Some are specific enough, e.g., the INS is sensitive to the differences in the sounds from glass and porcelain, yet others are broad enough, e.g., the impact sounds in Table 4 are broadly seen to have high INS, high $\Delta\omega_T$ and low ΔE.

3.7 Summary

At the beginning of the chapter, we presented the arguments that form the basis of the thesis; environmental sounds are important, diverse, and, by implication embody myriad of signal characteristics which cannot possibly be covered by an STFA-only representation. We explored the elements of this representation, the sound descriptors, and pointed out issues which arise when these elements are viewed as arising out of an STFA-only representation. We also suggested non-STFA methods for generating these elements.

To organize the sounds for proper analysis, we proposed a sound profile which groups together sounds with certain similar characteristics and, by hypothesis, the sounds with similar sound profile are expected to be most amenable to similar and more appropriate analysis methods for proper representation. Thus, the sound profile guided towards proper analysis methods, some of which we discussed. This abstract process was summarized with the analysis hierarchy in Figure 3, which outlines the process of generation of a sound profile and lists proposed methods for further analysis. We experimentally validated the first component of the analysis hierarchy, the sound profile, on sounds from our own ad-hoc collection and the RWCP database.

In the next chapter, we will use the information form the sound profile and proceed towards the next part of the analysis hierarchy—the use of sound-appropriate analysis methods. We will use a particular set of environmental sounds, the sounds generated by human footsteps. The use of the proper analysis methods guided by the sound profile will provide a more meaningful sound representation in a non-Fourier domain. The non-STFA representation will be used to generate the *features* suitable for a detection task and further extended to infer information and attempt tasks related to the source of these sounds—the humans and their gait.

CHAPTER 4

ACOUSTIC GAITS

4.1 Introduction

From the CLEAR evaluations [56] and our own investigations in Table 1, it is clear that 'footsteps' is one of the challenging environmental sounds that is difficult to detect and is frequently confused with other sounds, such as 'applause' and 'door close'. We now use the elements from the newly formulated framework to acquire and analyze the 'footstep' sounds as a first systematic test of the framework's effectiveness. As we will show, the application of this framework allows us to improve the traditional detection of footstep sound and uncovers this signal as a source of information which can potentially characterize the gait of a person—not unlike the speech sound which embodies the lexical structures. This chapter draws on our previously published work [96, 74, 97].

4.2 The generation of the footsteps sound

We show the major anatomical components of a human foot—the heel, the metatarsals-phalanges (MTP), and the toes—and part of a typical human gait cycle in Figure 12. A normal gait cycle begins when the heel of one foot makes contact with the ground and ends when that same foot touches the ground again. Each cycle consists of two phases: a stance phase (60% of a gait cycle) in which a foot is in contact with the ground as shown in the lower part of Figure 12, and a swing phase (40% of a gait cycle) in which the same foot swings forward until the next contact with the ground. We only show the stance phase and omit the swing phase in Figure 12 as the foot generates no sound during the swing phase, though the interaction of clothes with the moving body parts may produce sounds which we ignore as noise, as in our current recording setup the noise level due to frictions from clothing or others is insignificantly low relative to footsteps. This is normally true unless in inordinary situations which are beyond the present scope of investigation.

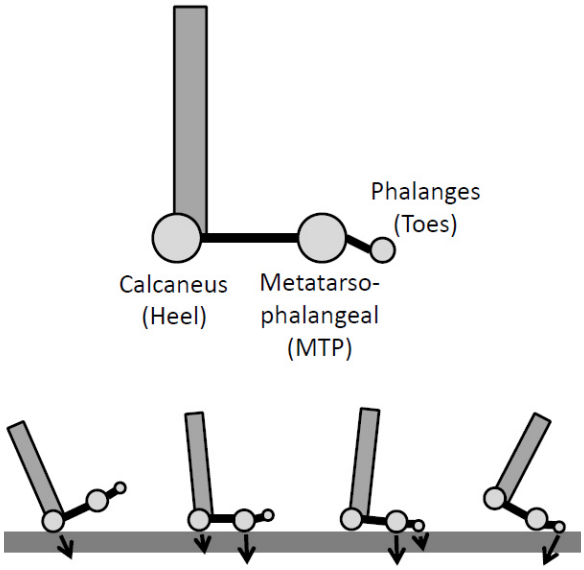


Figure 12: The schematic of a foot identifying the parts of the acoustically relevant foot structure (Top) and their interaction with the surface as the stance phase proceeds (Bottom). The downwards arrows highlight the parts of the foot structure that are most acoustically relevant at various instances of the stance phase. Adapted from [97].

Figure 13 displays the sound waveform recorded from a series of footsteps. As the foot makes initial contact with the ground, in Figure 12 made with the heel, the audio waveform in Figure 13 shows a significant increase in amplitude. We notice another increase, though not as prominent, when the hump at MTP joint makes contact with the floor. There may also be a third notable burst of acoustic activity, depending on the walking style and the footwear of the person, when the MTP joint structure lifts off from the surface. The rest of the cycle ensues when the next foot makes contact.

In general, each footstep creates a burst of acoustic activity, which consists of multiple sub-bursts followed by a gradual decay. In this interaction, the anatomical structure of the foot as well as the person's walking behavior (habitual under normal conditions or adaptive when the person's body conditions are temporarily different from normal) affects the characteristics of these bursts. External factors such as footwear, noise level, and floor type may also effect the burst characteristics but their effect can be controlled for in a practical situation. This series of seemingly rough bursts, irregularly separated and interspersed with

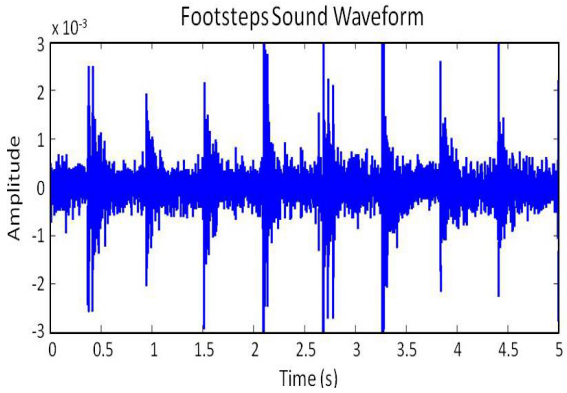


Figure 13: A time waveform of the 'footsteps' sound.

background sounds, needs to be processed for gait analysis as the key characteristics of the gait may not be readily observed in the raw waveform.

4.3 Footsteps sound in the light of the signal profile

Audio has also been used to study footstep sounds in the laboratory [98]. Barring very few exceptions [99], the general emphasis has been on using spectral features derived from STFA to characterize the footstep sound [100, 54]. However, as we have argued earlier, these conventional STFA features are designed to characterize human speech or music and are inadequate in characterizing footstep sounds [74]. Therefore, the tasks that have been investigated with footstep sounds, such as their detection among other sounds and biometrics extraction, have had limited success [74, 96].

A sound from the 'footsteps' class sound is shown in Figure 13. From Table 2, we note that this sound has a high INS when compared to the speech phonemes or the harmonic 'phone ringing', which indicates a high degree of non-WSS. We also note the SNS at 350 ms; much higher than other regular impact sounds such as 'keyboard typing' and 'door knock'. Further traversing through the analysis hierarchy in Figure 3 for a non-WSS sound, we note that this sound lacks spectral content as it has a very high value for $\Delta\omega_T$, when compared to the prototypical oscillatory signal, 'chirp'. Thus, we conclude that a combination of impulsive signal structure, long time scales of non-WSS, and a non-spectral

structure would render the conventional STFA-based features ineffective for this sound.

Next we turn to the perceptual aspect. Based on the observations of this class from the UPC and FBK databases [72, 73], the perceptually identifying features include a distinctive asymmetric peak shape due to the sudden rise, because of the contact of the foot with the ground, and a gradual fall, mainly because of the room impulse response. Similarly, the minimum interval between two successive peaks is perceptually important. For almost all such sounds in these databases, this interval of repetition is between 0.4-1 s. If we change the repetition interval to less than 0.4 s, the sound is no longer perceived as human footsteps. Moreover, if the rate is more than 1 s, then it is perceived as disjointed strikes. Thus, the time scales of perceptual significance can be up to 1 s, which is of the same order of magnitude as the time scale of non-WSS, as shown in Figure 2. Here, we notice the connection between the two time-scales—one derived from perception and the other from measurements of non-WSS, mediated by the Fourier phase.

Based on the analysis hierarchy, perceptual observations, and the apparent consistency between the two, we can infer that this sound belongs to the NSS class and that the temporal structure is more important than the short-time spectrum. Moreover, in the absence of a semantic hierarchical structure, bridging the gap between the long term repetition and very short duration impulses is problematic in the Fourier domain. Thus, it is reasonable to analyze this signal in the time domain with its envelope, as suggested in Figure 3.

In the following section, we show that this analysis can be used to improve the various tasks that have been attempted for footsteps sound—detection and biometric extraction. This places the analysis of this sound on a more rigorous footing which enables us to attempt a new task for 'footsteps' sound—the gait analysis.

4.4 Footsteps sound detection

Within the signal processing literature, the 'footsteps' sound detection [74] is treated as part of the broader acoustic event detection (AED) problem. In the AED task, the goal is

to process the acoustic signals collected by a set of distant microphones and convert them into symbolic descriptions corresponding to a listener's perception of the different sound events that are present in the signals and their sources.

In Section 3.1.1, we demonstrated the of challenge of AED in mismatched conditions and showed the substantial decrease in performance under such conditions. We now show that non-STFA features and modeling can substantially improve the performance of footstep sound in an AED task.

We have explained in previous sections that the poor performance can be because conventional STFA fails to account for the very short term and longterm perceptual cues, i.e., the irregular repetition of the self similarity of the audio signal and the sense of duration [101]. The sub-frame features are ignored in STFA and the modeling of long term behavior is left to the states in the HMM which model the temporal sequence of the events with their transitions. However, a well-known limitation of the HMM is that the underlying Markov assumption constrains the state occupancy duration to be exponentially distributed independent of the data distribution [102]. This problem is also accentuated with general audio signals as, when compared with speech, such signals do not have a hierarchical language model which could provide a high level description of the audio signal.

4.4.1 Acoustic Modeling—Signal shape and self similarity

We define shape as the evolution, the dynamics of signal parameters along time. If a signal does not change along with time, we could say that it has a constant shape. The basic question is: the evolution of *which* parameters need to be taken into account and over which time-scales to define signal shape. We address this from a perceptual point of view. Estimation of the these temporal parameters requires energy estimates that give local measurements while preserving long term trends.

In general, the idea of including temporal information into audio processing system is not new. In [103], the authors suggest using amplitude modulation features extracted in

1 sec analysis windows for robust speech detection. This approach is motivated by previous studies that indicate that this type of information is explicitly coded in the auditory cortex. The physiologically inspired analysis approach for audio classification presented in [104] is based on an advanced model of the auditory system. The authors propose modeling of the neural response over analysis window of the same 1 sec duration. Inspired by auditory scene analysis, a number of auditory features based on temporal analysis of the waveform were derived from amplitude histograms, amplitude onset maps, spectral and harmonic profiles of the waveform in 1 sec window. These have been shown to help in sound detection [105].

Taking cue from these studies, we begin our effort to model the long-term temporal behavior with the MFCCs. In Figure 14, we show a self-similarity plot of MFCC, ΔMFCC and $\Delta\Delta$MFCC which is obtained by calculating the reciprocal of the Euclidean vector distance between frames. The grayscale intensity gives the similarity between frames centered at time location on the x-axis and y-axis. We notice that there is only a single prominent line at the main diagonal indicating self-similarity at zero lag τ, which does not provide any information. Thus MFCCs do not represent the long term self-similarity of the signal or the onset of the peaks.

This is not surprising as MFCCs are based upon energy distributions in the frequency bands of a short time-interval and the variation in this distribution across frames characterizes the signal. A fundamental problem with this approach is the choice of Fourier frequency as the attribute. As we have stated, frequency in the Fourier sense has a fundamental limit on its temporal resolution and the time frequency uncertainty principle, thus signals which change rapidly suffer form inadequate representation in this domain. Thus defining the shape of the signal in terms of its Fourier frequency spectrum loses the temporal resolution.

With the failure of the frame based spectral approach, we turn to instantaneous energy

measurements to characterize its shape. The Hilbert envelope provides one such measurement of the signal energy through the Hilbert transform (cf. Section 3.3.2).

We extract the Hilbert envelope, $|x_a(n)|$, of the signal $x[n]$ [57] and then low pass filter it at 20Hz to obtain the smoothed temporal envelope, $x_{env}[n]$, of the signal. The Hilbert envelope for the 'footsteps' sound in Figure 13 is shown in Figure 22(b). The repeatability of the 'footsteps'sound can be demonstrated with autocorrelation function, $r[.,.]$, of Hilbert envelope $x_{env}[n]$ shown in, Figure 15, where the energy in each frame is normalized before calculating the autocorrelation as follows:

$$r_{M,L}^{x_{env}}[n_1, n_2] = \frac{\sum_{l=-L/2}^{L/2} x_{env}[Mn_1 - l]x_{env}[Mn_2 - l]}{\sum_{l=-L/2}^{L/2} x_{env}[Mn_1 - l] \sum_{l=-L/2}^{L/2} x_{env}[Mn_2 - l]}, \tag{28}$$

where $x[n]$ is the sampled $x(t)$, L is the frame duration and M is the frame shift. In Figure 15, the 'footsteps'demonstrate a self-similarity at 0.6 s seen clearly with a diagonal starting at this point. The contrast improvement in Figure 15 over that of Figure 14 is also obvious. We also observe that the diagonal lines are blurred at certain points.

Figure 16 shows the autocorrelation of the smoothed TKEO energy of the signal using the definition in (28), which is a marked improvement over the Hilbert envelope in Figure 15. Hilbert envelope brings out the apparent self-similarity, but its focus is only on the signal amplitude, which leads to a less sharp representation. TKEO provides the most crisp representation of this self-similarity which is robust against intra-class variations in 'footsteps'because it includes instantaneous frequency, in addition to signal amplitude, in the energy estimate —combining the local and long-term measurement in a single estimate.

4.4.1.1 Waveform self-similarity features

As per our perceptual observations, the discriminative characteristic of the 'footsteps'sound in the time domain is the irregular repetition of certain self-similarities. One can notice that other sounds also exhibit the repetitive behavior: 'applause' consists of the series of hands clapping; 'door knock' consists of the successively hitting the door with human hand, etc.

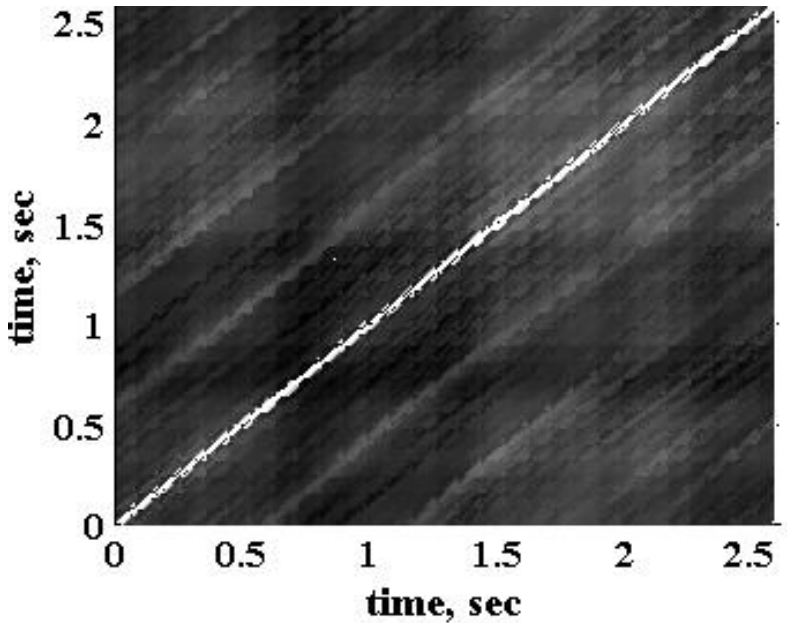


Figure 14: Autocorrelation function of a 'footsteps' sound using MFCC, ΔMFCC, and ΔΔMFCC features.

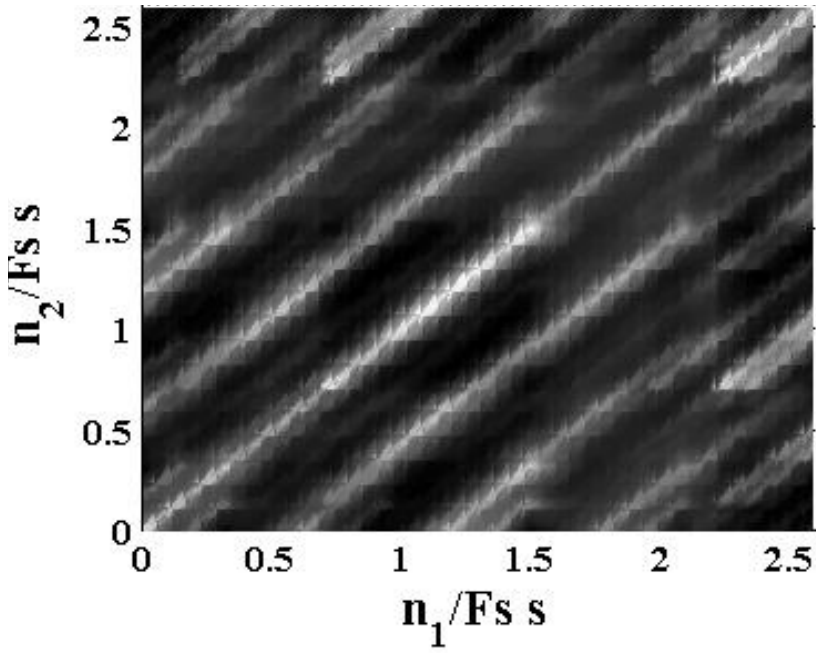


Figure 15: $r^{x_{env}}_{M,L}[n_1, n_2]$ of a 'footsteps'sound with M/F_s = 10 msec, L/F_s = 800 msec and F_s=16KHz.

The basic concept behind repetition estimation is the similarity measurement.

We define similarity in terms of autocorrelation $r^{x_{TKEO}}_{M,L}$ of energy measurement from TKEO, where M/F_s = 10 msec, L/F_s = 800 msec and F_s=16KHz is the sampling rate.

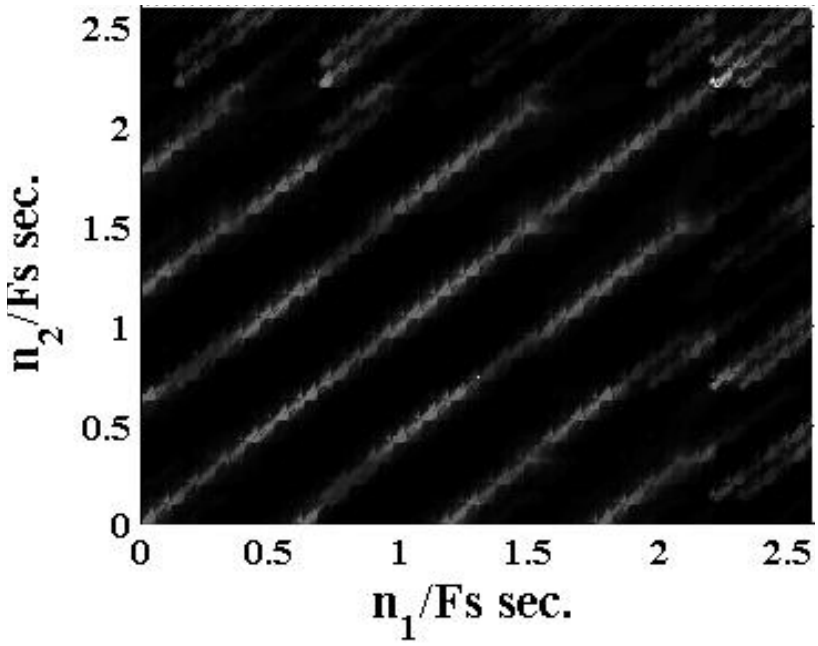


Figure 16: $r_{M,L}^{x_{TKEO}}[n_1, n_2]$ of 'footsteps'with M/F_s = 10 msec, L/F_s = 800 msec and F_s=16KHz.

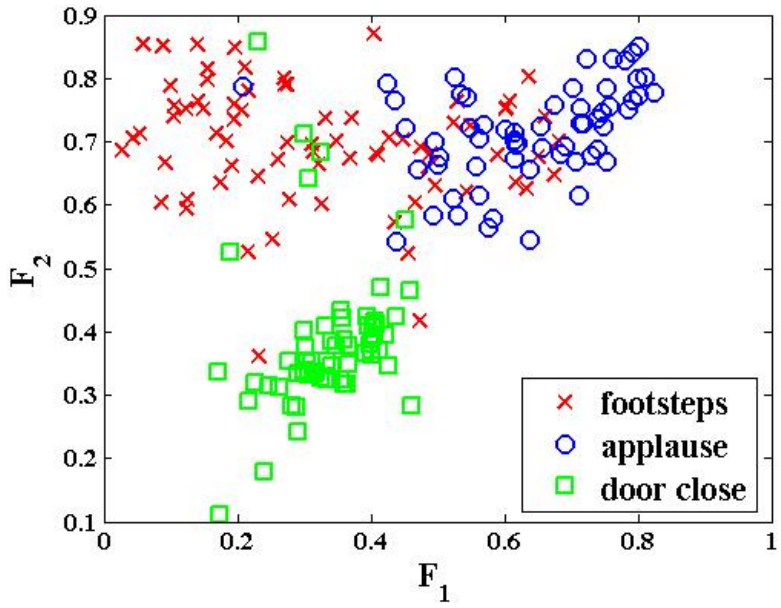


Figure 17: Distribution of different sound events with similarity features F_1 and F_2 for UPC database.

We compute two features: $F_1[n]$ and $F_2[n]$ that represent the degree of the waveform self-similarity in short and long intervals, respectively. The feature $F_1[n]$ represents the maximum of the autocorrelation function for a frame centered at n within a time lag of 0.2-0.4 sec, i.e., $F_1[n] = \max_{\tau_1} r_{M,L}^{x_{TKEO}}[n, n + F_s\tau_1]$ where $\tau_1 \in [0.2, 0.4]$ sec and $F_2[n]$ is for the same frame within a time lag of 0.4-0.9 sec, i.e., $F_2[n] = \max_{\tau_2} r_{M,L}^{x_{TKEO}}[n, n + F_s\tau_2]$ where $\tau_2 \in [0.4, 0.9]$ sec.

In Figure 17 we show the distribution of sound events in the UPC database using features F_1 and F_2. As one can expect, the 'footsteps'exhibit low short-term degree of self-similarity and high degree of long-term self-similarity. Events such as 'applause' show high degree of self-similarity in short and long terms. On the other hand, features F_1 and F_2 show low degree of self-similarity for the sound events that are not repetitive like 'door close'.

4.4.1.2 Explicit duration HMM

For modeling the temporal duration, we use HMM as a ready model for temporal duration characterization. HMMs incorporate the duration temporal structure of audio with the state transitions and have been shown to be particularly powerful in modeling sounds in which such temporal structure is important, such as speech. Ergodic (full-connected) or left-to-right topologies can be chosen for general sound events. In either case, a well-known limitation of the HMM is that the underlying Markov assumption constrains the state occupancy duration to be exponentially distributed according to $P(d) = (1 - a_{ii})a_{ii}^{d-1}$, where d is the duration, and a_{ii} is the self-transition probability.

In Figure 18 we show the waveforms of 'footstep' sound together with its Viterbi state alignment that was obtained using ergodic 2-state 'footstep' HMM applied to this sound (**1** and **2** are the two states of the HMM model). The first state corresponds to the impact sound constituting the 'footstep'and another state corresponds to the background sound between two successive peaks. We note that the duration occupancy in each of these two states has certain temporal constraints. Up to 0.2 sec, the 'footsteps'sound occupies the first state and

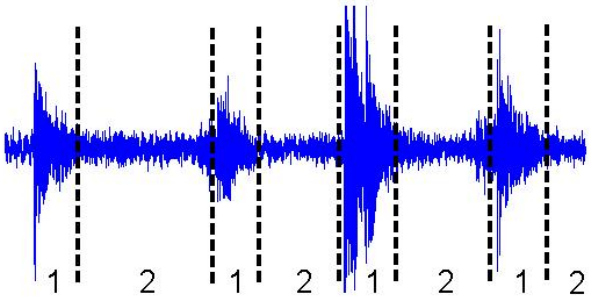


Figure 18: State alignments of ‘footsteps’sound event.

then for the interval 0.4 - 0.9 sec, it remains in the second state. We model these constraints for ‘footsteps’sound event using Ferguson’s explicit duration HMM (EDHMM) [102]. No state in EDHMM in Figure 19 has self transitions, hence direct modeling of per-state duration distributions using state transition probabilities parameters (α_i, β_i, γ_i, δ_i) becomes possible. In Figure 19 all states marked with the same number (**1** or **2**) have the same observation probability distributions but the transition probabilities between states monotonically change from left to right: α’s and γ’s decrease, while β’s and δ’s increase.

The acoustic model of ‘footsteps’ sound is created in 2 steps. On the first step a two emitting state HMM is trained using standard Baum-Welch algorithm. Then, each state is substituted by EDHMM consisting of 20 and 90 states (corresponding to the impact sounds and the sounds between them). At each state the observation probabilities of EDHMM are represented by the GMM trained on the first step but the transition probabilities are re-estimated using training data.

4.4.2 Experimental Results

The detection results for the ‘footsteps’ sound are presented in Figure 20. We used two approaches to incorporate the irregular repetition and the sense of duration analysis concepts: feature level fusion and EDHMM modeling. These approaches are compared with the baseline results for ‘footsteps’in Table 1.

In feature level fusion, the features F_1 and F_2 are appended to the initial 39 MFCCs

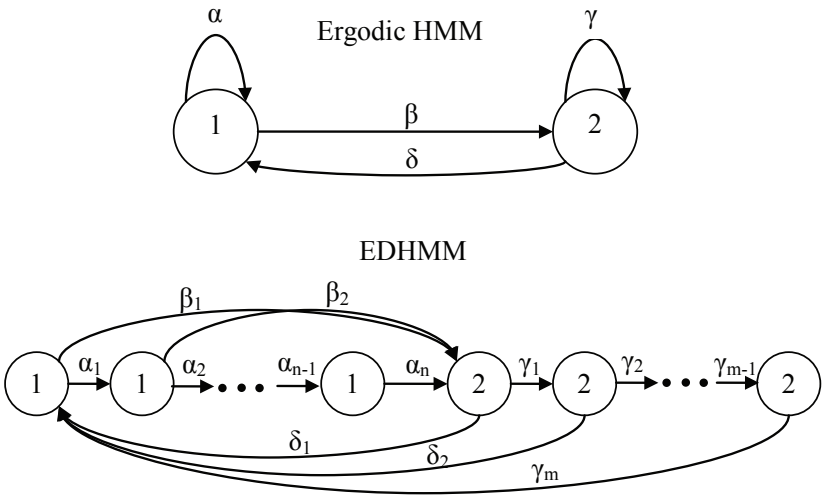


Figure 19: Ergodic and explicit duration HMM

to form the composite 41-dimensional feature vector. In EDHMM, the 'footsteps' acoustic model is built as described in Section 4.4.1.2. The estimation and inference of EDHMM parameters is performed using the forward-backward algorithm within the *allowable* duration intervals —chosen from previous perceptual observations to be 0.01-0.2 sec for peak and 0.4-0.9 sec for background sound between peaks. Note that both EDHMM and the baseline ergodic 2-state HMM have the same observation distributions in the corresponding states; the only difference between models lies in the transition probabilities between states. We achieved 14% of EER reduction in the case of feature level fusion approach and 27% of EER reduction in the case of EDHMM in cross-environmental scenario. Owing to our construction of EDHMM and the design of features, the error rate reduces significantly mainly because the temporal information, expressed in the form of long-term energy evolution, is less sensitive to the cross-environmental effects presented in different rooms.

4.5 Gait Analysis

Our interest in person-specific footstep sound derives from the observation that the footstep sound carries information about its source—the gait of the person, which is clinically used as an indicator in diagnosis of and rehabilitation from certain neurological diseases [97],

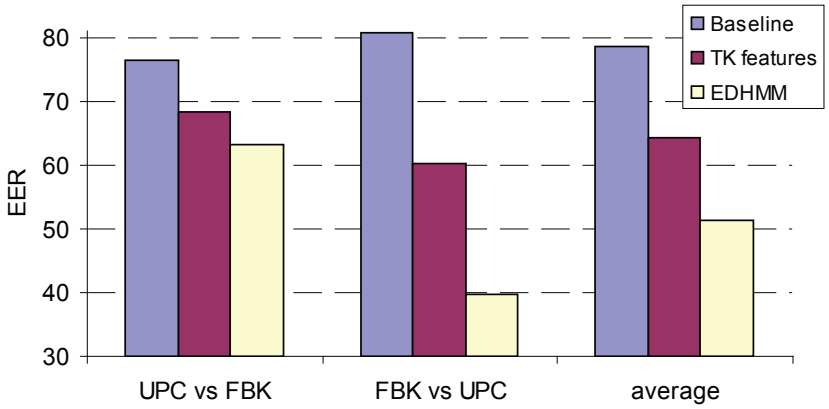


Figure 20: Comparison results for 'footsteps'sound event.

among others, or as a biometric to identify that person. We believe that the footstep sound can provide useful personal gait parameters and the new framework is ideally suited to attempt this task.

4.5.1 Acoustic Gaits Database

The environmental sound collections that we have mentioned so far—UPC, FBK, BBC, and RWCP—do not differentiate between the sound of footsteps generated by different people. In these databases designed for acoustic or sound event detection, the 'footstep' sound is treated as one of many acoustic event classes and hence the differences within a class are ignored. Hence, those are not suitable for gait analysis. Among databases recorded for gait analysis, audio modality is generally ignored. For example, under DARPA's humanID project, many research groups recorded more than a dozen databases for gait analysis but all of them were video only [106]. A recent database that specifically addresses the 'footsteps' sound for gait analysis in the context of person identification is TUM Gait from Audio, Image and Depth (GAID) database [54] which includes the audio modality, in addition to video and depth. However, the annotations for each individual 'footstep'sound are not provided and audio is noisy and not suitable for an exploratory study of this type.

We recorded and manually annotated a new database with clean audio suitable for this task. The acquisition process also entails the annotation, localization, and enhancement

of the audio event, which can be challenging with a such a large environmental sound database. For example, not withstanding the issues of linearity, the impulse response can no longer be tied to the room, but has to be estimated for each microphone with respect to the location of the audio event in the room. The type, location, and directionality of the microphone also contribute to this non-uniformity of the impulse response across the multiple audio channels. Standard localization and enhancement techniques rely on the uniformity of the room impulse response for further processing [107]. We will discuss the distributed acoustic localization problem in the next chapter.

The footsteps were recorded in a low-reverberant room ($RT_{60} \approx 200$ ms) schematically shown in Figure 21. In total, sixteen microphones were employed for recordings: eight cardioid and eight omni-directional attached in pairs to the walls of the room. The distance between the microphones in a given pair is approximately 7 cm, except for pairs (7,8) and (15,16) which are approximately 8.9 cm apart. All the microphones were placed at a height of 48.9 cm from the ground, except for the pair (9,10) which was placed at a height of 101.6 cm. Additionally, one video camera with fish-eye lens was used to facilitate the annotation process. The audio signals are recorded in pcm format at 48kHz sampling rate, F_s, with 24-bit resolution. The ratio between the peaks of 'footsteps'and the ambient noise in the recordings is around 20dB. We recorded the database in two scenarios. In Scenario-I, ten sessions corresponding to ten participating volunteers (1 female and 9 males) were recorded. Each session consisted of the subjects walking around the recording laboratory at his/her natural speed, making 15 rounds clockwise and 15 rounds anti-clockwise. In Scenario II, each participant recorded at least three sessions, each with different footwear, and walked 10 rounds clockwise and 10 anti-clockwise in each session. A total of 22 sessions were recorded for Scenario-II with seven subjects (all male). When a subject walks in the clockwise direction, the left foot will always be closer to the microphones mounted on the walls than the right foot, and vice versa for the anti-clockwise direction. Making the subjects walk both two directions and using balanced averaging allow us to

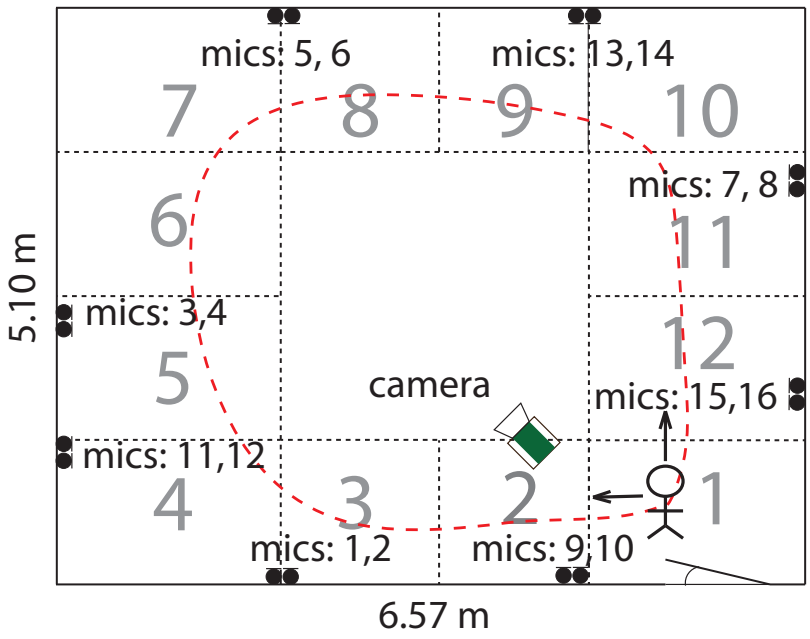


Figure 21: Recording Room Schematic

eliminate this asymmetry in distance and helps with detection and statistical analysis of the asymmetry of the footstep energy E in Appendix A.

Additionally, each recording session includes the 'footsteps'of two and three people walking simultaneously. A session is about 12-15 min long. For convenience, as shown in Figure 21, the room was divided into 12 zones so each individual footstep of the subject may take place in a particular zone. However, to verify the internal statistical consistency of the parameters in a recording session, we further group the 12 zones into 4 zone groups to get a statistically significant balanced sample of the parameters (cf. Section A.1).

The database was annotated by two annotators using the ELAN annotation tool [108] that allows creating, editing, visualizing and searching annotations for video and audio data. The annotation file includes the starting time of each 'footstep'sound together with the location labels in terms of the zone in which the footstep occurred and the foot (left or right) which produced the sound. The text labels were manually assigned based on information from the video camera. 10 sessions from Scenario-I and 13 sessions from Scenario-II were annotated. The annotations are used to coarsely localize the footstep sounds in our current

analysis.

4.5.2 Acoustic Gait Profile

The acoustic gait profile (AGP) of a person is a time-domain representation of a waveform obtained after applying a transformation, which can be linear or non-linear, on successive footstep sound bursts. The bursts in the raw waveform become identifiable pulses within an AGP and the gait specific characteristics naturally embedded in the waveform bursts—the relative energy, shape, timings, and duration—become available. The AGP is similar to other clinical diagnostic tools such as the Electrocardiogram (ECG), where the plot can be further processed by a skilled clinician or a computer to extract relevant information.

We wish to find a possible feature representation from the acoustic signal of the footsteps of a person, which preserves the prominent characteristics, as outlined before. Since these characteristics require fine temporal precision and measurements of energy over the long run, we propose to extract the temporal profile or temporal envelope of the waveform. This is achieved by a suitable signal processing technique. A *temporal profile* is an instantaneous energy representation of the signal. The *smoothed temporal profile* is a lowpassed version of the *temporal profile*. We termed the smoothed temporal profile extracted from transformed footstep sounds as the AGP in Section 4.5.2. In the following we will discuss in more details the three methods which can be used to obtain AGP.

Let us denote the lowpass filter operator with cut-off frequency set at c Hz as $\mathcal{L}_c()$. It is a fourth-order digital Butterworth filter. We can obtain the simplest profile by the basic square energy estimate (SEE). That is, for a time-domain discrete signal $x[n]$, $n \in \mathbb{Z}$, $AGP_{SEE}[n] = \mathcal{L}_c(x^2[n])$ is the SEE AGP. The HT is the method that provides another way to extract the profile. $AGP_{HT}[n] = \mathcal{L}_c(|x_a[n]|^2)$ is the Hilbert AGP.

$AGP_{TK}[n] = \mathcal{L}_c(\psi_d(x[n]))$ is the TKEO AGP. In Figures 22(a) to 22(c), we show the SEE, HT, and TKEO AGPs, respectively, of the 'footsteps'sound shown in Figure 13, with $c = 20$ Hz. When compared with the profiles from Hilbert transform (HT) (Figure 22(b))

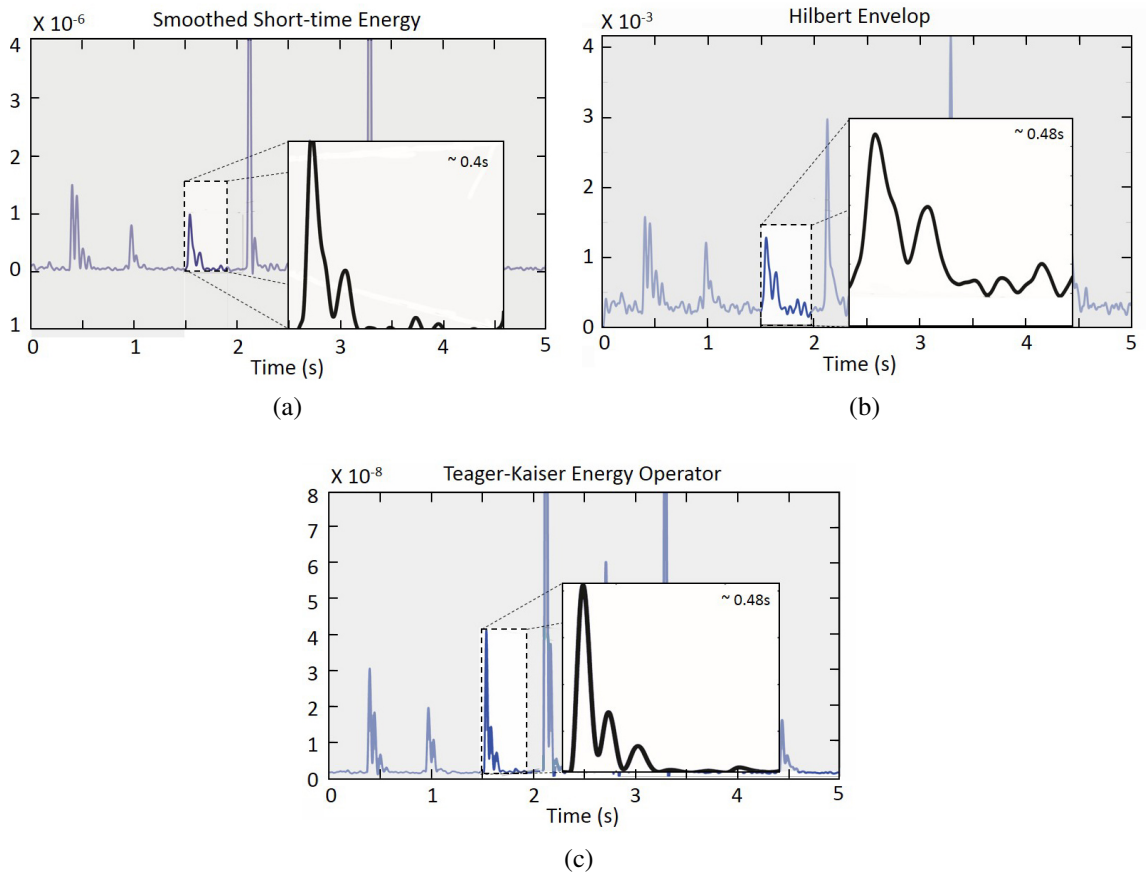


Figure 22: The SEE (a), HT (b) and TKEO (c) processed energy profiles of the raw recorded footstep shown in Figure 13. Adapted from [74, 97].

and square energy estimate (SEE) (Figure 22(a)), the TKEO profile clearly shows three sub-bursts in the footstep sound around the 1.5 sec mark, while HT and SEE profiles miss the last sub-burst. TKEO preserves the three sub-bursts as three pulses. TKEO AGP provides the most crisp representation of the sound, as the pulses are clearly delineated. This is because the decay of the sound burst is similar to an exponential decay and $\psi_d(e^{-bn}) = 0, \forall b \geq 0$. Moreover, TKEO combines the instantaneous frequency, in addition to signal amplitude, in the energy estimate—combining the local and long-term measurement in a single estimate.

4.5.2.1 Acoustic Gait Analysis

Acoustic gait analysis is the processing of the AGP to automatically extract the gait characteristics. The following general observations guide the acoustic gait analysis framework.

Consistency of successive footsteps. In a normal sequence of steps a walking subject's gait demonstrates a certain degree of statistical consistency, natural fluctuations notwithstanding. An effective analysis method will reflect this statistical consistency in its results. This can be manifest in the continuity of the walk, the dynamic balance, or the variability in cadence and speed.

Presence of multiple pulses in a step burst. Each sound burst due to a footstep is found to consist of one, two or more sub-bursts. They arise as different parts of the foot make contact with the walking surface. The energy distribution, temporal duration, and the number of sub-bursts are correlated with a particular walking style of an individual. The proper analysis of these sub-bursts yields the *pulses*.

Asymmetry between left and right foot. The acoustic characteristics produced by the two feet may be asymmetric. These asymmetries may be modulated by footwear, floor surface, carry-on objects or a pathological condition.

The nature of the sensing mechanism in acoustic gait analysis limits it to the stride and ground reaction forces level. As we will discuss in the next section, we can infer certain

facts about the posture and kinematic parameters but we cannot directly perform posture and kinematic analyses, which are beyond the scope of acoustic gait analysis and this paper.

Of particular interest is the number of pulses in one footstep sound, C. These sounds arise when the different parts of the foot—the heel, the toes and the metatarsals-phalanges—make contact with the ground during walking and, interestingly, are audible as distinct impacts. The time-resolution and TKEO-based energy measurements provided by the AGP allow us to detect them in a footstep sound. The features are directly correlated with the physical attributes of the source in a way that a cepstral coefficient is not; this partly explains the inadequacy of systems based on similar STFA features, e.g., in [52, 54, 55], and their struggle to model such sounds. We plot the distribution of C and E_L/E_R for different subjects in Figures 25 and 26, respectively.

4.5.2.2 AGP measurements

We now detail the procedure used for obtaining the raw measurements from the AGP. Let n_i be the starting time epoch of a footstep sound in samples, where $i \in \{1, \ldots, N\}$ and N is the total number of footsteps. n_i is determined from the annotations, as described in Section 4.5.1, though it can also be estimated automatically using the techniques in Section 4.4. Let δ_1 be the maximum duration, in samples, of the first pulse for any footstep AGP. From our observations, we set $\delta_1/F_s = 200$ msec, where F_s is defined in Section 4.5.1. The highest magnitude of the first pulse of the i^{th} footstep sound, E_1^i, is determined from the AGP with the following

$$\mathrm{E}_1^i = \max_{n \in [n_i, n_i+\delta_1]} \{AGP_{\{.\}}[n]\}. \tag{29}$$

The time sample epoch of E_1^i, $n_i^{\mathrm{E}_1}$, is straightforward to obtain with $n_i^{\mathrm{E}_1} = \arg\max_{n \in [n_i, n_i+\delta_1]} \{AGP_{\{.\}}[n]\}$. To determine the magnitude and time sample epoch of the second pulse of the i^{th} footstep sound, E_2^i, we calculate an adaptive threshold, for a particular session m and channel p with a global mean $\mu_{m,p}$. The mean is estimated by iterated expectations over the interval $[n_i + \delta_1, n_{i+1}] \, \forall \, i$, i.e., $\mu_{m,p}^i = E[x[n_i + \delta_1 : n_{i+1}]]$ and

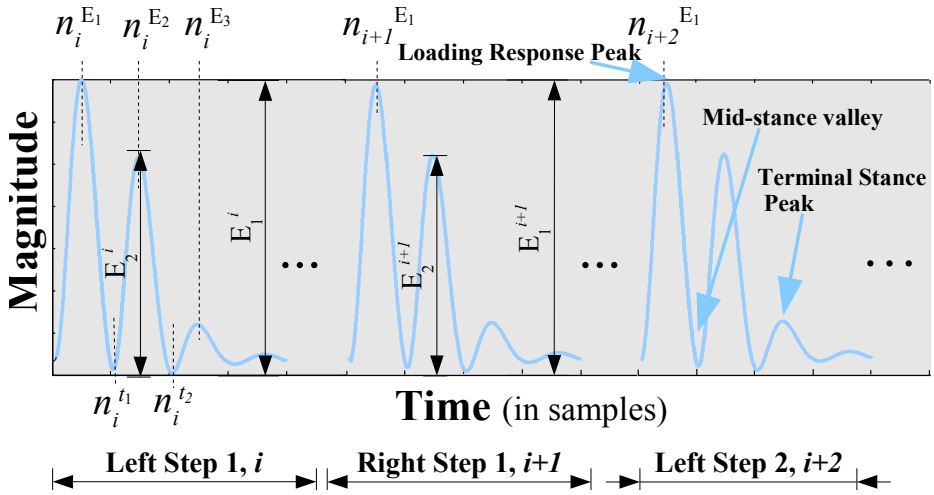


Figure 23: An illustration of the measurements that can be extracted from the AGP along the magnitude and time axes. Adapted from [97].

$\mu_{m,p} = E[\mu^i_{m,p}]$, where $E[.]$ is the expectation operator and : is the range operator. Thus,

$$\mathrm{E}^i_2 = \max_{n\in[n_i+\delta_1, n_{i+1}]} \{AGP_{\{.\}}[n] | AGP_{\{.\}}[n] > \beta\mu_{m,p}\}, \tag{30}$$

where $\beta = 1.20$. E^i_2 is undefined if there is no value above this threshold. The time sample epoch of E^i_2, $n_i^{\mathrm{E}_2}$, is obtained with $n_i^{\mathrm{E}_2} = \arg\max_{n\in[n_i+\delta_1, n_{i+1}]}\{AGP_{\{.\}}[n] | AGP_{\{.\}}[n] > \beta\mu_{m,p}\}$. Likewise, we can determine the third pulse magnitude, E^i_3, and the time sample epoch, $n_i^{\mathrm{E}_3}$, if any.

The time epochs of the first trough, $n_i^{\mathrm{t}_1}$, and second trough, $n_i^{\mathrm{t}_2}$, which occur after the first and second pulses, respectively, are estimated with $n_i^{\mathrm{t}_1} = \arg\min_{[n_i^{\mathrm{E}_1}, n_i^{\mathrm{E}_2}]}\{AGP_{\{.\}}[n] | AGP_{\{.\}}[n] > 0\}$ and $n_i^{\mathrm{t}_2} = \arg\min_{[n_i^{\mathrm{E}_2}, n_i^{\mathrm{E}_3}]}\{AGP_{\{.\}}[n] | AGP_{\{.\}}[n] > 0\}$, when the respective intervals are defined. We are not interested in the magnitude of the troughs. Figure 23 illustrates the magnitude and time epoch measurements that can be taken from an AGP.

The database includes the footstep sounds from multiple channels and in this study we only use the sound from the single channel closest to the source, due to the preliminary nature of this study. This channel was identified using video. The multi-channel capability is explored in Chapter 5 which, along with automatic detection of footsteps, can lead to a better collection of the initial sound and will provide acoustic localization and tracking,

Table 5: A summary of the acoustic gait parameters

Parameter	Definition
C	The number of pulses in one footstep sound.
D_1	$(n_i^{\mathrm{E}_2} - n_i^{\mathrm{E}_1})/F_s$, needs $C \geq 2$.
D_2	$(n_i^{\mathrm{E}_3} - n_i^{\mathrm{E}_2})/F_s$, needs $C \geq 3$.
T_1	$(n_i^{\mathrm{t}_1} - n_i^{\mathrm{E}_1})/F_s$
T_2	$(n_i^{\mathrm{t}_2} - n_i^{\mathrm{E}_2})/F_s$
E_L and E_R	E_1^i and E_1^{i+1}, respectively, when i is the index of a left footstep.
S_1	$(n_{i+1}^{\mathrm{E}_1} - n_i^{\mathrm{E}_1})/F_s$
S_{2L} and S_{2R}	$(n_{i+2}^{\mathrm{E}_1} - n_i^{\mathrm{E}_1})/F_s$, i is the index of a left footstep for S_{2L}.

which is currently accomplished with video.

4.5.2.3 The Acoustic Gait Parameters

Table 5 gives a subset of the AGP parameters that can be obtained from the raw measurements in Section 4.5.2.2. They are automatically collected for every subject with each footstep and, over a large number of footsteps, generate a statistical distribution. The number of pulses per footstep sound, derived from the sub-bursts, deserves special mention. We use the term *cardinality*, C, of a footstep to indicate the number of these pulses in a footstep sound. C = 1,2, and 3 corresponds to *single*, *double*, and *triple* footsteps, respectively.

Most of the parameters available from acoustic gaits are spatio-temporal gait parameters [109], such as stride-time or cadence. Others, such as E_L/E_R and C, are not strictly time or distance parameters but both of these parameter types are sourced from the same time-varying sound pressure measurement. Therefore, the latter category of parameters also inherits a spatio-temporal character.

4.5.3 Clinical Significance

The parameters identified in the last section are by no means exhaustive but they do provide a reasonable subset that are used for GA in clinical settings. We first notice that the AGP in Figure 22(c) resembles, though not identical to, the vertical ground reaction forces (GRF) plots [110, 111] traditionally used for GA in the stance phase of a gait cycle. GRF

Table 6: Mapping between the AGP and GRF terms

Acoustic gait term	Gait analysis term
E_L/E_R	Balance asymmetry
D_1	N/A
D_2	N/A
T_1	Mid stance Interval
T_2	Terminal Stance interval
AGP	Vertical GRF pattern
Cadence/velocity	Cadence/velocity
Cardinality	Forward/Backward lean

shows the magnitude and direction of loading applied to the foot structures during locomotion. The AGP provides the magnitude of such forces and the information about the stance phase of the gait cycle. Various epochs within the stance phase indicated by the AGP, such as the initial contact, the loading response, the mid stance valley and the terminal stance, can be easily mapped to the GRF plot. For example, E_L/E_R is the "balance asymmetry" in conventional gait analysis, T_1 is the mid-stance interval and T_2 the terminal stance interval, to name a few. The AGP provides this information with off-the-shelf microphone sensors which are cheaper and easier to set up than force plates or pressure mats, which are normally used in gait laboratories. Table 6 shows such a mapping.

Body-weight balancing is an important factor included in gait assessment tests such as BBS and POMS. Interestingly, most people with normal weight systematically distribute more weight on one foot than the other as they walk [112]. This asymmetry appears as energy asymmetry between the AGP of the right and the left footstep quantified by the ratio of sound energy, E_L/E_R.

In Figure 26 we show the normalized statistical distribution of E_L/E_R for different healthy people with a normal walking style. This parameter was obtained from the database that we collected from normal walking subjects (see Section 4.5.1). A value of E_L/E_R other than 1 corresponds to an asymmetry in gait. We notice that the distribution of this ratio can be well approximated by a log-normal distribution with the parameters of the log-normal distribution given in Figure 26. Clearly, we are able to identify subtle asymmetries in

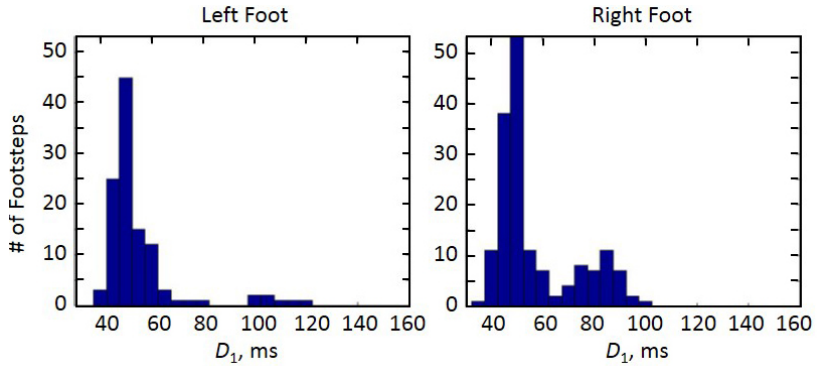
weight distribution for healthy individuals—E_L/E_R is distributed more evenly for Subjects 7 and 9 when compared to E_L/E_R for Subjects 3 and 4, which is skewed towards the left and right, respectively. These measurements can be extended to evaluation of foot and gait pathologies [113].

Stride time, stride length, cadence (steps/min), walking speed (meters/sec) and their stability are fundamental gait assessment parameters [114, 115]. It is straightforward to obtain the stride time, cadence, and, given the total walking distance, the stride length from S_{2R} and S_{2L}.

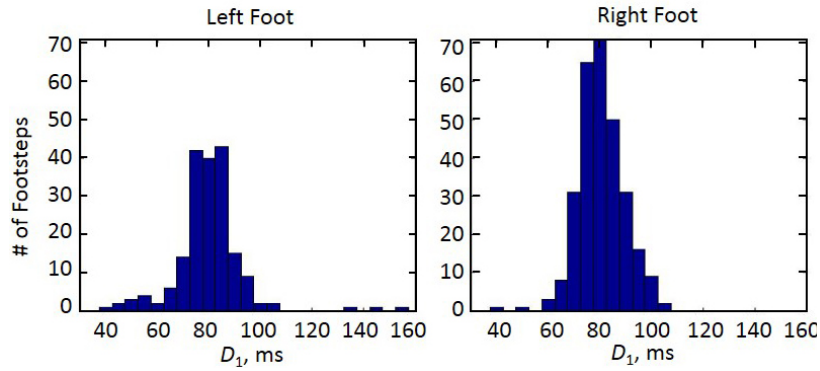
The elderly, unless they have diseases such as osteoporosis with kyphosis, can walk upright with no forward lean [116], although some healthy people walk normally with a forward lean. In either case, forward lean would almost always lead to increased proportion of $C > 1$ cardinality steps and may suggest pathological issues for unhealthy subjects. In Figure 25, we plot the proportion of C of left-footsteps for several healthy subjects with normal gait during the session. This plot shows that subject 1 has the highest number of left-footsteps with $C = 3$, while subject 10 does not have any. Most of the left-footsteps of subject 5 have $C = 2$.

In Figure 24, we plot the distribution of D_1, for subject 1 from left and right foot. D_1 is a proxy for step height and foot clearance parameters, which are used in POMS gait assessment tests.

Figures 24 to 26 illustrate the statistical distribution of a subset of acoustic gait parameters that can be extracted from the AGP and show the symmetry, or otherwise, of these parameters from the left and right foot. The clustering of the distributions around a central point indicates a consistency in a given session for a given person and we fully quantify this statistical consistency in Appendix A and also compare the consistency of respective parameters extracted with the analysis methods.



(a)



(b)

Figure 24: Left and right foot distribution of D_1 for subjects 1 and 7 in (a) and (b), respectively. The distribution of both feet is clustered around 50 ms. Note the lateral asymmetry due to a second prominent distribution mode for the right foot. For subject 7, the distribution of both feet is clustered around 80 ms. Adapted from [96, 97].

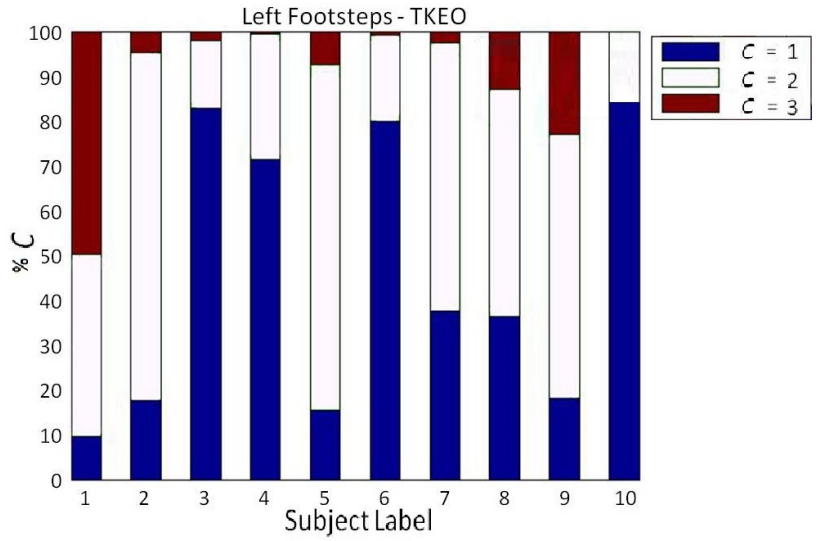


Figure 25: Percentage of C of 'footsteps'for 10 subjects produced by the left foot. Adapted from [97].

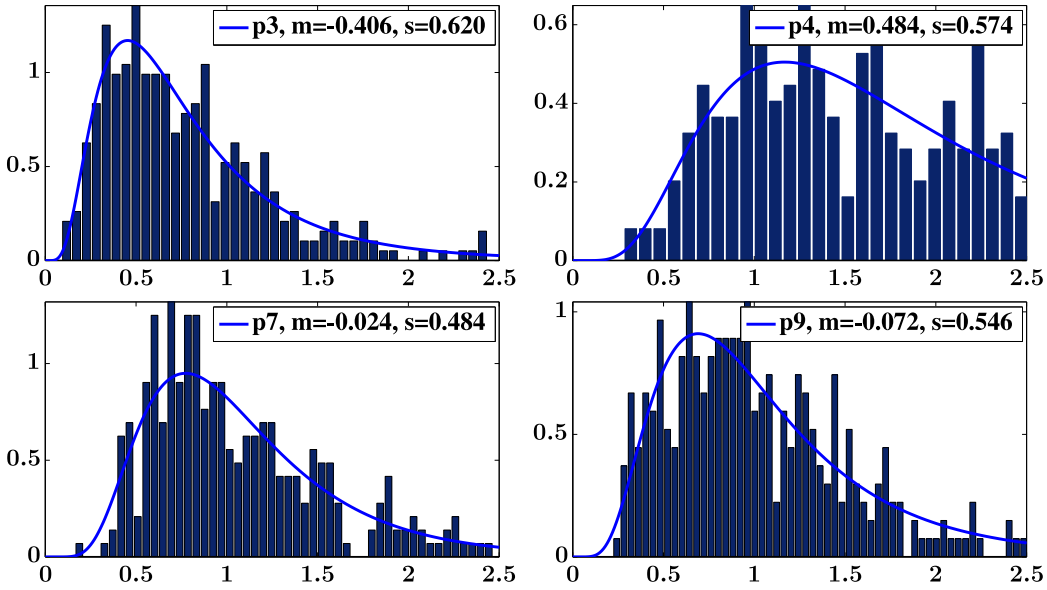


Figure 26: The distribution of E_L/E_R for 4 subjects estimated with TKEO AGP. The number followed by **p** is the subject label, while **m** and **s** are the parameters of the best-fit lognormal distribution. Adapted from [97].

4.5.4 Biometrics in the AGP

A biometric is a physical characteristic of an individual that can be used for person identification and verification. A person possess many biometrics and a few have been used to identify an individual such as finger prints, facial characteristics, palm prints, retinal images, speech, etc. Gait is one such characteristic which is unique to a person [117] and it is attractive as a biometric as it does not need the cooperation of the person and is, arguably, not easy to fake in the standard surveillance scenarios. It is also less intrusive. Nevertheless, it is a challenging problem in signal processing to use human gait as a biometric, even though human subjects have been shown to be capable in identifying people using gait only [118].

The biometric gait characteristics can be modeled with features derived from spectral, spatio-temporal, kinematic, or kinetic domains. Spectral flux, harmonicity and cepstrum are the spectral features, which are mainly derived from the short-time Fourier power analysis (STFA) of the signal. Spatio-temporal features include average walking velocity, stride length, step length, step time, cadence, stance phase time, swing phase time, single support

(when only one foot is in contact with the floor), double support (when both feet are in contact with the floor), and stride width. Kinematic features are derived from the study of angles between the ankle, hip and kneejoints. Finally, kinetic analysis examines moments, energy and power at these joints.

These features can be extracted from various modalities. By far the most widely used modality has been video [119, 120, 106], especially under the auspices of DARPA's HumanID project [112]. In some scenarios seismic signals are used to model and identify a person's gait [121, 122].

Audio has also shown promise in this task [98] and has the advantage of being less intrusive, less expensive, and more amenable to nocturnal surveillance tasks. Tanaka and Inoue [100] use the STFA to extract the frequency and pitch of a footstep while [123, 124] use MFCCs, spectral envelope and append walking intervals. Spectral features designed for speech and music processing are also used in [125, 54] for this task. An acoustic Doppler radar [126] has been used to identify kinematic features of the gait. In [127], the authors use the spectro-temporal modulation features while in [99] the structural information, i.e., 'footsteps' frequency and 'footsteps' signal energy is used to identify the person and detect whether that person is going up or down on a staircase.

In Figures 24 to 26 the distribution of the AGP parameters for different people provides a ready set of features that can characterize and dicriminate the walking style of an individual person. Thus, these AGP parameters, which we have used to characterize the gait, can be used as biometric markers to identify a person as these parameters are correlated with the physical properties of the walking person. In the following, we describe the experiments based on this ideas and show that these parameters can indeed work as biomarkers for an individual.

4.5.4.1 Experiments

For the following experiments, we divided the database (cf. Section 4.5.1) into two parts: 80% were used for training 'footsteps' models for each individual person and the remaining 20% were used for testing. Cross-validation is applied to average the recognition results. All the parameters are extracted from the TKEO AGP of the corresponding 'footstep' sounds.

The task of person identification is defined as finding a person $m \in \{1 \dots M\}$ that most likely produced a test sequence of N 'footsteps'$\{f_1, f_2, \dots f_N\}$. $M = 10$ is the total number of persons in the database. Note that the AGP of each individual 'footstep' sound f_j of cardinality C_j corresponds to either left or right foot. For each person m the likelihood, $P_m(x, C, L)$, of the AGP of a 'footstep'sound with cardinality C and produced by the foot L with the interval between humps $x = (D_1; D_2)$ is given by equation:

$$P_m(x, C, L) = \sum_{c=1}^{3} \sum_{l=1}^{2} \delta_{C,L}(c = C, l = L) p_m(c, l) \mathcal{N}_m(x, c, l) \tag{31}$$

Where $\delta_{C,L}(.)$ is an indicator function equal to 1 when $c = C, l = L$ and 0 otherwise; $p_m(c, l)$ is the prior probability that the AGP of the sound of footstep produced by the person m and the foot l has cardinality c; $L = 1$ for left foot and $L = 2$ for right foot. $\mathcal{N}_m(x, c, l)$ is the posterior probability distribution from the GMM that models the interval, x, between the humps. Note that for $c = 1$ (single 'footstep') both values D_1 and D_2 are assumed to be equal to 0 and $\mathcal{N}_m = 1$. The log-likelihood of a sequence of N 'footstep' sounds is computed as a sum of log-likelihoods from the AGP of each individual 'footstep' sound f_j.

Given the AGP of a set of N 'footstep' sounds we also model the energy ratio R between left and right foot for each person m. We group the AGP from N 'footstep' sounds into K chunks in such a way that the AGP of each chunk includes 'footsteps' from the same zone in the room, where $K = 12$ is the number of zones. This way the AGP of the left and right 'footsteps' from the same chunk occur at similar distance to the microphone. Assuming each chunk Ω_k includes V_k left and W_k right 'footsteps' with the corresponding energies

$(E_{1,left}, \ldots, E_{V,left}, E_{1,right}, \ldots, E_{W,right})$, the energy ratio can be defined as,

$$R_k = \frac{\frac{1}{V_k}\sum_{i=1}^{V_k} E_{i,left}}{\frac{1}{W_k}\sum_{j=1}^{W_k} E_{j,right}}. \tag{32}$$

Given a set of features R_k, where $k = 1, \ldots, K$ we train a GMM for each person m. During testing the log-likelihood from this model is combined with the log-likelihood obtained from (31).

The human identification results are presented in Figure 27. We show the recognition rate as the function of number of individual 'footsteps'used for identification during testing. As one may notice, 45% of correct identification is achieved just using 3 footsteps, and the maximum identification rate, 95%, is achieved by using more than 50 footsteps from each individual during testing.

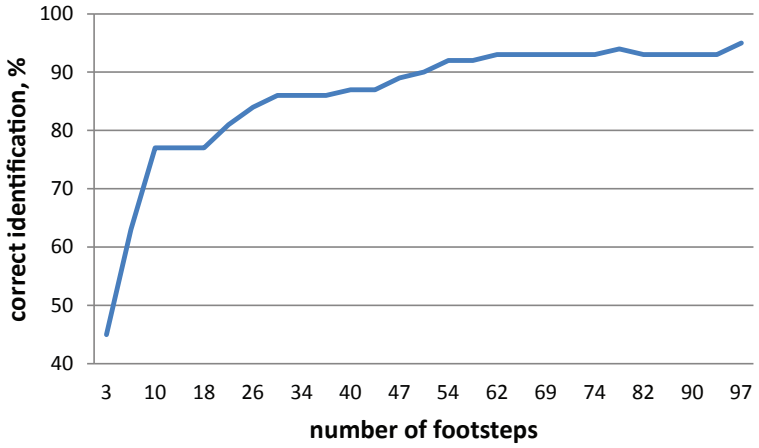


Figure 27: Person identification results as a function of a number of 'footsteps'used during testing. Adapted from [96].

4.6 Discussion and Summary

We began this chapter by performing an analysis of the sound made by regular footsteps of a human. Following the sound analysis hierarchy in the light of the characteristics of this particular sound, we argued that the spectral characteristics of this sound are less important

than its temporal properties. The temporal concepts are incorporated at feature and signal model level for detection of 'footstep'audio signal. The TKEO operator allowed us to accurately represent the self-similarity as it calculates energy as a functions of amplitude and frequency while we used EDHMM to model the duration of these repetitions. We further provided evidence for this claim in a footstep sound detection task when we explicitly modeled the duration and recurring character of the sound to improve the detection accuracy of 'footsteps'. The results indicate a significant improvement over the state of the art for the 'footsteps' sound in cross-environmental detection task.

By building a source model for the footstep sound, we introduce the novel approach of using audio to extract quantitative characteristics of the human gait. The analysis with these suitable methods uncovered a physically meaningful structure embedded within the sound and established a connection between the sound and its sources—the foot and its interaction with the ground. We used the connection to study this regular interaction, which provides a window into a person's gait. It enables in situ data acquisition and testing of the subject without wearing any instruments—as such instruments may alter the gait characteristics to be measured. We show that the sound from footsteps does encapsulate information useful for gait analysis and that this information can be consistently extracted.

With the results in Appendix A, we can be reasonably confident that the parameters provided by acoustic gaits are internally consistent and reliable enough over a session to be usable in gait analysis. In Section A.2, where we compare the parameters estimated within each zone group of a given session, we validate a premise of acoustic gaits that the sounds are picking up the characteristics related to gaits of the person. The TKEO profile provides better estimates when compared with other methods as evidenced by its consistency within the session and the low standard error of its estimates.

The results in Section A.3 indicate that E_L/E_R, which is energy based, is sensitive to footwear variation, more so when estimated with TKEO and HT methods. On the one hand, this shows that these methods are able to pick-up subtle variations in gait. On the

other hand, it shows that this statistic is not a good biometric due to its dependence on footwear, when estimated with TKEO. D_1, which is duration based, seems to be better at footwear independence due to its low relative variation. In case footwear dependency is to be avoided (as in some critical situations), acoustic gaits with bare foot (i.e, footwear free) would be advisable.

The difference between an energy based and duration based statistic is understandable because the latter is tied to the length of the footwear, while the former depends more strongly on the construction – such as the presence or absence of heel and its width and height – and material of the footwear. This implies that a normalization over footwear is essential for making acoustic gaits practically viable. However, at this point the data and analysis that we present does not allow us to point out the exact origin and extent of variations in the parameters due to footwear.

The variation in the parameters which is caused by footwear variation bears directly upon the use of these parameters as biometrics. Although, the biomarkers have been shown to be particularly useful—95% of a maximum accuracy is achieved in a human identification task, with 45% of accuracy is achieved by using just 3 footsteps from a person—the duration based biomarkers are more useful for the biometric identification of a person.

Our results for biometrics, although not definitive, are nonetheless an interesting step in the identification of person using footsteps sound without using short-time spectral features. A much harder problem is that of recognizing individuals performing more complex walks such as random walk and running. However, the acoustic footstep features proposed represent only a part of the information that could be obtained. The identified novel features are constructed to be room and shoe invariant since they are based on timing and energy ratio information from audio signal. Future work can test this empirically with a more diverse database.

The proposed approach assumes that 'footsteps'are accurately detected, roughly localized to a region and that the left and right foot assignment is made. In the study, the

annotations provided this information. This is necessary in light of the relatively unexplored nature of the acoustic gaits problem and the preliminary nature of the study. A more complete system should include 'footsteps'detection and localization from the audio signal itself. These tasks can clearly be performed from the database–the former, for example using the methods in Section 4.4 and the latter can be extracted from the microphone array, which we will explore in the next chapter. The lateral assignment of a footstep to left or right can only be provided by video, though in some limited scenarios, such as going around in a circle, it is possible to extract this information from audio.

With the novelty of the study in mind, the particular recording setup was designed to provide maximum data for subsequent analyses. Thus, the design of the experiment is overdetermined in a few respects and the results can be replicated with a reduced or simplified setup. E.g., the consistency of the parameters indicates that we can reduce the number of rounds in our recording scenarios by 30-40%, the subject can also walk in a straight line, and the number of microphones can be reduced as long as the walking area is adequately covered and the sounds are accurately localized.

The analysis and results in this chapter allows us to augment the current mostly qualitative gait assessment tests with objective and quantitative parameters, identify and extract gait characteristics in a reliable and consistent manner, and ensure that they are physically correlated with gait spatio-temporal characteristics. The proposed approach also makes available some new parameters which are not available with qualitative gait assessments. However, the use of this technique in clinical conditions is predicated on further studies, which are necessary to compare and correlate the results with current assessment methods. The new parameters also need validation for consideration as part of the gait assessment criteria.

CHAPTER 5

LOCALIZATION AND TRACKING OF FOOTSTEPS

5.1 Introduction

In Section 4.5.1 we described the acoustic gaits database and noted that the footstep sounds were digitally recorded with 16 audio channels. Up till now, our focus has been on proper sound analysis and, as a result, we have not exploited the multi-channel capability of the database. In Figure 21, we show the schematic of the room, the demarcations of the 12 zones, and the placement of the microphones around it. The manual annotations provide the zone number from which the footstep sound originates. Using these coarse locations in the previous chapter, which specify the location within a rather large rectangular zone, we used the audio channel closest to the zone in which the footstep sound originated to perform sound analysis.

Localizing the source of the footstep sound from the multiple audio channels can automate the process of finding the closest microphone and obviate the need for any manual annotations. This can improve the overall viability of the acoustic gaits as a stand alone system. We can further enhance the signal, e.g., by beamforming, if we can track the movement of the person's footsteps. In addition to signal enhancements, the location and tracking estimates can also inform acoustic gait analysis by providing distance metrics, such as step length, in addition to time measurements without relying on manual annotations.

In the following we will localize and track the movements of the person through the sounds of their footsteps and compare the performance with manual annotations. The limited resolution of the manual annotations, which locate footstep sounds source in a rather large rectangular zone, only allow us to evaluate the localization and tracking performance at the zone level. We will set up the problem in Section 5.2 as a distributed source localization problem and explore various solutions in Section 5.3. We will extend the localization

to tracking by incorporating the constraints on the movement of the subject and the geometry of the room in Section 5.4. We provide details on our ground truth labels in Section 5.5. We describe experiments on the algorithms in Section 5.6, discuss the results of these experiments and conclude with a summary in Section 5.7.

5.2 Time Delay Estimation

The goal of sound localization is to determine location of the sound source with respect to certain global coordinates. At first glance this may not seem the most suitable approach to address our problem. In particular, if we suppose our objective is only to determine the audio channel closet to the sound source, the most straightforward way to proceed might be to determine the audio channel which has the highest instantaneous energy at any given time instant.

While intuitive, this approach assumes that all microphones in the distributed system are closely placed; that they are identically constructed; and the sounds picked up by the microphones are amplified to same levels by similar amplifiers. Our recording setup satisfies none of these assumptions and, thus, a system based on energy thresholding is bound to be suboptimal. The impulsive character of the sounds that we are trying to localize may also be problematic in this case as the high energy levels only last for less than half a second and do not allow for a sustained comparison across all microphones.

Instead, we focus on the temporal properties of the signal which are largely independent of the particular amplitudes for this task. The temporal properties are also subject to issues such as room reverberation. Reverberation refers to the perception of persistence of a sound. It arises due to the effect of reflected sound waves which arrive at the ear or microphone shortly after the sound wave from the actual source and creates a lingering impression for the sound even after the sound waves from the actual source have ceased to emit. The extent of reverberation is determined by the room geometry and construction. It is quantified by the time required for the power of sound to fall below a certain threshold,

e.g., RT_{60} is a measure of room reverberation which is the time required for a sound to fall 60 dB below its initial power. It can be very challenging to localize and track sound sources in highly reverberant rooms. In our case, as indicated in Section 4.5.1, such effects can be safely ignored as the recording room has low reverberation time. Our choice of algorithms will be guided by low-reverebrant, low noise tracking scenario for a single sound source. A review of the methods for dealing with such problems can be found in [128, 129].

Consider a set of M microphone pairs, positioned at known but arbitrary locations along the periphery of a rectangular room. We assume plane wave propagation. Each microphone pair constitutes a *sensor*, enumerated as $m = 1, \ldots, M$. For the m^{th} sensor, $\mathbf{p}_{m1}$ and $\mathbf{p}_{m2}$ are locations of the two microphones in the pair and $\boldsymbol{\ell}_{s,t} = [x_{s,t}, y_{s,t}]$ is the location of the single source at time t, both with respect to the center of the room. The microphones at $\mathbf{p}_{m1}$ and $\mathbf{p}_{m2}$ in each sensor are separated by a distance d_m. With each sensor, we can estimate the time delay of arrival (TDOA) at time t, $\tau_{m,t}$ as follows,

$$\tau_{m,t} = \frac{|\mathbf{p}_{m1} - \boldsymbol{\ell}_{s,t}| - |\mathbf{p}_{m2} - \boldsymbol{\ell}_{s,t}|}{v}, \tag{33}$$

where we neglect the multi-path effects and $v = 340$ m/s is the speed of sound wave. This delay can be converted into the corresponding angle of arrival,

$$\theta_{m,t} = \arccos \frac{v\tau_{m,t}}{d_m}, \tag{34}$$

where $\theta_{m,t}$ is the angle of arrival (AOA) measured from the midpoint, $\boldsymbol{\ell}_m = [x_m, y_m]$, of the microphones. For the rest of the this chapter, we will drop the subscript t from symbols for clarity. For two sensor pair with $m = i$ and j, $\boldsymbol{\ell}_s$ can be calculated by solving the following equation:

$$\begin{bmatrix} -\tan\theta_i & 1 \\ -\tan\theta_j & 1 \end{bmatrix} \begin{bmatrix} x_s \\ y_s \end{bmatrix} = \begin{bmatrix} y_i - x_i \tan\theta_i \\ y_j - x_j \tan\theta_j \end{bmatrix} \tag{35}$$

where (35) calculates the point of intersection of the two lines described by the sensor locations and angles subtended with respect to their center. Such a system with two sensor pairs is shown in Figure 28.

Of course, $\boldsymbol{\ell}_s$, being the source location, is unknown in (33). Thus, we need to estimate the delay. Let $x_{mn}(t)$ denote the signal frame received at the n^{th} microphone of the m^{th} sensor at time, t. $n \in \{1, 2\}$ as each sensor is a microphone pair. Let $X_{mn}(\omega, t) = \mathcal{F}(x_{mn}(t))$ denote the Fourier domain representation of the signal. Then, the TDOA estimate with the well-known generalized cross correlation or GCC method [130] is given by:

$$f_m(\tau, t) = \int G_m(\omega) X_{m1}(\omega, t) X^*_{m2}(\omega, t) e^{j\omega\tau} d\omega \tag{36}$$

where $G_m(\omega) \in \mathbb{R}^+$ is a weighting term. (36) is calculating a weighted cross correlation between the two signals. The weighting term provides a sharp dominant peak for the cross correlation function and reduces spurious peaks due to room reverberation [130]. A common choice for the weighting term is:

$$G_m(\omega) = \frac{1}{|X_{m1}(\omega)||X_{m2}(\omega)|} \tag{37}$$

which results in the Smoothed Coherence Transform GCC or SCOT-GCC localization algorithm [128]. The weighting function in SCOT-GCC acts as a pre-whitening filter in the frequency domain of the cross correlation function which becomes a sharp peak in the time domain. For the m^{th} sensor, the TDOA estimate at time t is given by:

$$\hat{\tau}_{m,t} = \arg\max_{\tau} f_m(\tau, t) \tag{38}$$

Using (38) in (34) and (35) provides an estimate of $\boldsymbol{\ell}_s$.

We can notice that the microphone levels do not play a role in (38). As long as relative microphone amplifications do not fluctuate in a given recording session, we do not need any normalization on the audio file. Because of this, once the TDOA information is available for all sensor pairs, we can combine the TDOA estimates all sensors without worrying

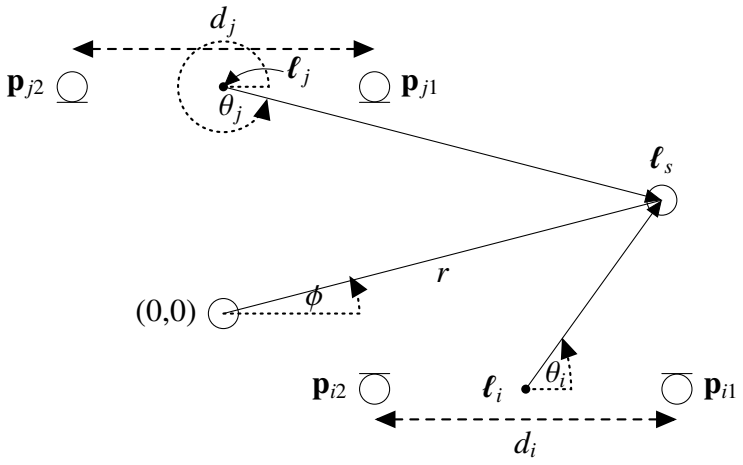


Figure 28: A schematic of the room with a sensor pair.

about their amplifications levels. We will now elaborate on combining the AOA estimates from the M sensors to estimate the location of the source.

5.3 Sound Source Localization

Since we have M sensors, using the procedure in the previous section can provide one estimate for each sensor pair, for a total of $\binom{M}{2}$ estimates of the AOA. Ideally, the intersection of the M lines for each of the AOA estimate should provide a unique location estimate, but that is seldom the case. This can be due to measurement errors and unrealistic modeling assumptions, among others. One way to address this problem is to estimate the location as the point which minimizes the mean square distance to each point of intersection. A straightforward extension of (35) into (39) provides such an estimate,

$$\begin{bmatrix} -\tan\hat{\theta}_1 & 1 \\ \vdots & \vdots \\ -\tan\hat{\theta}_M & 1 \end{bmatrix} \begin{bmatrix} \hat{x}_s^{LSE} \\ \hat{y}_s^{LSE} \end{bmatrix} = \begin{bmatrix} y_1 - x_1\tan\hat{\theta}_1 \\ \vdots \\ y_M - x_M\tan\hat{\theta}_M \end{bmatrix}. \tag{39}$$

The linear system in (39) is overdetermined and can be solved by calculating the pseudo-inverse of the matrix on the RHS of equation. This gives us a least square error (LSE) estimate. The LSE estimate is highly susceptible to noise, reverberation and measurement errors which produce outliers in AOA estimates.

Instead of combining all the information into a single estimate, we can adopt a statistical approach.I.e., we explicitly estimate the pairwise $\binom{M}{2}$ locations from the sensors using (35), and then further generalize this by noting that we need not be limited to permutations of pairs of sensors only. We can use AOAs from k-sensors at a time, and slightly modify (39) to estimate the location. Thus, we generate $\binom{M}{k}$ permutations for the k-tuples from the M sensors, where $k = 2 \dots M$. To that end, let $\alpha^{k,p}(i) : [2, M] \times [1, \binom{M}{k}] \times [1, k] \longrightarrow [1, M]$ be the permutation function which returns the index of the sensor in the p^{th} permutation of the k-tuple for the i^{th} member of the permutation set, where $1 \leq i \leq k$.

$$\begin{bmatrix} -\tan\hat{\theta}_{\alpha^{k,p}(1)} & 1 \\ \vdots & \vdots \\ -\tan\hat{\theta}_{\alpha^{k,p}(k)} & 1 \end{bmatrix} \begin{bmatrix} \hat{x}_s^{k,p} \\ \hat{y}_s^{k,p} \end{bmatrix} = \begin{bmatrix} y_{\alpha^{k,p}(1)} - x_{\alpha^{k,p}(1)} \tan\hat{\theta}_{\alpha^{k,p}(1)}, \\ \vdots \\ y_{\alpha^{k,p}(k)} - x_{\alpha^{k,p}(k)} \tan\hat{\theta}_{\alpha^{k,p}(k)} \end{bmatrix}, k = 2 \dots M \,, p = 1 \dots \binom{M}{k}. \quad (40)$$

Thus, from M sensors we have a total of $K = \sum_{k=2}^{M} \binom{M}{k}$ location estimates. To get the best location estimate from the K potential location, we adopt a statistical approach and form a pseudo log-likelihood:

$$F(\hat{x}_s, \boldsymbol{x}) = \sum_{k=2}^{M} \sum_{p=1}^{\binom{M}{k}} q_{k,p} \log(\mathcal{N}(\boldsymbol{x} : \hat{x}_s^{k,p}, \sigma^2)), \quad (41a)$$

$$F(\hat{y}_s, \boldsymbol{y}) = \sum_{k=2}^{M} \sum_{p=1}^{\binom{M}{k}} q_{k,p} \log(\mathcal{N}(\boldsymbol{y} : \hat{y}_s^{k,p}, \sigma^2)), \quad (41b)$$

and the value that maximizes this likelihood, the maximum likelihood estimate (MLE), gives the most likely location, i.e.,

$$\hat{x}_{MLE} = \arg\max F(\hat{x}_s, \boldsymbol{x}), \quad (42a)$$

$$\hat{y}_{MLE} = \arg\max F(\hat{y}_s, \boldsymbol{y}), \quad (42b)$$

where $\mathcal{N}(x; m, \sigma^2)$ denotes a Gaussian distribution with mean m and variance σ^2 evaluated at x. (42a) assumes that each estimate is independent of each other. $\hat{\boldsymbol{\ell}}_s = (\hat{x}_{MLE}, \hat{y}_{MLE})$ is the location estimate. $q_{k,p} < 1$ is the prior probability that the $(k, p)^{th}$ potential location is the true source location.

5.4 Source Tracking with Particle Filters

The above algorithms could work well for localization of the single sound source, but they do not exploit the regularity of the movement and temporal correlations between consecutive estimates embedded in our problem. The tracking approach uses these kinematic characteristics between successive localization estimates to improve the overall performance.

In order to analyze and make inference about the locations and its change with time in a dynamic system, at least two models are required: First, a model that describes the evolution of the system's state with time (the system model) and, second, a model relating the noisy measurements to the state (the measurement model). The state of a system at a particular time characterizes the system at that time and encapsulates some of its history up to that point in time. We will use a probabilistic formulation to specify the state.

If the successive locations are linearly related to each other, then the models for the dynamic system can be linear. For the location tracking problem, a linear relationship implies that the source is moving in a more or less straight lines. Kalman filter, among a few others, provide optimal estimates for certain constrained linear systems in the sense that they minimize the Bayes error for probabilistic models when noise is assumed to be Gaussian. For non-linear dynamic models, the most general probabilistic approach is suggested by particle filtering algorithms. In the following we will describe and implement this approach to solve our problem. Our development will follow [131].

5.4.1 The Probablistic Tracking Problem

To define the problem of tracking, consider the evolution of the state sequence $\boldsymbol{\gamma}_t$ of a target given by

$$\boldsymbol{\gamma}_t = \boldsymbol{f}_t(\boldsymbol{\gamma}_{t-1}, \boldsymbol{v}_{t-1}), \tag{43}$$

where $\boldsymbol{f}_t$ is a function of the state $\boldsymbol{\gamma}_{t-1}$, $\{\boldsymbol{v}_{t-1}\}$ is an i.i.d. noise sequence. The objective of tracking is to recursively estimate $\boldsymbol{\gamma}_t$ from measurements

$$\boldsymbol{\rho}_t = \boldsymbol{h}_t(\boldsymbol{\gamma}_t, \boldsymbol{n}_t), \tag{44}$$

where $\boldsymbol{h}_t$ is a function of the state $\boldsymbol{\gamma}_t$, $\{\boldsymbol{n}_t\}$ is an i.i.d. measurement noise sequence. In particular, we seek estimates of $\boldsymbol{\gamma}_t$ based on the set of all available measurements $\boldsymbol{\rho}_{0:t} = \{\boldsymbol{\rho}_i : 0 < i < t\}$, up to time t. Thus the tracking problem is to recursively calculate some degree of confidence in the state $\boldsymbol{\gamma}_t$ at time t, given the data $\boldsymbol{\rho}_{0:t}$. Thus, we construct a pdf $p(\boldsymbol{\gamma}_t|\boldsymbol{\rho}_{0:t})$ with the assumption that $p(\boldsymbol{\gamma}_0|\boldsymbol{\rho}_0) = p(\boldsymbol{\gamma}_0)$, which is also known as the prior probability distribution, is available. Then, in principle, the pdf may be obtained, recursively, in two stages: prediction and update. Suppose that the required pdf $p(\boldsymbol{\gamma}_{t-1}|\boldsymbol{\rho}_{0:t-1})$ at time $t-1$ is available. The prediction stage involves using the system model (43) to obtain the prior pdf of the state at time via the following,

$$p(\boldsymbol{\gamma}_t|\boldsymbol{\rho}_{0:t-1}) = \int p(\boldsymbol{\gamma}_t|\boldsymbol{\gamma}_{t-1})p(\boldsymbol{\gamma}_{t-1}|\boldsymbol{\rho}_{0:t-1})d\boldsymbol{\gamma}_{t-1}. \tag{45}$$

(45) uses the fact that (43) describes a Markov process of order one. The probabilistic model of the state evolution is defined by the system equation (43) and the known statistics of $\boldsymbol{v}_{t-1}$. At time t, the measurement $\boldsymbol{\rho}_t$ may be used to update the prior distribution (update stage) via Bayes' rule, i.e.,

$$p(\boldsymbol{\gamma}_t|\boldsymbol{\rho}_{0:t}) = \frac{p(\boldsymbol{\rho}_t|\boldsymbol{\gamma}_t)p(\boldsymbol{\gamma}_t|\boldsymbol{\rho}_{0:t-1})}{p(\boldsymbol{\rho}_t|\boldsymbol{\rho}_{0:t-1})} \tag{46}$$

$$p(\boldsymbol{\rho}_t|\boldsymbol{\rho}_{0:t-1}) = \int p(\boldsymbol{\rho}_t|\boldsymbol{\gamma}_t)p(\boldsymbol{\gamma}_t|\boldsymbol{\rho}_{0:t-1})d\boldsymbol{\gamma}_t. \tag{47}$$

where the denominator in (47) depends on the likelihood function defined by the measurement model (44) and the known statistics of $\{\boldsymbol{n}_t\}$. In the update stage (46), the measurement is used to modify the prior density to obtain the required posterior density of the current state. The recurrence relations (45) and (46) form the basis for the solution. This solution is optimal in the Bayesian sense when the respective distributions represent the true distributions of the data. Furthermore, the recursive propagation of the posterior density is only a conceptual solution in that in general, it cannot be determined analytically but particle filters approximate the solution (45) and (46) for the case when $\boldsymbol{f}_t$ and $\boldsymbol{h}_t$ are not linear.

5.4.2 Particle Filter Algorithm

We describe the so-called sequential importance sampling (SIS) version of particle filtering (PF) algorithm [131]. The key idea is to represent the required posterior density function by a set of random samples, the particles, with associated weights and then compute estimates based on these samples and weights. As the number of samples becomes very large, this characterization becomes an equivalent representation to the usual functional description of the posterior pdf, and the SIS filter approaches the solution (45) and (46).

Let $\{\boldsymbol{\gamma}_{0:t}^i, w_t^i\}_{i=1}^{N_{pf}}$ denote a random measure that characterizes the posterior pdf $p(\boldsymbol{\gamma}_{0:t}|\boldsymbol{\rho}_{0:t})$ where, $\{\boldsymbol{\gamma}_{0:t}^i, i = 0, \ldots, N_{pf}\}$ is a set of support points with associated weights $\{w_t^i, i = 0, \ldots, N_{pf}\}$ and $\boldsymbol{\gamma}_{0:t} = \{\boldsymbol{\gamma}_j; 0 < j < t\}$ is the set of all states up to time t. The weights are normalized to sum to one, $\sum_i w_t^i = 1$. Then, the posterior density can be approximated as

$$p(\boldsymbol{\gamma}_{0:t}|\boldsymbol{\rho}_{0:t}) = \sum_{i=1}^{N_{pf}} w_t^i \delta(\boldsymbol{\gamma}_{0:t} - \boldsymbol{\gamma}_{0:t}^i). \tag{48}$$

We therefore have a discrete weighted approximation to the true posterior, $p(\boldsymbol{\gamma}_{0:t}|\boldsymbol{\rho}_{0:t})$. The weights are chosen using the principle of importance sampling [132]. For a recursive estimation of the weights, we have

$$w_t^i \propto w_{t-1}^i p(\boldsymbol{\gamma}_t|\boldsymbol{\rho}_{0:t}) = \frac{p(\boldsymbol{\rho}_t|\boldsymbol{\gamma}_t^i) p(\boldsymbol{\gamma}_t^i|\boldsymbol{\rho}_{0:t-1}^i)}{\chi(\boldsymbol{\gamma}_t|\boldsymbol{\gamma}_{0:t-1}, \boldsymbol{\rho}_t)}, \tag{49}$$

where $q(.)$ is the importance density such that $\boldsymbol{\gamma}^i \sim \chi(\boldsymbol{\gamma}), i = 1, \ldots, N_{fp}$. Now the posterior density can be approximated as

$$p(\boldsymbol{\gamma}_t|\boldsymbol{\rho}_{0:t}) \approx \sum_{i=1}^{N_{pf}} w_t^i \delta(\boldsymbol{\gamma}_t - \boldsymbol{\gamma}_t^i), \tag{50}$$

with the weights defined in (49). The SIS algorithm thus consists of recursive propagation of the weights and support points as each measurement is received sequentially. It can be shown that the approximation (50) approaches the true posterior as $N_{pf} \to \infty$. A pseudo-code description of this algorithm is given by algorithm 1.

5.4.3 Particle Filtering Formulation

We now specify the state, $\boldsymbol{\gamma}_t$, state update model, $\boldsymbol{f}_t$, the measurement, $\boldsymbol{\rho}_t$, measurement model $\boldsymbol{h}_t$, and the measurement likelihood, $p(\boldsymbol{\rho}_t|\boldsymbol{\gamma}_t)$ for the general formulations in (43), (44) and (46).

We first define a first-order model of the source location state at time t as:

$$\boldsymbol{\gamma}_t = [r_t, \phi_t, \dot{\phi}_t]^T \tag{51}$$

Where $[r_t, \phi_t]^T$ are the polar coordinates of the true location. We use the polar coordinates from now on instead of the Cartesian coordinates because the source dynamic model is more naturally represented in the polar coordinates for our scenario as the source is constrained to move in a roughly ellipsoid shape. $[\dot{\phi}_t]$ is the angular source velocity. $\boldsymbol{\rho}_t = [\hat{r}_t^i, \hat{\phi}_t^i]_t$, $i = 1, \ldots, K$ are the polar coordinate transformations of the estimates in (40). $\boldsymbol{h}_t$ is a straightforward transformation which omits $\dot{\phi}_t$. Following [133], we set $p(\boldsymbol{\rho}_t|\boldsymbol{\gamma}_t) = F(\boldsymbol{\rho}_t, \boldsymbol{\gamma}_t)$, where $F(.,.)$ is pseudo-likelihood in (41) modified for polar coordinates.

We will now elaborate on the non-linear source dynamic model $\boldsymbol{f}_t$. The source's dynamic model is based on the Langevin stochastic process [134] which has been show to be useful as a model for tracking problems [135, 133]. The basic idea is to model a source's location as a Brownian motion which is driven by the stochastic nature of its velocity. Since the sound source in our problem is following a roughly elliptical path, we model the angle and angular velocity updates with the following Langevin equations given in (52a) and (52b):

$$\dot{\phi}_t = a_\phi \dot{\phi}_{t-1} + b_\phi F_\phi, \tag{52a}$$

$$\phi_t = \phi_{t-1} + \Delta T \dot{\phi}_t, \tag{52b}$$

$$r_t = r_{t-1} + \Delta T \dot{r}_t, \tag{52c}$$

$$a_\phi = \exp(-\beta_\phi \Delta T), \tag{52d}$$

$$b_\phi = v_\phi \sqrt{1 - a_\phi^2}, \tag{52e}$$

where ΔT is the frame increment interval, β_ϕ is the estimate update interval, and v_ϕ is the radial velocity. F_ϕ is a normally distributed random variable.

For expressing the r_t update in terms of $\dot{\phi}_t$ in (52c), we begin by noting that:

$$\dot{r}_t = \frac{dr}{dt} = \frac{dr}{d\phi}\frac{d\phi}{dt} = \frac{dr}{d\phi}\dot{\phi}_t \tag{53}$$

We now introduce the constraint that the source follows an elliptical path, with major and minor axes a and b, respectively and derive a relation for $\frac{dr}{d\phi}$, i.e.,

$$d(r_t^2) = d(a^2\cos^2(\phi_t) + b^2\sin^2(\phi_t)), \tag{54a}$$

$$2rdr = d((a^2 - b^2)\cos^2(\phi) + b^2) \tag{54b}$$

$$2rdr = d(\frac{a^2 - b^2}{2}(1 + \cos(2\phi)) + b^2) \tag{54c}$$

$$2rdr = d(\frac{a^2 + b^2}{2} + \frac{a^2 - b^2}{2}\cos(2\phi)) \tag{54d}$$

$$rdr = \frac{b^2 - a^2}{2}\sin(2\phi)d\phi \tag{54e}$$

$$\frac{dr}{d\phi} = \frac{b^2 - a^2}{2r}\sin(2\phi) \tag{54f}$$

Plugging (54f) in (53), we get:

$$\dot{r}_t = \frac{b^2 - a^2}{2r_t}\sin(2\phi_t)\dot{\phi}_t, \tag{55}$$

Put $\dot{r}_t$ from (55) in (52c) and solving for r_t, we get the following,

$$r_t = \frac{r_{t-1}}{2} + \frac{\sqrt{r_{t-1}^2 + 4d(\phi_t, \dot{\phi}_t)}}{2}, \tag{56}$$

where $d(\phi_t, \dot{\phi}_t) = \frac{b^2-a^2}{2}\Delta T\dot{\phi}_t\sin(2\phi_t)$. (52) and (56) constitute the specifications of the model for sound source movement, $\boldsymbol{f}_t$. Algorithm 1 gives the pseudo-code of the particle filtering algorithm modified for the current problem. In the next section, we will evaluate the techniques and algorithms that we discussed in this section on our database and compare the performance with the ground truth indicated by our manual annotations.

Algorithm 1: A Particle filtering algorithm modified for the current problem [133, 131].

Form an initial set of particles $\{\boldsymbol{\gamma}_0^{(i)}\, i = 1, \ldots, N_{pf}\}$ and initialize them with uniform weights $\{w_0^{(i)} = 1/N_{pf}, i = 1, \ldots, N_{pf}\}$. Then, as each frame of data is received ;

1 Resample the particles from the previous frame $\{\boldsymbol{\gamma}_{t-1}^{(i)}\}$ according to their weights $\{w_{t-1}^{(i)}\}$ to form the resampled set of particles $\{\tilde{\boldsymbol{\gamma}}_{t-1}^{(i)}, i = 1, \ldots, N_{pf}\}$;

2 Predict the new set of particles $\{\boldsymbol{\gamma}_t^{(i)}\}$ by propagating the resampled set $\{\tilde{\boldsymbol{\gamma}}_{t-1}^{(i)}\}$ according to the source dynamical model ((52));

3 Form the likelihood function with (41):

$$F(\boldsymbol{\rho} \mid \boldsymbol{\gamma})$$

4 Weight the new particles according to the likelihood function:

$$w_t^{(i)} = F(\boldsymbol{\rho}_t \mid \boldsymbol{\gamma}_t^{(i)})$$

and normalize so that $\sum_i w_t^{(i)} = 1$;

5 Compute the current source location estimate $\hat{\boldsymbol{\rho}}_s$ as the weighted sum of the particle locations:

$$\hat{\boldsymbol{\rho}}_{s,t} = \sum_{i=1}^{N} w_t^{(i)} \boldsymbol{\rho}_{\boldsymbol{\gamma}}^{(i)}$$

6 Store the particles and their respective weights $\{\boldsymbol{\gamma}_t^{(i)}\, i = 1, \ldots, N_{pf}\}$;

5.5 Ground Truth Determination

Before we describe the results of our experiments on the location estimation and tracking algorithms in the previous section, we discuss the criteria for evaluating the results of the said algorithms and explain the rationale behind the choice of the ground truth. The choice of ground truth bears significantly on subsequent discussions on the experimental results.

In general, datasets for developing distributed localization and tracking algorithms are synthetically generated using delays between audio channels to simulate the TDOA which, as we noted in the previous section, specifies the location. These prior locations are the ground truth labels and, by construction, quite accurate. Noise and other distortions, such as room impulse responses, can be added later to the audio to allow the evaluation of a localization algorithm under non-ideal conditions. This type of location generation is not a possibility for the acoustic gaits database.

We want subjects to walk with minimum restrictions so as to be as close to a realistic walking scenario as possible. This necessarily precludes precise prior labeling as we cannot ask our subjects to be at a particular location at a particular time instant. Therefore, our initial goal is to select an audio channel for every footstep sound within the room so that the effects of noise and reverberations could be minimized. This goal is sufficient for the exploratory study in Section 4.5, as it does not require precise locations. Hence, the type of location labels that we manually assigned—zone labels, were motivated by a compromise between localizing the footstep and minimizing the manual work required to add coarse locations.

We manually assigned the zone location label to each footstep sound using the recorded video. The determination is made based on the heel of the foot. The footstep is labeled to be in a particular zone if the heel of the foot touching the ground is in that zone. If the heel is on the line between the zones, then we select the next zone in the direction of movement because, for a heel on the edge of the zone, the next sound generation event from of the MTP on the ground is surely going to be from the next zone in the direction of movement.

The use of zone-based locations essentially imposes a discrete label on a continuous location variable. This has implications for the evaluation as minor displacements of the location estimate may become errors, if the location estimates are near zone boundaries.

5.6 Experimental Results and Discussion

In our database, we have 16 channel audio data which implies $M = 8$ sensor pairs. Following [135], we set $\Delta T = 160$ ms, $\beta_\phi = 10$/s and $v_\phi = 0.5$ rad/s. We set $a = 2.33$ m and $b = 1.93$ m from physical measurements of the room.

In Table 7 we show the zone accuracy rates for each session estimated with the LSE algorithm (39), MLE algorithm (42a), the particle filter tracking, and the average for each method over all sessions. The zone accuracy rate is the number of times the location zone was estimated correctly by the algorithm divided by the total number of footsteps in that zone for that session. The ground truth is available to us from manual annotations in the database which were entered by utilizing the available video. The particle filtering results have been averaged over 10 runs.

The results seem to indicate that MLE outperforms all methods, closely trailed by the particle filtering versions. We also note that the best performance is not very good at less than 80% zone accuracy rate. Further, it seems to imply that the prior information and added complexity of the PFL algorithm does not translate into improved performance with respect to this evaluation metric.

If that is indeed the case then it seems more appropriate that we ignore PFL and apply a post processing on the best system's, i.e., the MLEs, estimate by inferring the movement direction—clockwise or counter clockwise, constraining the footstep location to lie within the zone boundaries and restricting consecutive steps to only belong to the current or next zone. Incorporating these boundary and sequential vicinity constraints within MLE in an ad-hoc manner may improve the zone accuracy rate. However, as we will show next, the core problem lies with the zone accuracy metric itself and this approach will not address it.

Table 7: The zone accuracy rates for different sessions and multiple methods. The number after PFL in the first row is number of particles, N_{pf}, in the particle filtering algorithm.

Session Labels	LSE	MLE	PFL-10	PFL-20	PFL-40	PFL-60	PFL-80
1	0.471	0.728	0.527	0.644	0.693	0.705	0.712
3	0.503	0.762	0.574	0.697	0.740	0.751	0.762
4	0.636	0.876	0.703	0.799	0.840	0.856	0.855
5	0.527	0.797	0.630	0.717	0.758	0.767	0.770
6	0.581	0.827	0.658	0.751	0.792	0.800	0.809
7	0.607	0.879	0.704	0.795	0.845	0.854	0.855
8	0.511	0.781	0.628	0.716	0.747	0.753	0.758
9	0.543	0.752	0.612	0.687	0.708	0.717	0.720
11	0.518	0.795	0.588	0.695	0.760	0.768	0.770
12	0.427	0.656	0.457	0.568	0.614	0.628	0.620
13	0.504	0.764	0.556	0.676	0.740	0.741	0.752
15	0.519	0.704	0.559	0.632	0.656	0.651	0.654
16	0.542	0.729	0.572	0.676	0.686	0.685	0.696
Average	0.530	0.773	0.598	0.696	0.737	0.744	0.749

To understand the reason behind this unexpected result, we take a closer look at the location of estimates. In Figure 29 we show the MLE and PFL-60 estimates for session 11. The distribution of the estimates across zones reveals that the PFL-60 estimates are more regular (note the elliptical shape) and almost all the PFL-60 estimates fall within the zone boundaries, as they should because all movements of the source were constrained to be within the zone boundaries. We can also notice that there are a few MLE estimates which fall outside these boundaries, especially around zone 7. Thus, the sub-par performance of PFL-60 is not really sub-par; in fact, the resolution of the estimate is much better than the resolution of the zone level evaluation metric and the seemingly sub-par performance is an artifact of the evaluation metric.

The confusion matrices in Tables 8 and 9 reinforce this observation and show that the estimates from MLE do lie outside the zone boundaries while that is not the case for PFL-60. We also notice from the confusion matrices that the errors are mostly limited to adjacent zones. This shows that the error rate is not a significant impediment to our goal of choosing the most suitable microphone for the a particular footstep sound in acoustic gaits since the

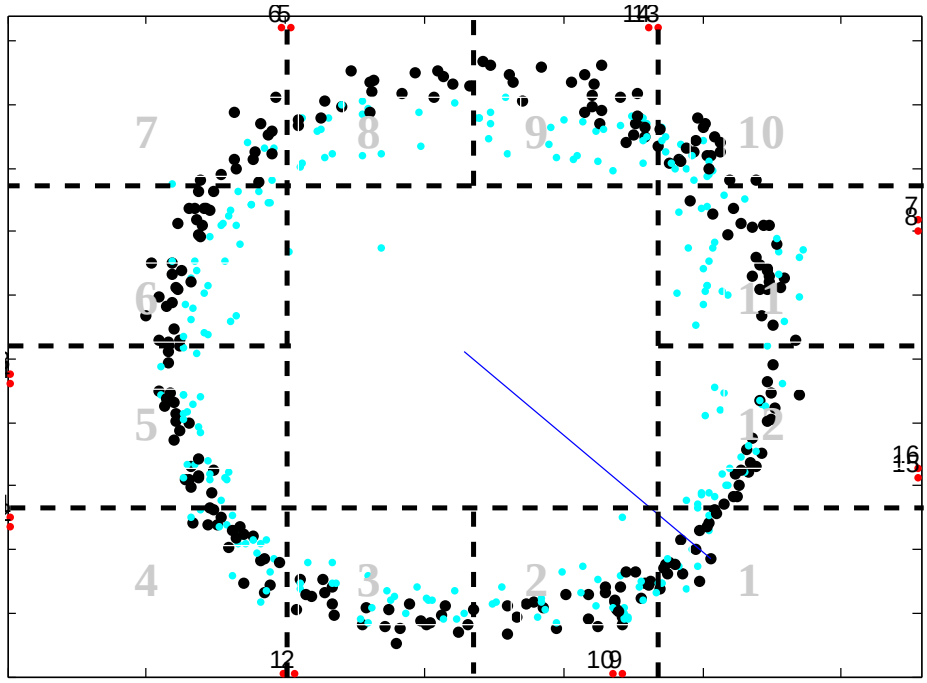


Figure 29: The black dots are the location estimates by PFL ($N = 60$) while the cyan ones are MLE estimated locations during session 11.

suitable microphone will still be chosen based on its proximity to the sound and not based on zone number alone.

Hence, the evaluation metric, the zone accuracy rate based on the ground truth indicated by the annotations, is not without its own issues. In the annotations, the footstep sound is assigned to a zone based on the point of contact of the heel with the floor and, for the many cases where the point of contact lines on a zone boundary, a displacement in estimates of few centimeters may cause a zone change. Also, the ground truth in terms of zone numbers is rather inaccurate itself and does not allow us to measure fine differences. The low accuracy rate can be explained by the results from the confusion matrices.

5.7 Summary

We used the multichannel audio recordings available in the acoustic gaits database to automate the process of choosing the best channel. We used the relative time delay of arrival at

Table 8: The confusion matrix for the 12 zones in our room, estimated with MLE for session 11. Zone number 13 is a label for any estimate which lies outside the 12 zones.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**Zone Acc. Rate**
1	8	4	0	0	0	0	0	0	0	0	0	1	0	0.61538
2	2	20	1	0	0	0	0	0	0	0	0	0	0	0.86957
3	0	1	19	0	0	0	0	0	0	0	0	0	0	0.95
4	0	0	2	15	1	0	0	0	0	0	0	0	0	0.83333
5	0	0	0	3	19	0	0	0	0	0	0	0	0	0.86364
6	0	0	0	0	4	21	1	0	0	0	0	0	1	0.77778
7	0	0	0	0	0	6	5	0	0	0	0	0	1	0.41667
8	0	0	0	0	0	2	0	17	7	0	0	0	0	0.65385
9	0	0	0	0	0	0	0	0	16	4	0	0	0	0.8
10	0	0	0	0	0	0	0	0	3	13	4	0	0	0.65
11	0	0	0	0	0	0	0	0	0	0	23	0	0	1
12	1	1	0	0	0	0	0	0	0	0	1	22	0	0.88
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0.79518

Table 9: The confusion matrix for the 12 zones in our room, estimated with PFL-60 for session 11. Zone number 13 is a label for any estimate which lies outside the 12 zones.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**Zone Acc. Rate**
1	8	4	0	0	0	0	0	0	0	0	0	1	0	0.61538
2	1	21	1	0	0	0	0	0	0	0	0	0	0	0.91304
3	0	1	19	0	0	0	0	0	0	0	0	0	0	0.95
4	0	0	3	13	2	0	0	0	0	0	0	0	0	0.72222
5	0	0	0	4	18	0	0	0	0	0	0	0	0	0.81818
6	0	0	0	0	4	22	1	0	0	0	0	0	0	0.81481
7	0	0	0	0	0	4	8	0	0	0	0	0	0	0.66667
8	0	0	0	0	0	1	1	17	7	0	0	0	0	0.65385
9	0	0	0	0	0	0	0	3	10	7	0	0	0	0.5
10	0	0	0	0	0	0	0	0	5	13	2	0	0	0.65
11	0	0	0	0	0	0	0	0	0	2	20	1	0	0.86957
12	7	0	0	0	0	0	0	0	0	0	1	17	0	0.68
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0.74699

the microphones of the sound wave to estimate an approximate location. We further incorporated the information embedded in the correlations between successive locations and the constraints on the movements to track the sound source. We used the coarse manual annotations of the locations of the sound source to judge the merits of the various approaches to

location estimation and tracking. The simulation results indicated that the computationally and conceptually complex PFL tracking did not show improved performance according to the zone accuracy rate criteria.

We presented arguments and evidence that the PFL is indeed improving performance in terms of its utility for acoustic gaits and that the problem is with the limited resolution of evaluation metric of zone accuracy rate. Even though the location estimation and tracking task is supported by the current annotations available within the database, the resolution limits a full evaluation of the various algorithms that are presented.

The location and tracking information can also be used to enhance the signal, e.g., by combining the multichannel audio by beamforming or reducing the noise in the audio signal for acoustic gaits. Further, it can also provide automated distance measurements for acoustic gaits. However, we do not attempt these tasks here. While it will be interesting to attempt these tasks and develop a metric for their evaluation in the signal processing context, a proper evaluation for such task within the acoustic gaits framework needs more extensive and precise annotations with parameters which are of interest to gait analysis community. This will also form a part of the overall validation of the overall acoustic gaits approach with respect to the conventional gait analysis metrics. The discretization of the evaluation metric and lack of an independent validation of the acoustic gaits limits the scope of the study to a localization and tracking task.

CHAPTER 6

CONCLUSION

6.1 Summary and Contributions

In this dissertation, we have proposed and demonstrated a representation and analysis framework to understand and analyze environmental sounds as a distinct category of sounds. Environmental sounds broadly lack the lexical structure embedded in other sounds such as speech and music. The proposed conceptual framework acknowledges and incorporates the particular aspects of environmental sounds which make them unique from both an ontological and a structural perspective. The proposed framework eschews the conventional single dimensional sound representation framework, the STFA, in favor of a multidimensional representation guided by a sound profile.

The contributions of the dissertation are summarized as follows:

I We proposed to repurpose the elements of an STFA representation with the help of non-STFA analysis methods. We demonstrated the need for this exercise by pointing out the latent assumptions which underlie these elements and showed that environmental sounds may not fulfill the criteria that makes these elements, and in turn the STFA representation that is built from these elements, useful otherwise.

II We purposed a multidimensional sound profile. This profile is designed to provide the missing lexical structure. We show that the sound profile does indeed reflect the complex structures embedded in the environmental sounds.

III We proposed to analyze the sounds with signal-suitable analysis methods. Together with II, this forms the analysis hierarchy where, the sounds are organized with the sound profile, and then propose to analyze similar sounds, in the sense of closer in the sound profile plane, with similar methods. Once we apply this process, we show that the performance of detection of footstep sounds on CLEAR's UPC and FBK

databases improves by 27% absolute, when compared to published results on these databases.

IV We recorded a database of human footstep sounds. Once properly analyzed with proposed methods and representations, guided by the proposed sound profile, this seemingly mundane environmental sound, nonetheless, proves to be a rich source of information and intelligence. The representation provided by the non-STFA methods is rich enough to use this sound as a basis for a novel application—the acoustic gaits, where we extract novel features for biometrics and gait analysis from footstep sounds.

V We proposed and evaluated a radial Langevin model for a particle filter tracking algorithm. The experimental results indicate that the proposed model can improve the tracking performance when compared to independent location based tracking.

6.2 Future Perspectives

The sound profile is an attempt, albeit incomplete, at developing a 'linguistics' of the environmental sounds. More work is necessary before we can be confident of the comprehensiveness of this approach for many sounds of interest to humans. Sparsity in a particular domain and regular/irregular recurrences of sound structures can also be included, but the relationship of these largely abstract mathematical and statistical constructs with perception needs to be investigated.

The signal analysis methods mentioned in this dissertation do not represent a comprehensive listing of such techniques. We have left out many, such as wavelets or the data driven empirical mode decomposition [136], or the data adaptive matching pursuit [137], or basis pursuit techniques [138], to name a few. Further, the deep neural networks (DNN) [139], if its proponents are to be believed, may render the choice of the analysis methods moot. However, we do not envision such a scenario as likely. The current stage of development of such networks in no way obviates the requirement for a perception and sound

centric analysis. In fact, DNNs may prove useful in using multi-time scale features from non-STFA methods as DNNs are widely believed to be neutral towards the unconventional features.

Clinical gait analysis can benefit substantially in terms of cost and ease of use from acoustic gait analysis. However, as it stands, our claim to validity of acoustic gait is supported by the physical correlation of the parameters to the sound production process and internal consistency of these parameters for a given person. A comparison of gait parameters, obtained by using acoustic gaits, to the current gait analysis methodologies is essential for acoustic gaits to be a viable clinical gait analysis system. Furthermore, it may not be straightforward to compare some acoustic gait parameters with other validated measures as some of the parameters may not be directly comparable to the validated gait measures due to the particular nature of the audio modality, though a correlation may be established.

The tracking performance could not be adequately judged due to limitations of the database location annotations. We do have medium quality single camera recordings that can help in improving the resolution of the annotation. This video tracking task can use the expertise from the image analysis and computer vision. The tracking task also did not make use of the non-STFA analysis methods, e.g., the sound can also be enhanced using the output of the TKEO operator as TKEO can possibly improve the onset timings of the footsteps sound. However, the lack of high resolution annotations may lead to a similar problem that we faced with particle filter tracking, i.e., any improvement may not be reflected at the level of zones.

Environmental sounds have a lot to offer and this dissertation represents a foray into organizing the diversity which accompanies their utility. As we have argued, these sounds form an integral component of human intelligence yet most of us take the information provided by these sounds for granted. Currently, one can record and store these environmental sounds by the tera-bytes but the tools to specifically understand and analyze these sounds do not go beyond human ears and spectral analyses. It is imperative that proper analysis

of these sounds be developed and more compelling applications and tasks be designed to highlight the information and intelligence potential of these sounds.

APPENDIX A

STATISTICAL ANALYSIS OF THE ACOUSTIC GAITS

A.1 Introduction

Within the acoustic gaits (AG) framework, we generate the parameters by analyzing the footstep sounds collected over a period of a few minutes. These multiple realizations of the footstep sound generation process lends itself naturally to a statistical description. We have already illustrated this when we showed the parameters in terms of their statistical distributions in Section 4.5.3. The statistical analysis in this appendix,

a) verifies the internal consistency of the parameters;

b) helps us understand the parameter variation when the conditions change for a given healthy subject; and

c) compares the results in (a) and (b) across the aforementioned three signal analysis methods to assess the relative discerning power of parameters estimated from each method.

We leverage our dual-scenario database recordings from healthy individuals, described in Section 4.5.1 to achieve the above. We achieve (a) by checking for *intra-session* consistency of the parameters estimated with recordings from Scenario-I and Scenario-II, and (b) by checking for *inter-session* consistency of the perimeters estimated from recordings in Scenario-II only. The parameters are estimated from the estimation methods for both cases and are compared to each other to achieve (c). In the subsequent discussion, each session label has the format 'S$y - x$', where $y = 1$ for sessions from Scenario-I and $y = 2$ for sessions from Scenario-II. $x \in \mathbb{N}$ is an arbitrary index.

The ratio E_L/E_R and D_1, as defined Table 5, are chosen as the study parameters in the following analyses. E_L/E_R is used instead of E_L or E_R directly, as each analysis method's energy estimate is not directly comparable. This becomes important when comparing the

estimates of this parameter with different methods, which is the goal (c) above. It also obviates the need for applying arbitrary normalizations for each scenario. A one-way balanced ANOVA design is used to achieve the objectives (a)-(c).

One of the assumptions behind ANOVA is the normality of the data but, as we show in Figure 26, this is not true of E_L/E_R derived from TKEO. To address this, we also checked the results with the Kruskal-Wallis (KW) test, which is similar to the standard ANOVA test but does not require the normality assumption. The results from both tests were similar, indicating that ANOVA is sufficiently robust against the violation of the normality assumption for this data. Thus, we used the standard ANOVA for subsequent tests.

A.2 Intra-session consistency and parameter stability

Let us denote the parameter of interest as ρ. It is collected from each of the 12 zones in the room during a given session. We group the statistics from the 12 zones into 4 zone groups (ZG) numbered 1 ...4. Zones 1, 2, and 3 are grouped as ZG1, zones 4, 5, and, 6 are grouped as ZG2 and, so on. This grouping into ZGs ensures adequate samples for each grouping and makes for a more balanced ANOVA design between the ZGs when compared with just the zones.

We assume the following fixed-effects one-way ANOVA model for parameter ρ estimated in i^{th} ZG, with j^{th} method at time t, $(\rho)_i^j(t)$:

$$(\rho)_i^j(t) = \mu^j(t) + ZG_{i,\rho}^j(t) + \epsilon_i^j(t) \tag{57}$$

where $i = 1, 2, 3$, and 4 is the ZG index and $j \in \{SEE, HT, TKEO\}$ is the label for the method used to estimate the parameter. SEE is the SEE profile, HT is the Hilbert profile and $TKEO$ is the TKEO profile. $\mu^j(t)$ is the grand mean for the j^{th} estimation method, and $ZG_{i,\rho}^j(t)$ is the effect on the ρ parameter estimated in the i^{th} ZG and j^{th} method at time t. $\epsilon_i^j(t)$ is the zero mean, unit variance Gaussian noise process for the i^{th} ZG and j^{th} method at time t. We set the significance level for ANOVA at $\alpha = 0.01$.

We expect that a parameter collected within a session is statistically consistent across

Table 10: The number of samples of the parameter E_L/E_R collected from each ZG.

Session Label	i =ZG1	i =ZG2	i =ZG3	i =ZG4
S1-3	35	35	34	39
S1-5	80	70	77	72
S1-6	83	78	72	88
S1-8	76	76	71	73
S2-1	58	56	57	59
S2-3	58	49	54	59
S2-4	57	50	53	58
S2-5	54	49	54	56
S2-7	55	54	55	57

the four ZGs. This ensures that the value of the parameter is internally stable. To test this, we set up the following null hypothesis:

$$H^j_{0\rho} : ZG^j_{1,\rho} = ZG^j_{2,\rho} = ZG^j_{3,\rho} = ZG^j_{4,\rho} \tag{58}$$

We begin with the perimeter $\rho = E_L/E_R$ in (57). Table 10 shows the number of estimated parameters for E_L/E_R in each ZG for a given session. In Table 11 we show the result of the ANOVA analysis on E_L/E_R for the sessions in Table 10. It shows that for a majority of sessions, the hypothesis $H^j_{0\rho}|_{\rho=E_L/E_R}$ is not rejected. This suggests that the variation across the ZGs is not statistically significant for those sessions. For sessions S1-5 and S1-6, the parameter estimated from SEE and HT shows significant variation at $\alpha = 0.01$, while the TKEO estimates are still insignificant. These results imply that the variation of the parameter as estimated from the audio analysis is not significant during a session, and that TKEO estimate is the most consistent among the three.

Table 11 also gives the mean value of the parameter and its standard error for each session and method. A value of $E_L/E_R < 1$ indicates an asymmetry in favor of the left foot and vice-versa for $E_L/E_R > 1$. We can notice that the asymmetry indicated by the three methods is consistent in indicating the dominance of one foot over the other within the error bounds. For example, in session S1-3 $E_L/E_R < 1$ for all methods, while for session S2-1 $E_L/E_R > 1$ for all methods, though the degree of this asymmetry is not consistent. We can also notice that the standard error is consistently the least for TKEO among all methods for

Table 11: The results of ANOVA with the E_L/E_R parameter, collected from four ZGs with the SEE, HT and TKEO profiles for the sessions in Table 10. The F-value column is the F-statistic for each one-way ANOVA, the p column gives the p-value and E_L/E_R gives the average estimate of the parameter along with its std. error. The p-values in bold are statistically significant w.r.t. $\alpha = 0.01$

	$j = SEE$			$j = HT$			$j = TKEO$		
Session Label	F-value	p	E_L/E_R	F-value	p	E_L/E_R	F-value	p	E_L/E_R
S1-3	2.86	0.039	0.70 ± 0.04	2.41	0.070	0.81 ± 0.03	1.09	0.355	0.62 ± 0.02
S1-5	8.38	**0.000**	1.33 ± 0.04	9.66	**0.000**	1.13 ± 0.05	3.02	0.030	1.22 ± 0.02
S1-6	5.41	**0.001**	1.16 ± 0.06	9.96	**0.000**	1.04 ± 0.05	2.16	0.093	0.99 ± 0.02
S1-8	1.86	0.136	1.22 ± 0.05	0.90	0.441	1.08 ± 0.05	0.32	0.814	1.15 ± 0.02
S2-1	0.95	0.416	2.09 ± 0.09	1.34	0.263	1.36 ± 0.08	0.08	0.969	2.27 ± 0.03
S2-3	2.94	0.034	0.96 ± 0.05	3.08	0.028	0.91 ± 0.04	1.46	0.226	1.01 ± 0.02
S2-4	3.00	0.031	1.38 ± 0.08	2.58	0.055	1.09 ± 0.06	0.83	0.479	1.23 ± 0.03
S2-5	1.58	0.197	1.23 ± 0.10	0.88	0.452	0.99 ± 0.07	0.11	0.953	1.02 ± 0.04
S2-7	0.35	0.791	1.29 ± 0.09	1.00	0.396	1.05 ± 0.08	0.78	0.504	1.17 ± 0.03

Table 12: The number of samples of the parameter D_1 collected from each ZG.

Session Label	i =ZG 1	i =ZG 2	i =ZG3	i =ZG4
S1-4	58	64	63	71
S1-5	148	138	134	147
S1-8	118	131	123	128
S1-9	153	159	146	163
S2-4	78	87	84	94
S2-5	79	79	70	85
S2-13	93	92	83	95

a given session.

Next we test consistency for parameter $\rho = D_1$ in (57). The setup, including the one-way ANOVA model, is identical to the case for parameter E_L/E_R, except that we do not use SEE to estimate D_1, i.e., $j \in \{HT, TKEO\}$ for $\rho = D_1$ in (57) and (58). Table 12 tabulates the number of parameter D_1 estimated in each ZG for a given session. All estimation methods will have the same number of parameters in a given session. Unlike E_L/E_R, the parameter D_1 relies on the presence of a *double step* — i.e., $C \geq 2$ — and is thus subject to the walking style of an individual. Hence, we only show sessions here where we could collect adequate number of double steps balanced across the ZGs. In Table 13 we give the results for the ANOVA analysis with the hypothesis $H^j_{0\rho}|_{\rho=D_1}$ in (58). The p-values suggest that D_1 estimated from TKEO profiles does not show significant variation within a session,

Table 13: The results of ANOVA with the D_1 parameter, collected from four ZGs with the HT and TKEO profiles for the sessions in Table 12. The F-value column is the F-statistic for each one-way ANOVA, the p column gives the p-value and D_1 gives the average estimate of the parameter along with its std. error. The p-values in bold are statistically significant w.r.t. $\alpha = 0.01$

	$j = HT$			$j = TKEO$		
Session Label	F-value	p	D_1 ms	F-value	p	D_1 ms
S1-4	4.50	**0.004**	84.17 ± 1.05	3.72	0.013	86.31 ± 1.50
S1-5	0.38	0.765	62.10 ± 0.48	1.43	0.235	64.04 ± 0.54
S1-8	3.85	**0.010**	106.90 ± 1.38	3.39	0.018	108.63 ± 1.36
S1-9	3.61	0.013	68.76 ± 0.88	0.47	0.700	70.09 ± 1.09
S2-4	8.03	**0.000**	99.90 ± 1.33	2.12	0.098	104.28 ± 1.11
S2-5	4.91	**0.002**	100.01 ± 1.23	0.14	0.934	105.15 ± 0.86
S2-13	0.39	0.761	69.19 ± 1.27	1.33	0.265	69.64 ± 1.42

while D_1 estimated from HT profile does show significant variation for some sessions such as S1-4 and S2-5. Thus we conclude that the intra-session consistency is much stronger for D_1 estimated from the TKEO profile than the HT profile.

The average values of the parameter and its standard error also appear in Table 13. D_1 estimated by the TKEO profile is consistently higher than the HT profile estimate, while the standard error is similar for both. We can also notice that the estimates of D_1 for session S2-4 and S2-5 are almost the same. This is an example where the same subject participated in the two sessions, but with different footwear. We will study this inter-session variation more throughly in the next section.

A.3 Inter-session consistency and footwear variation

We now discuss the statistical tests for the case when the same subject is recorded in different sessions with different footwear.

Let ρ denote the parameter of interest. We assume the following fixed-effects one-way ANOVA model for parameter $(\rho)^j_{i_{k,l}}(t)$, which is ρ estimated in $i^{th}_{k,l}$ session, for k^{th} subject, with j^{th} method at time t:

$$(\rho)^j_{i_{k,l}}(t) = \mu^{j,k}(t) + S^{\,j}_{i_{k,l},\rho}(t) + \epsilon^j_{i_{k,l}}(t) \tag{59}$$

Table 14: The results of ANOVA with the E_L/E_R parameter, collected from three subjects estimated from the SEE, HT, and TKEO profiles for the sessions in Table 10. The F-value column is the F-statistic for each one-way ANOVA and the p column gives the p-value. The p-values in bold are statistically significant w.r.t. $\alpha = 0.01$.

	$j = SEE$		$j = HT$		$j = TKEO$	
Subject No. k	F-value	p	F-value	p	F-value	p
3	1.94	0.145	6.31	**0.002**	7.93	**0.000**
6	0.21	0.651	0.02	0.884	2.30	0.130
8	7.03	**0.001**	7.36	**0.001**	31.32	**0.000**

where $i_{k,l}$ is the session label for the l^{th} session of the k^{th} subject and $j \in \{SEE, HT, TKEO\}$ is the label for the methods used to estimate the parameter. $\mu^{j,k}(t)$ is the grand mean for the j^{th} estimation method and k^{th} subject, and $S^{j}_{i_{k,l},\rho}(t)$ is the effect on $(\rho)^{j}_{i_{k,l}}(t)$. The footwear variation is achieved by keeping j and k constant and varying l. $\epsilon^{j}_{i_{k,l}}(t)$ is the zero mean, unit variance Gaussian noise process for the $i^{th}_{k,l}$ session and j^{th} method at time t. As before, we set the significance level for ANOVA at $\alpha = 0.01$.

We first set $\rho = E_L/E_R$ in (59). To check the consistency, or otherwise, of the E_L/E_R estimate across the footwear variant sessions, we set up the following null hypothesis for three subjects with $k \in \{3, 6, 8\}$:

$$H^{j,k}_{0E_L/E_R} : S^{j}_{i_{k,1},E_L/E_R} = S^{j}_{i_{k,2},E_L/E_R} = S^{j}_{i_{k,3},E_L/E_R} \tag{60}$$

$i_{3,\{1,2,3\}}$ = {S2-4, S2-5, S2-6}, $i_{6,\{1,2\}}$ = {S2-6, S2-8}, and $i_{8,\{1,2,3\}}$ = {S2-9, S2-10, S2-11} are the session labels for the three subjects. Note that there are only two sessions for subject $k = 6$. In Table 14, we give the results of the ANOVA analysis on E_L/E_R for these subjects. Table 14 shows that $H^{j,k}_{0E_L/E_R}$ is not rejected for subject 6 for all methods, while it is rejected for the majority of the rest. To analyze this further, we performed post-hoc tests at the same, Bonferroni corrected, significance levels. The results are given in Table 15. It shows that sessions S2-4 and S2-6, for subject 3, are not significantly different from each other while session S2-5 is significantly different from the other two for subject 3. Subject 8 not only shows variation with footwear, but the differences among the sessions are also significant for different methods.

Table 15: The table provides the average E_L/E_R estimate and its standard error from the SEE, HT, and TKEO profiles for each session with the given subject labels.

Subject No. k	Session label $i_{k,l}$	$j = SEE$	$j = HT$	$j = TKEO$
		E_L/E_R	E_L/E_R	E_L/E_R
3	S2-4	1.36 ± 0.07	1.22 ± 0.06	1.09 ± 0.03
	S2-5	1.17 ± 0.07	0.97 ± 0.06	0.96 ± 0.03
	S2-6	1.32 ± 0.08	1.26 ± 0.06	1.08 ± 0.03
6	S2-7	1.29 ± 0.08	1.17 ± 0.06	1.05 ± 0.03
	S2-8	1.24 ± 0.08	1.04 ± 0.06	1.06 ± 0.03
8	S2-9	1.02 ± 0.06	0.88 ± 0.05	0.97 ± 0.02
	S2-10	1.33 ± 0.06	1.33 ± 0.05	1.08 ± 0.02
	S2-11	1.15 ± 0.05	0.85 ± 0.04	1.08 ± 0.02

Table 16: The results of ANOVA with the D_1 parameter, collected from two subjects with HT and TKEO profile. The F-value column is the F-statistic for each one-way ANOVA and the p column gives the p-value. The p-values in bold are statistically significant w.r.t. $\alpha = 0.01$.

	$j = HT$		$j = TKEO$	
Subject No.	F-value	p	F-value	p
2	13.55	**0.000**	0.48	0.487
3	4.48	0.012	3.07	0.047

Next, we fix $\rho = D_1$ as the parameter of interest in (59). As before, $j \in \{HT, TKEO\}$ for $\rho = D_1$ in (59) is the label for the methods used to estimate D_1.

We now determine the effect on the estimate of the footwear across different sessions by setting up the following null hypothesis for two subjects with labels $k \in \{2, 3\}$:

$$H_{0D_1}^{j,k} : S^{j}_{i_{k,1},D_1} = S^{j}_{i_{k,2},D_1} = S^{j}_{i_{k,3},D_1} \tag{61}$$

$i_{2,\{1,2\}}$ = {S2-1, S2-2} and $i_{3,\{1,2,3\}}$ = {S2-3, S2-4, S2-5} are the session labels for the two subjects. Note that there are only two sessions for subject $k = 2$. In Table 16, we give the results of the ANOVA analysis on D_1 for these subjects. As we explained in the last section, the selection of these sessions was dictated by the availability of sufficient number of estimates of D_1 balanced across the sessions.

Table 16 shows that $H_{0D_1}^{j,k}$ is rejected for subject no. 2 for HT estimate, while it is on the margin for subject no. 3 for both estimation methods. To analyze this further,

Table 17: The table provides the average D_1 estimate and its standard error from the two methods for each session with the given subject labels.

Subject No. k	Session Label $i_{k,l}$	$j = HT$	$j = TKEO$
		D_1	D_1
2	S2-1	74.58 ± 0.88	79.49 ± 1.04
	S2-2	78.61 ± 0.64	80.42 ± 0.80
3	S2-3	96.29 ± 1.01	102.45 ± 0.77
	S2-4	100.25 ± 1.10	104.98 ± 0.86
	S2-5	100.04 ± 1.16	104.85 ± 0.94

we performed post-hoc tests at the same, Bonferroni corrected, significance levels. The results are given in Table 17. It shows that sessions S2-4 and S2-5, for subject 3, are not significantly different from each other while session S2-3 is significantly different from the other two for subject 3. The differences are statistically significant but, for a given method, the absolute difference in values are less than 5% for this parameter.

A.4 Summary

The data analysis of parameters E_L/E_R and D_1 in Section A.1 shows that the values do not significantly vary within a walking session ($p < 0.01$) for parameters obtained with TKEO profile. When the footwear is changed, E_L/E_R shows statistically significant ($p < 0.01$) variations, while D_1 remains unchanged.

REFERENCES

[1] J. E. Ancell, "Effect of external sound fields on hearing tests in audiometric booths," *J. Acoust. Soc. Am.*, vol. 30, no. 7, pp. 694–695, 1958.

[2] N. J. Vanderveer, "Ecological acoustics: Human perception of environmental sounds," Doctoral Thesis, Cornell University, 1979.

[3] D. Mitrovic, M. Zeppelzauer, and H. Eidenberger, "Analysis of the data quality of audio descriptions of environmental sounds," *Journal of Digital Information Management*, vol. 5, no. 2, p. 48, 2007.

[4] B. Gygi, "Factors in the identification of environmental sounds," Doctoral Thesis, Indiana University, 2001.

[5] S. Chu, S. Narayanan, and C.-C. Kuo, "Environmental sound recognition with time-frequency audio features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1142–1158, 2009. [Online]. Available: 10.1109/TASL.2009.2017438

[6] R. E. Turner, "Statistical models for natural sounds," Doctoral Thesis, University College London, 2010.

[7] B. De Coensel and D. Botteldooren, "A model of saliency-based auditory attention to environmental sound," in *Proc. Intl. Congress on Acoustics, ICA*, 2010.

[8] P. Gaunard *et al.*, "Automatic classification of environmental noise events by hidden markov models," in *Proc. IEEE ICASSP*, vol. 6, 1998, pp. 3609–3612.

[9] W. A. Yost, "Auditory image perception and analysis: The basis for hearing," *Hearing Research*, vol. 56, no. 1-2, pp. 8–18, 1991.

[10] R. D. Patterson, "Auditory warning sounds in the work environment," *Philos. Trans. R. Soc. B.*, vol. 327, no. 1241, pp. 485–492, 1990.

[11] B. C. Pijanowski *et al.*, "Soundscape Ecology: The Science of Sound in the Landscape," *BioScience*, vol. 61, no. 3, pp. 203–216, 2011.

[12] K. Van Den Doel, P. G. Kry, and D. K. Pai, "FoleyAutomatic: Physically-based sound effects for interactive simulation and animation," in *Proc. Annual Conference on Computer Graphics and Interactive Techniques*, 2001, pp. 537–544.

[13] V. Cevher, R. Chellappa, and J. McClellan, "Vehicle Speed Estimation Using Acoustic Wave Patterns," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 30 –47, 2009.

[14] T. G. Horner, "Engine Knock Detection Using Spectral Analysis Techniques With a TMS320 DSP," 1995. [Online]. Available: http://www.ti.com/lit/an/spra039/spra039.pdf

[15] B. Samimy and G. Rizzoni, "Mechanical signature analysis using time-frequency signal processing: application to internal combustion engine knock detection," *Proc. IEEE*, vol. 84, no. 9, pp. 1330 –1343, 1996.

[16] I. Howard, "A Review of Rolling Element Bearing Vibration 'Detection, Diagnosis and Prognosis',," Defence Science and Technology Organisation, Australia, Technical Report DSTO-RR-0013, 1994.

[17] M. Davy, "Application of Time-Frequency Techniques to Sound Signals: Recognition and Diagnosis," in *Time-Frequency Analysis*, F. Hlawatsch and F. Auger, Eds. ISTE, 2010, pp. 383–408.

[18] N. Checka *et al.*, "Multiple person and speaker activity tracking with a particle filter," in *Proc. IEEE ICASSP*, vol. 5, 2004, pp. V–881.

[19] D. Giannoulis *et al.*, "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.

[20] J. Behar *et al.*, "SleepAp: An automated obstructive sleep apnoea screening application for smartphones," in *Computing in Cardiology Conference (CinC), 2013*, 2013, pp. 257–260.

[21] S. Shimai, K. Fukuda, and M. Terasaki, "Pleasantness-unpleasantness of environmental sounds and gender difference in evaluation," *Perceptual and Motor Skills*, vol. 76, no. 2, pp. 635–640, 1993.

[22] G. W. Cermak, "Multidimensional analyses of judgments about traffic noise," *J. Acoust. Soc. Am.*, vol. 59, no. 6, p. 1412, 1976.

[23] L. N. Solomon, "Semantic approach to the perception of complex sounds," *J. Acoust. Soc. Am.*, vol. 30, no. 5, pp. 421–425, 1958.

[24] E. Björk, "The perceived quality of natural sounds," *Acta Acustica united with Acustica*, vol. 57, no. 3, pp. 185–190, 1985.

[25] G. R. Kidd and C. S. Watson, "Sound quality judgments of everyday sounds," *J. Acoust. Soc. Am.*, vol. 106, no. 4, p. 2267, 1999.

[26] J. C. Bartlett, "Remembering environmental sounds: The role of verbalization at input," *Memory & Cognition*, vol. 5, no. 4, pp. 404–414, 1977.

[27] C. Y. P. Chiu and D. L. Schacter, "Auditory priming for nonverbal information: Implicit and explicit memory for environmental sounds," *Consciousness and Cognition*, vol. 4, no. 4, pp. 440–458, 1995.

[28] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[29] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[30] D. Gerhard, "Audio signal classification: history and current techniques," Dept. of Computer Science, University of Regina,Regina, Canada, Tech. Rep., 2003.

[31] G. S. Ohm, "Ueber die definition des tones, nebst daran geknüpfter theorie der sirene und ähnlicher tonbildender vorrichtungen," *Annalen der Physik und Chemie*, vol. 59, no. 8, pp. 513–565, 1843.

[32] H. L. F. Helmholtz, *On the Sensations of Tone as a Physiological Basis for the Theory of Music*, 3rd ed. London: Longmans, Green, and Co., 1895.

[33] R. Plomp, "The ear as a frequency analyzer," *J. Acoust. Soc. Am.*, vol. 36, no. 9, p. 1628, 1964.

[34] H. Nyquist, "Certain factors affecting telegraph speech," *Bell System Technical Journal*, vol. 3, pp. 324–346, 1924.

[35] H. Fletcher, "A space-time pattern theory of hearing," *J. Acoust. Soc. Am.*, vol. 1, no. 3A, pp. 311–343, 1930.

[36] H. Fletcher, "Auditory patterns," *Reviews of Modern Physics*, vol. 12, pp. 47–65, 1940.

[37] J. Peterson, "Combination tones and other related auditory phenomena," Doctoral Thesis, University of Chicago, 1908.

[38] J. F. Schouten, "The residue revisited," in *Frequency Analysis and Periodicity Detection in Hearing*, R. Plomp and G. Smoorenburg, Eds. Leiden: A.W. Sijthoff, 1970.

[39] S. Rosen, "Temporal information in speech: Acoustic, auditory and linguistic aspects," *Philos. Trans. R. Soc. B.*, vol. 336, no. 1278, pp. 367–373, 1992.

[40] N. Viemeister and C. Plack, "Time analysis," in *Human psychophysics*, W. Yost, A. N. Popper, and R. R. Fay, Eds. Springer-Verlag, 1993, pp. 116–154.

[41] R. V. Shannon, "The relative importance of amplitude, temporal, and spectral cues for cochlear implant processor design." *American Journal Of Audiology*, vol. 11, no. 2, pp. 124–127, 2002.

[42] S. Greenberg, "Auditory function," in *Encyclopedia of Acoustics*, M. J. Crocker, Ed. John Wiley & Sons, Inc, 1997, vol. 3, pp. 1301–1323.

[43] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 5th ed. Academic Press, 2003.

[44] R. J. Zatorre and P. Belin, "Spectral and temporal processing in human auditory cortex," *Cerebral Cortex*, vol. 11, no. 10, pp. 946–953, 2001.

[45] T. D. Griffiths *et al.*, "Encoding of the temporal regularity of sound in the human brainstem," *Nature Neuroscience*, vol. 4, no. 6, pp. 633–637, 2001.

[46] H. Luo *et al.*, "Concurrent encoding of frequency and amplitude modulation in human auditory cortex: Encoding transition," *J Neurophysiol*, vol. 98, no. 6, pp. 3473–3485, 2007.

[47] W. Buxton, "Introduction to this special issue on nonspeech audio," *Human Computer Interaction*, vol. 4, no. 1, pp. 1–9, 1989.

[48] N. Saint-Arnaud, "Classification of sound textures," Master's Thesis, Massachusetts Institute of Technology, 1995.

[49] F. I. Klassner, III, "Data reprocessing in signal understanding systems," Ph.D. dissertation, University of Massachusetts Amherst, 1996.

[50] B. Feiten and S. Günzel, "Automatic indexing of a sound database using self-organizing neural nets," *Computer Music Journal*, vol. 18, no. 3, pp. 53–65, 1994.

[51] H. Thornburg, "Detection and modeling of transient audio signals with prior information," Doctoral Thesis, Stanford University, 2005.

[52] S. Cavaco and J. Rodeia, "Classification of similar impact sounds," in *Image and Signal Processing, Intl. Conf. on*, A. Elmoataz *et al.*, Eds. Springer, 2010, pp. 307–314.

[53] M. Casey, "MPEG-7 sound-recognition tools," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 737 –747, 2001.

[54] J. T. Geiger *et al.*, "Gait-based person identification by spectral, cepstral and energy-related audio features," in *Proc. IEEE ICASSP*, 2013.

[55] A. Temko *et al.*, "CLEAR evaluation of acoustic event detection and classification systems," in *Multimodal Technologies for Perception of Humans*, R. Stiefelhagen and J. Garofolo, Eds. Springer, 2007, pp. 311–322.

[56] R. Stiefelhagen *et al.*, "The CLEAR 2007 evaluation," in *Multimodal Technologies for Perception of Humans*, R. Stiefelhagen, R. Bowers, and J. Fiscus, Eds. Springer, 2008, pp. 3–34.

[57] L. Cohen, *Time-Frequency Analysis*. Englewood Cliffs N.J: Prentice Hall PTR, 1995.

[58] B. Boashash and P. O'Shea, "A methodology for detection and classification of some underwater acoustic signals using time-frequency analysis techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 11, pp. 1829–1841, 1990.

[59] D. Dufournet and P. Jouenne, "MADRAS, an intelligent assistant for noise recognition," in *INTERNOISE*, vol. 3, 1997.

[60] R. Hennequin, R. Badeau, and B. David, "NMF with time-frequency activations to model non stationary audio events," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 744–753, 2011.

[61] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 7, pp. 504– 516, 2002.

[62] M. A. Casey, "Auditory group theory with applications to statistical basis methods for structured audio," Doctoral Thesis, Massachusetts Institute of Technology, 1998.

[63] A. S. Bregman, *Auditory Scene Analysis*. The MIT Press, 1990.

[64] D. Ellis. (2014) Computational Auditory Scene Analysis. [Online]. Available: http://www.ee.columbia.edu/~dpwe/talks/BKN-CASA-2014-07.pdf

[65] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recognition Letters*, vol. 30, no. 14, pp. 1281–1288, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865509001603

[66] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, *Signals & systems*, 2nd ed. Upper Saddle River, N.J: Prentice Hall, 1997.

[67] B. Lathi, *Modern digital and analog communication systems*, 3rd ed. New York: Oxford University Press, 1998.

[68] A. Papoulis, *Probability, random variables, and stochastic processes*, 4th ed. Boston: McGraw-Hill, 2002.

[69] N. Wiener, "Generalized harmonic analysis," *Acta Mathematica*, vol. 55, no. 1, pp. 117–258, 1930. [Online]. Available: http://link.springer.com/article/10.1007/BF02546511

[70] W. Gardner, "Introduction to Einstein's contribution to time-series analysis," *IEEE ASSP Magazine*, vol. 4, no. 4, pp. 4–5, 1987.

[71] R. M. Warren, "Auditory illusions and perceptual processing of speech," in *Principles of Experimental Phonetics*, N. J. Lass, Ed. Mosby, 1996, pp. 435–466.

[72] (2008) UPC-TALP database of isolated meeting-room acoustic events. [Online]. Available: http://catalog.elra.info/product_info.php?products_id=1053

[73] (2008) FBK-Irst database of isolated meeting-room acoustic events. [Online]. Available: http://catalog.elra.info/product_info.php?products_id=1093

[74] M. U. B. Altaf, T. Butko, and B.-H. Juang, "Perceptually motivated temporal modeling of footsteps in a cross-environmental detection task," in *Proc. IEEE ICASSP*, 2013.

[75] F. Crawford, *Waves*, ser. Berkeley Physics Course. McGraw-Hill, 1965, vol. 3.

[76] P. T. Landsberg, *Thermodynamics and statistical mechanics*. Oxford University Press, 1978.

[77] J. F. Kaiser, "Some useful properties of teager's energy operators," in *Proc. IEEE ICASSP*, vol. 3, 1993, pp. 149–152.

[78] C. P. Clark, "Effective coherent modulation filtering and interpolation of long gaps in acoustic signals," Masters Thesis, University of Washington, 2008.

[79] B. Yegnanarayana, D. Saikia, and T. Krishnan, "Significance of group delay functions in signal reconstruction from spectral magnitude or phase," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 3, pp. 610– 623, 1984.

[80] L. D. Alsteris and K. K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digit. Signal Process.*, vol. 17, no. 3, pp. 578–616, 2007.

[81] M. Kazama *et al.*, "On the significance of phase in the short term fourier spectrum for speech intelligibility," *J. Acoust. Soc. Am.*, vol. 127, no. 3, pp. 1432–1439, 2010.

[82] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. I. Fundamentals," *Proc. IEEE*, vol. 80, no. 4, pp. 520–538, 1992.

[83] J. N. Oppenheim and M. O. Magnasco, "Human time-frequency acuity beats the fourier uncertainty principle," *Physical Review Letters*, vol. 110, no. 4, 2013.

[84] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *Proc. IEEE ICASSP*, vol. 1, 1990, pp. 381–384.

[85] M. U. B. Altaf and B.-H. Juang, "Audio signal classification with temporal envelopes," in *Proc. IEEE ICASSP*, 2011, pp. 469 –472.

[86] S. Nakamura *et al.*, "Design and collection of acoustic sound data for hands-free speech recognition and sound scene understanding," in *Proc. IEEE Intl. Conf. on Multimedia and Expo*, vol. 2, 2002, pp. 161–164.

[87] L. Cohen, "Time-frequency distributions-a review," *Proc. IEEE*, vol. 77, no. 7, pp. 941–981, 1989.

[88] C. Corduneanu, *Almost Periodic Oscillations and Waves*. Springer, 2009.

[89] P. Flandrin and P. Borgnat, "Revisiting and testing stationarity," in *Journal of Physics: Conference Series*, vol. 139, 2008.

[90] S. Kay, "A new nonstationarity detector," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1440–1451, 2008.

[91] M. B. Priestley and T. S. Rao, "A test for non-stationarity of time-series," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 31, no. 1, pp. 140–149, 1969.

[92] J. B. Gao, "Detecting nonstationarity and state transitions in a time series," *Physical review E*, vol. 63, no. 6, pp. 2021–2028, 2001.

[93] P. Borgnat *et al.*, "Testing stationarity with surrogates: A time-frequency approach," *IEEE Trans. Signal Process.*, vol. 58, no. 7, pp. 3459–3470, 2010.

[94] J. Lin, "Divergence measures based on the Shannon entropy," *Information Theory, IEEE Transactions on*, vol. 37, no. 1, pp. 145–151, 1991. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=61115

[95] (2014) BBC sound effects library original series. [Online]. Available: http://www.sound-ideas.com/bbc-1-60-hd.html

[96] M. U. B. Altaf, T. Butko, and B.-H. Juang, "Person identification using biometric markers from footsteps sound," in *Proc. INTERSPEECH*, 2013, pp. 2934–2938.

[97] M. U. B. Altaf, T. Butko, and B.-H. Juang, "Acoustic Gaits: Gait Analysis with Footstep Sounds," *To appear in IEEE Trans. Biomed. Eng.*, 2015. [Online]. Available: http://dx.doi.org/10.1109/TBME.2015.2410142

[98] K. Mäkelä, J. Hakulinen, and M. Turunen, "The use of walking sounds in supporting awareness," in *Proceedings of ICAD*, 2003, pp. 144–147. [Online]. Available: http://dev.icad.org/websiteV2.0/Conferences/ICAD2003/paper/35%20Makela.pdf

[99] D. Alpert and M. Allen, "Acoustic gait recognition on a staircase," in *World Automation Congress (WAC)*, 2010, pp. 1–6.

[100] M. Tanaka and H. Inoue, "A study on walk-recognition by frequency analysis of footsteps," *Trans. lEE of Japan*, vol. 119-C, no. 6, pp. 762–763, 1999. [Online]. Available: https://www.jstage.jst.go.jp/article/ieejeiss1987/119/6/119_6_762/_pdf

[101] K. Saifuddin, T. Matsushima, and Y. Ando, "Duration sensation when listening to pure tone and complex tone," *Journal of Temporal Design in Architecture and the Environment*, vol. 2, no. 1, pp. 42–47, 2002.

[102] M. Johnson, "Capacity and complexity of HMM duration modeling techniques," *IEEE Signal Process. Lett.*, vol. 12, no. 5, pp. 407 – 410, 2005.

[103] J. Bach, J. Anemüller, and B. Kollmeier, "Robust speech detection in real acoustic backgrounds with perceptually motivated features," *Speech Communication*, vol. 53, no. 5, pp. 690–706, 2011.

[104] S. Ravindran, K. Schlemmer, and D. Anderson, "A physiologically inspired method for audio classification," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 9, pp. 1374–1381, 2005.

[105] M. Büchler *et al.*, "Sound classification in hearing aids inspired by auditory scene analysis," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 2991–3002, 2005.

[106] S. Sarkar *et al.*, "The HumanID gait challenge problem: Data sets, performance, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 162–177, 2005. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1374864

[107] Y. Huang, J. Benesty, and G. W. Elko, "Source localization," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds. Springer US, 2004, pp. 229–253.

[108] P. Wittenburg *et al.*, "ELAN: a professional framework for multimodality research," in *Intl. Conf. on Language Resources and Evaluation*, 2006.

[109] J. H. Hollman, E. M. McDade, and R. C. Petersen, "Normative spatiotemporal gait parameters in older adults," *Gait & posture*, vol. 34, no. 1, pp. 111–118, 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3104090/

[110] M. M. Rodgers, "Dynamic foot biomechanics," *J. Orthop. Sports Phys. Ther.*, vol. 21, no. 6, pp. 306–316, 1995.

[111] D. Sutherland, "The evolution of clinical gait analysis part III - kinetics and energy assessment," *Gait & Posture*, vol. 21, no. 4, pp. 447–461, 2005. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0966636204001201

[112] M. S. Nixon, T. N. Tan, and R. Chellappa, *Human Identification Based on Gait*. Springer, 2006.

[113] M. J. Hessert *et al.*, "Foot pressure distribution during walking in young and old adults," *BMC geriatrics*, vol. 5, pp. 8-1–8-8, 2005. [Online]. Available: http://www.biomedcentral.com/1471-2318/5/8/

[114] J. H. Hollman *et al.*, "Does walking in a virtual environment induce unstable gait?: An examination of vertical ground reaction forces," *Gait & Posture*, vol. 26, no. 2, pp. 289–294, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0966636206002827

[115] S. Fritz and M. Lusardi, "White paper: "Walking speed: the sixth vital sign"," *J. Geriatr. Phys. Ther.*, vol. 32, pp. 2–5, 2009.

[116] J. O. Judge, "Gait disorders in the elderly," in *The Merck Manual of Diagnosis and Therapy*, 19th ed., R. S. Porter and J. L. Kaplan, Eds. Merck Sharp & Dohme Corp., 2011, p. 3085.

[117] J. E. Cutting and L. T. Kozlowski, "Recognizing friends by their walk:Gait perception without familiarity cues," *Bulletin of the Psychonomic Society*, vol. 9, no. 5, pp. 353–356, 1977.

[118] K. M. Ostrosky *et al.*, "A comparison of gait characteristics in young and old subjects," *Physical Therapy*, vol. 74, no. 7, pp. 637–644, 1994.

[119] L. Lee and W. E. L. Grimson, "Gait analysis for recognition and classification," in *Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, 2002, pp. 148–155.

[120] J. J. Little and J. E. Boyd, "Recognizing people by their gait: The shape of motion," *Videre: Journal of Computer Vision Research*, vol. 1,2, pp. 2–32, 1998.

[121] R. J. Orr and G. D. Abowd, "The smart floor: a mechanism for natural user identification and tracking," in *Proceedings of the Conference on Human Factors in Computing System*, 2000, pp. 275–276. [Online]. Available: http://dl.acm.org/citation.cfm?id=633453

[122] M. Alwan *et al.*, "Method and system for the derivation of human gait characteristics and detecting falls passively from floor vibrations," U.S. Patent 7 857 771, 2010. [Online]. Available: http://www.google.com/patents?id=vjrwAAAAEBAJ

[123] Y. Shoji, T. Takasuka, and H. Yasukawa, "Personal identification using footstep detection," in *International Symposium on Intelligent Signal Processing and Communication Systems ISPACS*, 2004, pp. 43 – 47.

[124] R. de Carvalho and P. Rosa, "Identification system for smart homes using footstep sounds," in *IEEE International Symposium on Industrial Electronics (ISIE)*, 2010, pp. 1639–1644.

[125] A. Itai and H. Yasukawa, "Footstep recognition with psyco-acoustics parameter," in *IEEE Asia Pacific Conference on Circuits and Systems, (APCCAS)*, 2006, pp. 992–995.

[126] K. Kalgaonkar and B. Raj, "Acoustic doppler sonar for gait recogination," in *IEEE Conference on Advanced Video and Signal Based Surveillance(AVSS)*, 2007, pp. 27–32.

[127] C. DeLoney, N. Mesgarani, and J. Fritz, "Person identification and gender recognition from footstep sound using modulation analysis," The Institute for Systems Research, University of Maryland, Technical Report 2008-17, 2008. [Online]. Available: http://drum.lib.umd.edu/handle/1903/8379

[128] G. Doblinger, "Localization and tracking of acoustical sources," in *Topics in Acoustic Echo and Noise Control*, E. Hänsler and G. Schmidt, Eds. Springer Berlin Heidelberg, 2006, pp. 91–122.

[129] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, P. M. Brandstein and D. B. Ward, Eds. Springer Berlin Heidelberg, 2001, pp. 157–180.

[130] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[131] M. S. Arulampalam *et al.*, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, 2002.

[132] N. Bergman, "Recursive Bayesian Estimation," Doctoral Thesis, Linköping University, Linköping, Sweden, 1999.

[133] D. B. Ward, E. Lehmann, and R. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. Speech, Audio Process.*, vol. 11, no. 6, pp. 826–836, 2003.

[134] D. S. Lemons and A. Gythiel, "Paul langevin's 1908 paper "On the theory of brownian motion" ["Sur la théorie du mouvement brownien," C.R. Acad. Sci. (Paris) 146, 530-533 (1908)]," *American Journal of Physics*, vol. 65, p. 1079, 1997.

[135] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *IEEE ICASSP*, vol. 5, 2001, pp. 3021–3024.

[136] N. E. Huang *et al.*, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. R. Soc. Lond. A*, vol. 454, pp. 903–995, 1998.

[137] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, 1993.

[138] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

[139] G. Hinton *et al.*, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.