DISSERTATION

Characterization of Multiple Time-Varying Transient Sources from Multivariate Data Sequences

Submitted by

Neil Wachowski

Department of Electrical and Computer Engineering

In partial fulfullment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2014

Doctoral Committee:

Advisor: Mahmood R. Azimi-Sadjadi

F. Jay Breidt
Kurt Fristrup
Ali Pezeshki

Abstract

Characterization of Multiple Time-Varying Transient Sources from Multivariate Data Sequences

Characterization of multiple time-varying transient sources using sequential multivariate data is a broad and complex signal processing problem. In general, this process involves analyzing new observation vectors in a data stream of unknown length to determine if they contain the signatures of a source of interest (i.e., a signal), in which case the source's type and interference-free signatures may be estimated. This process may continue indefinitely to detect and classify several events of interest thereby yielding an aggregate description of the data's contents. Such capabilities are useful in numerous applications that involve continuously observing an environment containing complicated and erratic signals, e.g., habitat monitoring using acoustical data, medical diagnosis via magnetic resonance imaging, and underwater mine hunting using sonar imagery.

The challenges associated with successful transient source characterization are as numerous as the application areas, and include 1) significant variations among signatures emitted by a given source type, 2) the presence of multiple types of random yet structured interference sources whose signatures are superimposed with those of signals, 3) a data representation that is not necessarily optimized for the task at hand, 4) variable environmental and operating conditions, and many others. These challenges are compounded by the inherent difficulties associated with processing sequential multivariate data, namely the inability to exploit the statistics or structure of the entire data stream. On the other hand, the complications that must be addressed often vary significantly when considering different types of data, leading to an abundance of existing solutions that are each specialized for a particular application. In other words, most existing work only simultaneously considers a subset of these complications, making them difficult to generalize.

The work in this thesis was motivated by an application involving characterization of national park soundscapes in terms of commonly occurring man-made and natural acoustical sources, using streams of "1/3 octave vector" sequences. Naturally, this application involves developing solutions that consider all of the mentioned challenges, among others. Two comprehensive solutions to this problem were developed, each with unique strengths and weaknesses relative to one another. A sequential random coefficient tracking (SRCT) method was developed first, that hierarchically applies a set of likelihood ratio tests to each incoming vector observation to detect and classify up to one signal and one interference source that may be simultaneously present. Since the signatures of each acoustical event typically span several adjacent observations, a Kalman filter is used to generate the parameters necessary for computing the likelihood values. The SRCT method is also capable of using the coefficient estimates produced by the Kalman filter to generate estimates of both the signal and interference components of the observation, thus performing separation in a dual source scenario. The main benefits of this method are its computational efficiency and its ability to characterize both components of an observation (signal and interference).

To address some of the main deficiencies of the SRCT method, a sparse coefficient state tracking (SCST) approach was also developed. This method was designed to detect and classify signals when multiple types of interference are simultaneously present, while avoiding restrictive assumptions concerning the distribution of observation components. This SCST method uses generalized likelihood ratios tests to perform signal detection and classification during quiescent periods, and quiescent detection whenever a signal is present. To form these tests, the likelihood of each signal model is found given a sparse approximation of an incoming observation, which makes the temporal evolution of source signatures more tractable. Robustness to structured interference is incorporated by virtue of the inherent separation capabilities of sparse coding. Each signal model is characterized by a Bayesian network, which captures the dependencies between different coefficients in the sparse approximation under the associated hypothesis.

In addition to developing two complete transient source characterization systems, this thesis also introduces several concepts and tools that may be used to aid in the development of new systems designed for similar tasks, or supplement existing ones. Of particular note are a comprehensive overview of existing general approaches for detecting changes in the parameters of sequential data streams, a new method for performing fusion of sequential classification decisions based on a hidden Markov model framework, and a detailed analysis of the 1/3 octave data format mentioned above. The latter is especially helpful since this data format is commonly used in audio analysis applications.

A comprehensive study is carried out to evaluate the performance of the developed methods for detecting, classifying, and estimating the signatures of signals using 1/3 octave soundscape data that is corrupted with multiple types of structured interference. The systems are benchmarked against a Gaussian mixture model approach that was adapted to handle the complexities of the soundscape data, as such approaches are frequently used in acoustical source recognition applications. Performance is mainly measured in terms of the receiver operator characteristics (ROC) of the test statistics implemented by each method, the improvement in signal-to-noise ratio they offer when estimating signatures, and their overall ability to accurately detect and classify signals of interest. It was observed that both the SRCT and SCST methods perform exceptionally on the national park soundscape data, though the latter performs best in the presence of heavy interference and is more flexible in new environmental and operating conditions.

Acknowledgements

Dr. Mahmood R. Azimi-Sadjadi certainly deserves my profound gratitude for being my advisor and good friend, where he devoted significant time and effort to supporting my research. It was consistently evident that he cares about the well-being of his students. I will never forget his flexibility during times of stress and turmoil, nor the numerous opportunities he offered me over the past several years. I would also like to thank my committee members Dr. Kurt Fristrup, Dr. Ali Pezeshki, and Dr. F. Jay Breidt for their time and support.

A special thanks goes to the Natural Sounds Program of the National Park Service for providing the financial support for this research under a cooperative agreement (contract #H2370094000), without which this work could not have been completed. I feel very fortunate to have been afforded the opportunity to conduct research with an associated mission that is very much in line with my personal interests and convictions. I am also grateful to those at the National Park Service who have assisted me in various aspects of my research by providing data and information, namely Emma Lynch, Damon Joyce, and Dan Mennitt.

I would like to express my appreciation for my past and present colleagues at the Colorado State University Signal and Image Processing Laboratory for providing a fun and supportive work environment. In particular, thanks to Bryan Thompson, Nick Roseveare, Derek Tucker, Mike McCarron, Gergely Kiss, Katalin Orova, Nick Klausner, Mike Kabatek, Yinghui Zhao, Amanda Dinstel, Soheil Kolouri, Adam Aboudan, Jarrod Zacher, and anyone else whom I may have forgotten. There are also *numerous* friends and acquaintances who deserve my gratitude for providing various forms of support over the years. My climbing partners played particularly important role in helping me to maintain my sanity, especially Duncan Ryan, Randy Hoover, Tina Charland, Jean-Christophe Prin, Bryan and Kathy Beavers, John Wilson, and certainly my wife, Shannon Wachowski. El Capitan is still going to be around for a little while, so I am told.

Speaking of my wife, I am indebted to her in countless other ways, such as the extraordinary level of patience she was able to maintain given the extended amount of time I have dedicated to school, and my constant need to pursue (often unconventional) outdoor adventure in order to recover and maintain balance. I am especially thankful given that she is nearly always very busy as well, and knows that I love her despite the fact that the amount of time we spent together was often limited.

Finally, I will forever be grateful to my family for their love and support throughout my life. The financial assistance provided by my parents throughout my undergraduate coursework gave me an opportunity that I have never taken for granted. Most recently, their persistence in uncertain and occasionally bleak times was remarkably inspiring.

Table of Contents

List of Tables

List of Figures

List of Notation and Symbols

Basic Definitions

Bold lowercase letter, e.g. $\mathbf{x}$	column vector
Bold uppercase letter, e.g. $\mathbf{X}$	matrix
Non-bold lowercase letter, e.g. n	scalar
Non-bold uppercase letter, e.g. N	constant
$\{x : \ldots\}$	set of x such that
$\{x_i\}_{i=n_1}^{n2}$	set $\{x_{n_1},\ x_{n_1+1},\ \cdots,\ x_{n_2}\}$
(a, b)	open interval
$[a, b]$	closed interval
$\hat{x}$	estimate of x
$\sim$	distributed as
$\overset{\text{IID}}{\sim}$	independent and identically distributed as
$\mathcal{N}\left(\mu, \sigma\right)$	Gaussian distribution with mean μ and variance σ
$\langle \mathbf{X} \rangle$	subspace spanned by the columns of matrix $\mathbf{X}$

Specific Definitions

k	discrete time index
t	continuous time instance
N	observation dimension
p	specific signal class
P	total number of signal classes
q	specific interference class
Q	total number of interference classes

$\mathbb{R}^n$	n-dimensional vector space over the field of the real numbers
$\mathbf{0}_n$	$n \times n$ zero matrix
$\mathbf{0}_{n_1 \times n_2}$	$n_1 \times n_2$ zero matrix
$\mathbf{I}_n$	$n \times n$ identity matrix
$\mathcal{H}_i$	ith hypothesis

Basic Operations

$\max\limits_{x} f(x)$	extrema value of $f(x)$ for all valid x
$\arg\max\limits_{x} f(x)$	specific value of x that achieves extrema value of $f(x)$
$\mathbf{X}^T$	transpose of matrix $\mathbf{X}$
$\mathbf{X}^H$	conjugate transpose, i.e., Hermitian transpose of matrix $\mathbf{X}$
$\mathrm{tr}\,(\mathbf{X})$	trace of matrix $\mathbf{X}$
$\det\,(\mathbf{X})$	determinant of matrix $\mathbf{X}$
$\mathbf{X}^\dagger$	Moore–Penrose pseudoinverse of matrix $\mathbf{X}$
$\lVert\mathbf{x}\rVert_p$	vector p-norm such that $\lVert\mathbf{x}\rVert_p = \left(\sum_{i=1}^{n} \lvert x_i\rvert^p\right)^{1/p}$
$\lVert\mathbf{X}\rVert_F$	Frobenius norm of matrix $\mathbf{X}$
$E\,[x]$	expected value of random variable x
$\otimes$	Kronecker product

CHAPTER 1

Introduction

1.1. Problem Statement and Motivations

The ability to characterize multiple time-varying transient sources from sequential multivariate data has a myriad of applications including speech recognition [1–7], habitat monitoring [8–16], medical diagnosis [17], and battlefield surveillance [18–21]. A transient source is one whose signatures are not continually present in the data, and hence, new observations in a sequential data stream must be constantly monitored to detect each source's time of arrival and subsequently (or simultaneously) extract more detailed information about its properties, such as its duration and class (i.e., type of source). Although detection and classification of the transient sources is the primary goal of this research, another desirable capability is estimating the signatures of a given source under conditions of significant background noise and/or the presence of competing sources, thereby parsing composites of source signatures into isolated representations.

Transient source characterization is often complicated by many factors that depend on the particular application. These factors include time-varying source signatures, a large number of possible source types, variable number of sources that may be simultaneously present leading to superimposed signatures, unknown arrival times and parameters (e.g., Doppler shift), the presence of ambient noise as well as environmental and operational variations, and variable structure and duration even among those sensed events associated with a single source type. The latter complication is due to the random nature of many sources, and can lead to 1) between-class similarities where sources belonging to different classes may produce acoustical signatures that are similar within a given time interval, and 2) within-class diversity where signatures can vary significantly enough such that it is difficult to describe them with a single model or set of parameters. Superimposed signatures are typically caused by the simultaneous presence of sources that are extrinsic and intrinsic to the sensed environment. Throughout this thesis these two categories are associated with the more general terms of *signal* and *interference*, respectively, whereas the term *source* means anything that produces signatures that do not fit the ambient noise model. An *event* refers to the signatures of a single source that often span several observations, due to the extent of such signatures often exceeding the temporal resolution of the observation sequence. In general, interference sources are generally considered a nuisance, as they hinder signal detection by shifting or obstructing discriminatory features, and hence, robustness to interference is a significant consideration of the research in this thesis.

The research in this thesis is motivated by a cooperative agreement (#H2370094000) with the Natural Sounds Program of the National Park Service (NPS) for enhancing the effectiveness of acoustical monitoring efforts in national parks. The Natural Sounds Program was established to manage acoustical environments in a way that balances access to parks with their long-term preservation as well as expectations of visitors. This program provides services to parks in the form of recreational planning assistance, acoustical data collection and analysis, and research projects in the areas of acoustical and social sciences. These services ultimately advance the overreaching purpose of national parks, which is to conserve the natural state of resources therein.

Establishing a scientific basis for the state of acoustical resources involves determining the composition of national park soundscapes in terms of commonly (and simultaneously) occurring man-made (e.g., aircraft) and natural (e.g., weather effects) sources. To accomplish this task, monitoring stations are typically deployed in certain locations in parks for months at a time to constantly record the soundscapes. However, due to the complexity of the soundscapes [22], the current approach to source characterization involves manual post observation and evaluation of large volumes of acoustical recordings. This process can be tedious and prohibitively time-consuming, which results in circumstances where the majority of the data remains unanalyzed, thereby limiting the ability of the NPS to achieve its assessment goals. Therefore, the developments in this thesis were designed to prevent such manual analysis.

Historically, a particular type of lossy data reduction has been used for efficient storage of acoustical information. This data reduction involves transforming consecutive windows of the streaming audio data by using a bank of filters and forming a "1/3 octave data vector" (see Appendix A) based on the energy in the filtered waveforms. Although storage limitations are no longer a large concern, the research in this thesis remains focused on solutions that suit the 1/3 octave data format so that a unified approach for analyzing new data, as well as data collected in the past, may be used. Furthermore, the NPS intends to implement one of the automated source characterization algorithms directly on monitoring stations for future deployment in national parks, and the computational savings offered by methods that operate on a lossy data format make this goal more realistic. Finally, developing methods to operate on sequential multivariate data allows for their application to other types of problems (such as those mentioned above) that inherently use this general data format. On the other hand, as detailed in Chapter 2, performing source characterization tasks using this data format presents a unique set of challenges above and beyond those associated with using raw audio data for the same purpose.

The remainder of this chapter is organized as follows. Below in Section 1.2 a survey of related previous work on transient source characterization is provided. Section 1.3 provides a comprehensive list of properties of effective solutions that are deemed necessary for conducting automated soundscape analysis. An outline of the objectives and contributions of this research follows in Section 1.4. Finally, an overview of the organization of this thesis is presented in Section 1.5.

1.2. Survey of Previous Work

This section presents a survey of previous research on the fundamental capabilities required to perform characterization of time-varying transient sources using sequential multivariate data. Work in the broad field of transient detection is first discussed, followed by methods for signal classification. Lastly, a survey on comprehensive systems that attempt to offer solutions to similar source characterization problems is presented.

1.2.1. Transient Detection

Transient detection involves determining if and when an abrupt change to some characteristic properties of the data occurs. This definition is vague, since what constitutes a change of interest may vary according to the application. In some cases, changes are not necessarily observed directly, but instead may be associated with some set of latent variables or parameters. Moreover, the temporal characteristics of a transient are application specific, e.g., it may be assumed the data changes permanently, or decays monotonically after a rapid onset. In this thesis a transient source is one whose signatures are not continually present in the data, but appear at some unknown time k_1, and cease to be extant at some other unknown time $k_0 > k_1$.

Although transient detection is generally used to constantly monitor an incoming stream of sequential data, some approaches analyze an entire data segment to determine whether a relevant change occurs at some point within [23–28]. For instance, the method in [23] applies a wavelet packet transform to an entire time series and identifies transients using the likelihood of a Hidden Markovian Tree that models transform coefficients. The method in [24] detects multiple transients by forming a likelihood ratio for a known family of transient parameterizations, given the entire data sequence. While these methods can potentially be applied to segments extracted from a sequential data stream, such as those used for the present soundscape characterization problem, this approach is unpractical since frequent splitting of acoustical events between segments is likely. Additionally, since the data in Chapter 2 has a low temporal resolution (one observation per second), using such approaches can result in a significant lag between the time an event occurs and when it is reported, since several hours of data would be needed before it could be analyzed. Some other methods assume multiple realizations of the data are available, i.e., that the environment is recorded using an array of microphones. For example, the work in [29] considers detection of a tapered, complex, transient oscillation observed coherently in two time series. Detection is performed by calculating a normalized cross-wavelet spectrum between the two time-series to perform a binary hypothesis test. Such methods cannot be applied to the soundscape data in this thesis, since only a single microphone is used to generate one realization of the acoustical scene.

Another very common approach to transient detection is to simply look for anomalies in the data based on some model of quiescent observations [30–32]. For instance, in [30] arbitrary transients are detected in audio waveforms by evaluating time-scale coefficients of wavelet atoms, and assuming the contributions of transients are characterized by a monotonic decay in the energy of these coefficients over time, regardless of their source. Another study [32] exploits the distribution of individual time-frequency components in a spectrogram, which is then transformed to a space appropriate for segmenting time-frequency slices containing noise only from those containing noise plus an unknown signal. However, for the present application, detection of arbitrary changes is not the goal, since the soundscape data is typically subject to continuous variations owing to the intermittent presence of many sources and phenomenon (see Chapter 2). For example, it is sometimes desirable to detect only signals (sources of extrinsic sound), while ignoring interference (sources of intrinsic sound).

Due to the high variability of signatures associated with any given source considered in this thesis, they can generally be considered as random. Consequently, the idea here is to look for a change in the model parameters of the data. The prevalence of fault detection applications, e.g., for quality control in manufacturing, has motivated the development of many transient detection methods that operate by estimating an unknown time when the parameters of the data change, using scalar observations and stopping after detection. Comprehensive coverage of this particular type of problem under various assumptions about sources of interest and environmental properties can be found in [33]. Perhaps the most fundamental tool for this type of transient detection is the classical Page's Test [34], also known as the cumulative sum (CUSUM) procedure, for determining a change from one independent distribution to another. Page's Test has min-max optimality in terms of average run length of the test, meaning it minimizes the worst case delay to detection given a constraint on the average delay between false alarms.

Despite the optimality of CUSUM, the assumed independence of observations prevents application in its original form to problems involving sources with signatures that span several observations. In [35], a version of Page's test that operates on dependent observations was implemented using a pair of Hidden Markov Models (HMM) [36] to characterize the distribution of the observations before and after the unknown change time, thus allowing for less restrictive assumptions on the structure of the transient signal and noise. An alternative approach was presented in [37] that models observations after the change time using the superposition of an HMM that is always present with a new HMM. This approach has greater utility when the signatures of the transient source are occluded by persistent interference of a single type. In [38], the generalized likelihood ratio test (GLRT) was reformulated for cases where the signal parameters are unknown, and the *a priori* signal distributions are used. While most of the mentioned methods are not designed to inherently perform simultaneous detection and classification, detect multiple transients within the same sequence, operate on multivariate observations, exploit temporal dependencies between observations, and/or offer robustness to interference, many of them can be extended to offer such features. The existing transient detection approaches that are most relevant to the soundscape characterization problem are discussed in more detail in Section 3.2.

1.2.2. Classification Using Sequential data

Once a transient source has been detected, it must be assigned a class label, which here denotes the type of acoustical source that produced the detected event. If detection considers both signal and interference sources, both may be assigned labels. Another common scenario is designing a detector to only be sensitive to signal sources, e.g., by filtering out interference or incorporating robustness capabilities, in which case only signal labels are assigned. The discussion here will mostly focus on the latter case.

The main difficulties associated with signal classification in the present application are that a single event often spans multiple adjacent observations, has unknown duration, and the types of interference that are simultaneously present may change. A rather simple way to perform classification under such circumstances is to analyze and assign a label to each observation within a detected event separately [15, 39, 40], and subsequently fuse the preliminary decisions [41–43] to yield a consistent event-wide class label. The matched subspace classifier [40, 44] is a common choice for assigning labels to vector observations, as it consists of a set of uniformly most powerful detectors and can operate in the presence of known subspace interference, though it assumes signals are deterministic but unknown. A subspace classifier that is robust to non-Gaussian noise [40] can be formed by weighting the residuals between a set of observations and their corresponding subspace approximations to minimize a user-specified discriminant function. Gaussian mixture models (GMM) can also be used for single-observation classification, which are common for applications involving environmental sound [15], wildlife call [13], and speech [45] recognition. It is the ability of GMMs to model arbitrary distributions with multiple modes that makes them suitable for recognizing features extracted from inconsistent acoustical signatures that are a mainstay of these applications. The main disadvantage of this type of approach is that the temporal dependencies between source signatures present in different observations is not directly exploited, which can offer significant discriminatory information. This problem may be alleviated to a degree through the use of template-based methods, e.g., the spectral-band matched filter in [46], especially if dynamic time-warping is used [12]. However, these latter methods typically require making even more limiting assumptions about the structure of acoustical events, due to the inflexibility of templates.

A more convenient way to perform classification using sequential data is to look for segments of data with similar properties so that a unified label may be assigned to a cluster of observations. One way this can be accomplished is by extending methods that looks for a change in the parameters of the data to accommodate an unknown parameter after the change [33]. The generalized likelihood ratio (GLR) algorithm [34] for composite or multiple hypothesis tests can be implemented as a direct extension of CUSUM, where the parameter associated with the null hypothesis is fixed, but that associated with the alternative hypothesis is unknown and assumed to be a member of known set of parameterizations. The different alternative hypothesis parameters correspond to different source types, and simultaneous detection and classification can be performed by replacing the unknown parameter with its maximum likelihood (ML) estimate, i.e., the assigned label corresponds to the ML source type. A variation of this approach, known as the weighted CUSUM algorithm [47], uses the cumulative distribution function of a probability measure for the unknown parameter to weight the likelihood ratio w.r.t. all possible values of the alternative parameter, thus incorporating *a priori* class probabilities. The main drawback to these CUSUM-based approaches is that there may be many parameterizations of the data under the alternative hypothesis, especially when multiple types of interference are intermittently present, leading to potentially frequent parameter switching that necessitates reinitializing the detection test statistic(s). Furthermore, there is no standard procedure for extending these methods to constantly detect and classify multiple sources.

Sparse representations have recently been employed [48–53] for performing detection and/or classification from multivariate observations by using only a few atoms from an overcomplete dictionary to represent sources of interest [54]. In [51], separate dictionaries are learned using K-SVD [55], that are capable of sparsely representing different classes of audio signals, and a support vector machine (SVM) is used to directly classify sparse coefficient vectors. A related approach is to use a dictionary that consists of training templates for different classes and assign a class label based on which sparse subset of templates provides the smallest reconstruction error. This approach was adopted in [48] for face recognition and extended to handle multiple observations in [49], though the latter assumes consistency of sparse representations among different observations. However, these methods process either a single observation or an ensemble of observations simultaneously, and hence, may be insufficient for continually detecting and classifying multiple signals using sequential data.

A natural way to extend methods that perform classification in a sparse domain to handle sequential multivariate data is to model the dependencies between atom coefficients [56–59] extracted from different observations. For instance, [56] proposes a general framework for modeling the structure of wavelet coefficients for natural signals using HMMs. Dependencies between the wavelet functions at different scales and translations are modeled using a Hidden Markov Tree and a Hidden Markov chain, respectively. The work in [58] performs source estimation from an audio waveform by modeling sparse coefficients in a way that exploits their prior distributions as well as their dependencies through the use of activity variables, where the latter also controls sparsity. This coefficient modeling concept can be used to extend many classification methods that are based on sparse coefficients, though most existing work only applies to time series data, does not consider simultaneous detection and classification, and/or does not consider the presence of interference.

1.2.3. Comprehensive Acoustical Source Characterization Systems

As detailed in Section 2.2, source characterization using natural soundscape data carries many unique challenges and, prior to the development of the methods introduced in this thesis, few if any approaches existed that simultaneously address all of them. The approach in [60] is perhaps the most applicable since it also considers environmental sound classification and is based on the 1/3 octave data format. In this work, a self-organizing map is used along with a locally-excitatory/globally-inhibitory oscillator network to identify co-occurring sound features extracted from a spectral representation and group sound fragments, respectively. The main disconnect relative to the goals in this thesis is that [60] detects any novel acoustical event, rather than specific sources of interest. Furthermore, the simultaneous presence of multiple sources is not considered, and each 1/3 octave vector is analyzed separately, meaning the overall structure of an entire acoustical event is not exploited.

Many studies attempt to perform source characterization using common generalized approaches [1, 7, 11, 12, 15, 20], such as those described above. HMMs are frequently used to directly model acoustical signatures or other types of data sequences with high variability [1, 8, 12, 56] due to their flexibility and exploitation of observation dependencies. In [1] mixture autoregressive HMMs are used to perform speech recognition, while [12] compares the effectiveness of HMMs with dynamic templates for recognizing bird song elements. The work in [15] applies a GMM (see Section 6.2) to features extracted using a matching pursuit framework [61] to perform environmental sound recognition, while in [11] artificial neural networks are applied directly to features extracted from acoustical waveforms for animal identification. As is evident from the application areas addressed by each of these studies, generalized methods typically fail to account for all of the complexities of natural soundscape data (e.g., the simultaneous presence of multiple source types), and hence, they usually only perform well on simple data when used by themselves. Alternatively, a significant amount of preprocessing can be performed to condition the data such that certain assumptions are satisfied, e.g., interference-free.

Many approaches have been developed that take a more targeted approach to acoustical source characterization [2–4, 9, 13, 16], by developing feature extraction, detection, and classification capabilities that directly exploit the properties of sources of interest. Methods for robust speech recognition are proposed in [2–4], where each study applies a well-known classification framework (e.g., HMM) to unique feature sets. For instance [3] proposes a peak selection method, as well as a new integration method for extracting periodicity information across different frequency channels, while [2] extracts features that are a combination of subband power and dominant subband frequency information. Bird species recognition is performed in [13] by posing the problem as one of parameter estimation. Three different parameterizations are compared including sinusoidal modeling, Mel-cepstrum parameters, and a vector of various descriptive features unique to each species. Although none of these methods account for the simultaneous presence of multiple types of interference, they are also fundamentally incompatible with the 1/3 octave data considered in this thesis since they presume access to the raw audio data.

Since time-frequency data representations (such as 1/3 octave vector sequences) are commonly used for acoustical source characterization tasks, several approaches have been developed specifically to operate on this type of data [5, 6, 8, 10, 18]. One common approach to such problems is to identify novel portions of time-frequency data based on certain features or statistical properties. The work in [5] identifies speech patterns using 2-D patches that are extracted from spectrograms and projected onto a 2-D discrete cosine basis. In [8], a framework is presented for automated detection and classification of sounds from birds, crickets, and frogs using a HMM to represent sequences of statistical features extracted from spectrograms. Classification of bird vocalizations is performed in [10] by developing a sum-of-sinusoids model for each vocalization type and calculating the degree to which parameters extracted from the data match a set of stored templates. Despite being somewhat compatible with the 1/3 octave data format, most of these methods are incapable of recognizing any general source type and are not robust to the presence of structured interference; such capabilities are essential for the problem at hand.

1.3. Objectives and Requirements

Clearly, there are many possible approaches for accomplishing the required source characterization tasks. Before explaining the contributions and contents of this thesis, it is helpful to establish those traits that are beneficial and/or necessary for a method to possess in order to ensure it is capable of offering acceptable performance when applied to the natural soundscape data. In no particular order, these traits are:

- Operates on sequential data, both for efficient processing of long data segments in post-mission analysis scenarios, and for direct implementation on acoustical monitoring stations for in-situ soundscape characterization.
- Inherently takes advantage of the multivariate nature of the data sequence by fully exploiting all available information within an observation, as well as the dependencies between them.
- Exploits the random and time-varying nature of source signatures in order to remain robust to significant within-class diversity and inconsistent signatures that are a mainstay of acoustical sources in natural environments.
- Accounts for the simultaneous presence of signal and interference, leading to superimposed signatures, either by treating the latter as a unique source to be characterized or remaining robust to its effects, thereby maximizing performance of the system for detecting and classifying signals.
- Capable of recognizing any general source type given adequate training data, as opposed to being designed from the ground up to exploit specific properties of certain sources that are difficult to generalize [5, 8, 10, 13]. This capability is especially important in light of the fact that different sites that are monitored by the NPS often contain different types of signals and interference, and hence, the system must be easily adaptable to unspecified environments.
- Capable of performing transient source detection and classification using a cohesive framework. As discussed in Section 1.2.1, a wide variety of techniques exist for performing transient detection alone, but many of these cannot be easily modified to incorporate classification capabilities, without simply appending a disjoint classifier.

As outlined in the next section, the contributions of this thesis were all developed with the intension of meeting or exceeding these requirements.

1.4. Contributions of the Present Work

1.4.1. A New Sequential Random Coefficient Tracking Method

As mentioned in the previous section, there is no prior work in the area of source characterization that satisfies all of the requirements for effective solutions. This is in spite of the fact that many real source characterization problems are also burdened with similar challenges, e.g., variable source signatures and heavy interference, due to uncontrolled environments. Therefore, the primary goal of this work is to develop methods that indeed meet the mentioned requirements and can be realistically implemented on acoustical monitoring stations for source characterization in national parks. First, a sequential random coefficient tracking (SRCT) framework is developed that applies a hierarchy of log-likelihood ratio tests to individual observations to determine their composition in terms of different source signatures. These signatures are modeled as random to capture variability between different events associated with the same source type. The parameters of the test statistics are generated by incorporating information from previous observations, thereby exploiting known dependency models. This allows for detection and classification of one type of signal and one type of interference that may be simultaneously present within an observation. The SRCT method also performs separation of the signal and interference portions of the measurement and produces estimates of their signatures in isolation; a property that is not shared by other methods mentioned before.

Since the SRCT method assigns class labels to each observation separately, this thesis also develops a HMM-based sequential decision fusion framework both to reduce detection/classification errors and to provide event-wide class labels, as desired by the NPS. This process finds the likelihood that a certain signal type is present, given a sequence of preliminary labels, using a procedure that is based on the CUSUM-like method proposed in [35]. This method is general in that it may be applied to a preliminary decision sequence generated by any sequential detection and classification framework. When applied to decision sequences made by classifiers that do not consider information in temporally adjacent observations (e.g., GMMs), this fusion provides a means to incorporate dependencies between the decisions.

1.4.2. A New Sparse Coefficient State Tracking Method

While the SRCT method performs very well when applied to data that adheres to the assumed source model, it has a few shortcomings that lead to reduced performance in certain situations. Specifically, the SRCT method assumes that only one type of interference may be present at a time. Additionally, the assumed subspace model may not be appropriate for some source types, especially for other similar applications. Lastly, since the SRCT method assigns class labels to individual observations separately, decision fusion must be performed to obtain event-wide class labels, which makes in-situ implementation of this method more difficult (though still feasible). To address these concerns, this thesis introduces a sparse coefficient state tracking (SCST) method, which also meets all of the requirements in Section 1.3. This approach draws from the concepts of classification in a sparse domain and modeling of sparse atom coefficients to yield a cohesive framework that places very few restrictions on the structure of the observations. The main advantage of this method is its applicability to data containing signal, interference, and noise components that may not necessarily follow models based on convenient parametric distributions, e.g., multivariate Gaussian. Instead, the idea is to simplify the data representation for realistic and accurate modeling and likelihood calculation. The SCST method can operate on observations that contain multiple types of interference, and inherently detects and classifies entire acoustical events without the need for decision fusion. The main disadvantages of the SCST method compared to the SRCT method is that the former is not capable of detection and classifying interference sources, and it is more computationally intensive.

The benefits of the proposed algorithms for detecting, classifying, and estimating the signatures of sources of interest are demonstrated using real NPS soundscape data. Results are presented in terms of the receiver operator characteristics of test statistics associated with each method, their overall ability to detect and classify entire acoustical events, and their ability to improve the SNR of 1/3 octave signal events corrupted with interference. Both the SRCT and SCST methods demonstrate exceptional performance in all aspects of the benchmarking, though each exhibits particular strengths and weaknesses. Importantly, these experiments represent the first successful comprehensive application of any source characterization method to an NPS soundscape data set, which is a testament to their ability to simultaneously perform a combination of functions that cannot be claimed by any other approach.

Development of the proposed source characterization methods also involved a rigorous analysis of the 1/3 octave data format collected by NPS acoustical monitoring stations. For this reason, this thesis also presents what is perhaps the first comprehensive analysis of the properties of 1/3 octave data, with a focus on its utility for representing natural acoustical environments. A general discussion on the benefits and deficiencies of this data format is presented in Section 2.2, while more detailed and rigorous information that is not essential for understanding the proposed source characterization algorithms is presented in Appendices A–C. This information is included since many audio scientists use 1/3 octave data representations to perform source characterization for reasons that are detailed in Section 2.2, and yet very little is understood about the properties of this data format as they pertain to developing rigorous detection and classification algorithms.

1.5. Thesis Organization

This thesis is organized as follows. Chapter 2 describes the acoustical monitoring stations and process used to collect the 1/3 octave soundscape data. This chapter also discusses the properties of the two data sets used to generate experimental results. Chapter 3 provides an introduction to the main concepts and approaches associated with detecting and classifying nonstationary acoustical sources using sequential data, both to introduce the fundamental ideas in the context of relative simple problems and to provide material that can be used as building blocks for constructing more comprehensive solutions. The proposed SRCT and SCST source characterization frameworks are detailed in Chapters 4 and 5, respectively. In Chapter 6, comprehensive results are presented that are obtained by applying the SRCT, SCST, and benchmark GMM-based methods to the two soundscape data sets. Finally, Chapter 7 concludes the studies carried out in this research and discusses possible directions for future work.

CHAPTER 2

Soundscape Data Collection and Properties

2.1. Introduction

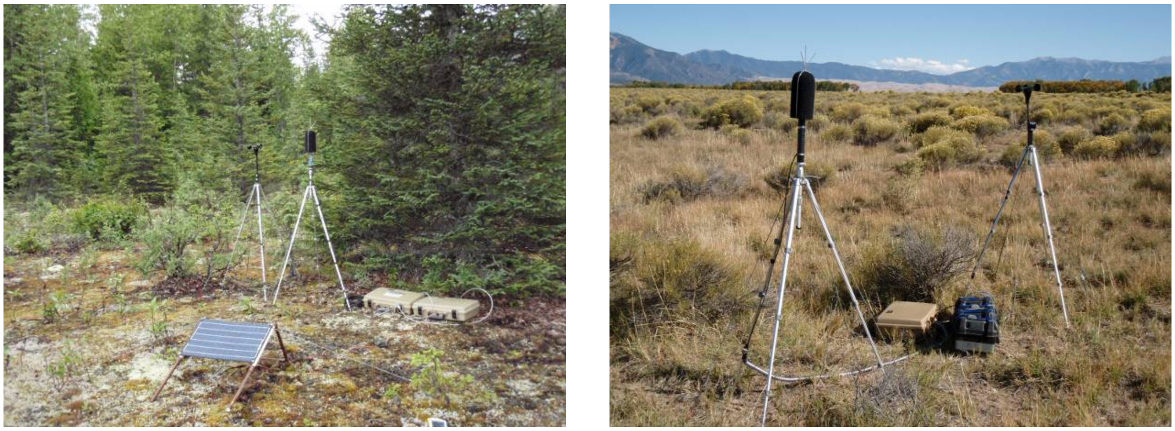
To characterize a diverse set of natural soundscapes, the National Park Service (NPS) deploys numerous acoustical monitoring stations in various remote settings throughout the U.S. parks. Currently, these monitoring stations collect single channel acoustical information stored as both compressed audio waveforms (e.g., MP3) and as a representative but lossy 1/3 octave vector sequence [62], where for older missions only the latter format was collected due to storage constraints. Soundscapes are monitored for extended periods of time — often months — before the stations are retrieved so that the recorded data can be manually analyzed. This chapter discusses all of the relevant information pertaining to the soundscape monitoring process and the resulting data. This is important because this soundscape data strongly motivated the development of the source characterization methods discussed in this thesis, and understanding the data collection system and monitored environments are instrumental in developing algorithms capable of providing automatic, accurate, and robust detection and classification of acoustical sources (mainly signals).

The outline of this chapter is as follows. Section 2.2 discusses the characteristics of the current acoustical monitoring setup, the data collection process, and properties of typical source types as well as their interactions in the acoustical environment. Section 2.3 provides details unique to the two data sets used in this study to generate experimental results, including a discussion on the types of sources captured, specific characteristics of the soundscapes, and associated challenges inhibiting successful signal detection and classification. Conclusions are then made in Section 2.4.

2.2. Overview of Acoustical Monitoring

2.2.1. Data Collection Process

To collect acoustical data for a given soundscape, a team of NPS employees physically transport a monitoring station to a desired location within a park. The primary goal for site selection is to use an area that is representative of the main attributes of an acoustical zone within a park, i.e., an area with unique acoustical properties owing to the presence of specific types of wildlife, weather patterns, air traffic density, etc. Section 2.2.3 presents examples of acoustical zones and variations between them. A site that has plenty of sunlight for solar panel use, and one that provides reasonable protection from high winds, is also desirable. An acoustical monitoring station consists of the following equipment: 1) Larson Davis 831 sound level meter [63] 2) single microphone with environmental shroud 3) preamplifier 4) ten 12 V lantern cell batteries 4) anemometer 5) MP3 recorder and 6) meteorological data logger (e.g., wind speed from the anemometer). Photos showing deployed acoustical monitoring stations can be seen in Figs. 2.1(a) and 2.1(b).



(a) KEFJ004 monitoring station.

(b) GRSA001 monitoring station.

Figure 2.1. Photos of the acoustical monitoring stations that collected the data used in this study.

The Larson Davis 831 sound level meter [63] is perhaps the single most important piece of equipment in the current data collection process, as it generates a 1/3 octave vector sequence

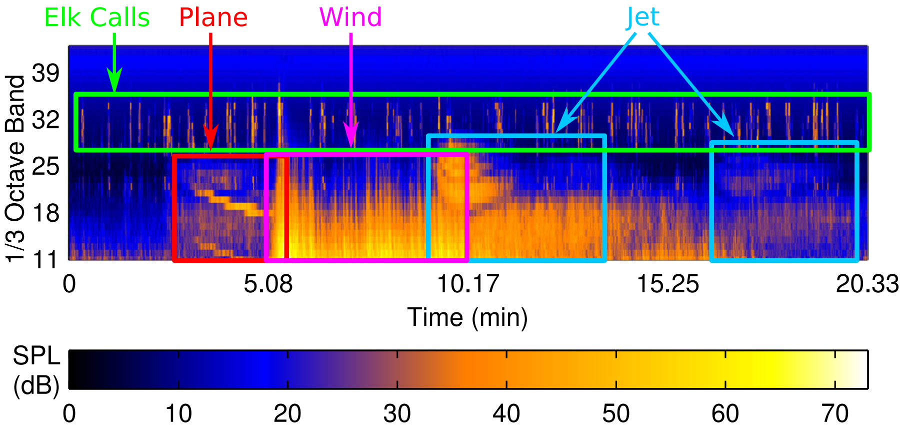


FIGURE 2.2. Example 1/3 octave data sequence and representation of some common acoustical events.

[62] from observed acoustical waveforms and stores it for subsequent analysis. More specifically, a 1/3 octave vector is extracted from every non-overlapping one-second time segment and has $N = 33$ elements that represent the average energy in different 1/3 octave frequency bands for the corresponding one second interval. An example of a 1/3 octave vector sequence captured by a NPS monitoring station is shown in Fig. 2.2, which also highlights a few signal and interference sources that are captured. Note that the vertical and horizontal axes in this figure represent frequency band (from lowest at the bottom to highest at the top) and time, respectively, while the color of each pixel in the representation denotes a specific sound pressure level (in dB). These display conventions will remain consistent throughout this thesis. More explicit details on this data format can be found in Appendix A. Note that storage of compressed audio (MP3 format) has recently become feasible, but this data format is not used in this thesis for reasons mentioned earlier in Section 1.1.

The 1/3 octave data format was adopted by the NPS for several reasons, namely because it is representative of how humans perceive sound, i.e., pitch is perceived as changing with the ratio of frequencies, rather than a linear increase [62]. More importantly, 1/3 octave data provides a convenient data reduction and more efficient representation for manual post-analysis of soundscape data. This is due to the fact that most common signal and interference sources tend to have

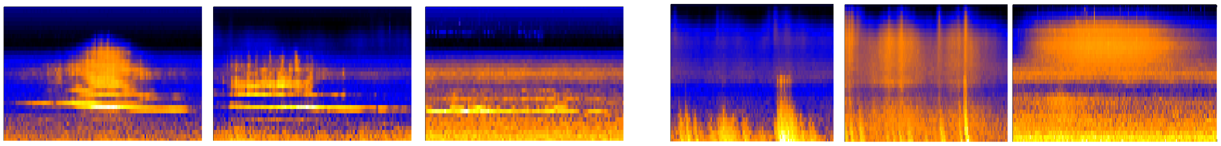
1/3 octave signatures that are easily distinguishable by humans, and is important since manual analysis was the only reliable way to analyze the data used in this thesis before the advent of the algorithms developed in Chapters 4 and 5. Another primary motivation behind using 1/3 octave data for soundscape characterization is a significant reduction in required storage capacity when compared to raw audio, i.e., 33 rather that 51,200 [63] samples are recorded every second. Beyond storage advantages, it is clear that it takes significantly fewer computational resources to process a 1/3 octave vector when compared to raw audio, which is especially important for sensor-level processing needed for analyzing sequential data in the field.

On the other hand, a typical approach to processing raw audio would be to first extract a set of salient features to represent each time segment, that are optimized for the intended tasks. In contrast, the samples that are retained in the 1/3 octave format are not necessarily the 33 features that are most useful for detection, classification, or estimation of transient acoustical sources. First off, as explained in Appendix A, the bandwidth of each 1/3 octave band grows exponentially with frequency [64], and hence, the low frequency resolution in each vector is far higher than the high frequency resolution. While this property is not inherently unfavorable when using orthogonal transforms such as wavelets, it may be detrimental for characterization of some sources in this study whose signatures are condensed to one or two bands, despite the fact that their signatures may undergo subtle variations that could be exploited. For instance, the first element of each 1/3 octave data vector collected by the sound level meter has a center frequency of 12.5 Hz and a bandwidth of 2.9 Hz, whereas the last element of each vector has a center frequency of 20 kHz and a bandwidth of 4.6 kHz. Since many of the sources of interest (signals) have signatures that lie within the mid-frequency range around 100 Hz - 1000 Hz (see next subsection), and the low frequency signatures of signals are often obstructed by interference, there is a significant amount of information in each vector that has very limited use for signal characterization. Looking at the bigger picture, the frequency domain might not be even the best elementary domain to work in for the mentioned tasks. Ideally, data reduction would involve the use of basis functions that provide separable representations of different source types to be characterized [65]. Finally, it has not been established that the average energy measure that each vector element represents [66] is indeed optimum for the tasks at hand. These concerns are exacerbated by the fact that the 1/3 octave transformation is obviously not invertible and many useful discriminatory features are lost in the process. Nonetheless, as shown in Chapter 6, 1/3 octave data does provide sufficient discriminatory information to use for the source characterization tasks.

2.2.2. General Source Properties

In order to successfully detect and classify sources of extrinsic sound (signals) in national parks, it is essential to understand the properties of their corresponding 1/3 octave signatures, as well as those of competing interference sources, that hinder these efforts. Note that, only generic properties of common signal and interference sources are discussed here, whereas the next section provides details on the sources that are captured in the data sets used for the performance evaluations presented in this thesis.

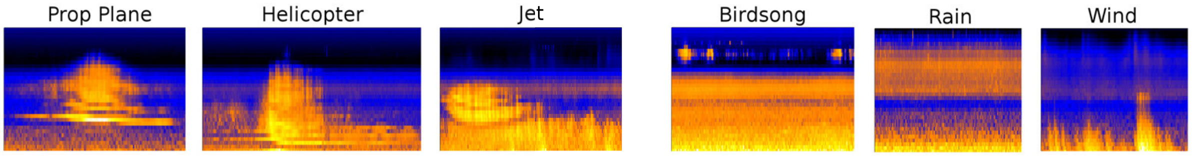
The primary challenge in soundscape characterization is that nearly every source type considered in this study (both signal and interference) emits nonstationary, highly variable, and often erratic 1/3 octave signatures. This causes within-class diversity where acoustical events associated with the same source type can be dramatically different, to the point where even an untrained human may have trouble making such associations. Examples of such scenarios can be seen in Figs. 2.3(a) and 2.3(b), for the propeller plane signal class and weather-related interference class (e.g., rain, thunder), respectively. As can be seen, each event shares some features with others within its respective category, but various phenomena (discussed below) cause glaring inconsistencies. For this reason, it is difficult or impossible to successfully employ techniques that assume deterministic source signatures.



(a) Propeller plane signal events.

(b) Weather-related interference events.

FIGURE 2.3. Examples of within class diversity.



(a) Signal events.

(b) Interference events.

FIGURE 2.4. Examples of 1/3 octave signatures of different source types.

Perhaps the most measurable cause of within class diversity for 1/3 octave events associated with signals is their movement w.r.t. the receiver which, depending on the speed and distance of the source relative to the receiver, causes a varying degree of nonlinear Doppler shift of the frequencies of the received acoustical waveforms. The specific impacts of Doppler are detailed in Appendix C but, suffice to say here, it is the cause of the momentary but severe increase in bandwidth near the middle of recorded events witnessed in many cases, as seen in Fig. 2.3(a). Of course, sources can have extremely different positions and motion parameters relative to the receiver, which not only causes variations in the Doppler effects described above, but also in the amplitude and duration of the signatures of such sources. The last event shown in Fig. 2.3(a) likely displays limited bandwidth variability because its trajectory was such that its velocity w.r.t. the receiver did not change drastically within a short time interval.

There are many additional causes of nonstationary signatures for signal sources, many of which are unique to specific signal classes. These signature characteristics are exemplified by the events shown in Fig. 2.4(a). For instance, some signal sources, such as helicopters, emit highly directional sounds so a received waveform may gradually increase in amplitude as the source approaches the receiver, but die off quickly once the source passes the receiver. Since many signals of interest are propeller-based aircraft, a significant portion of the acoustical signatures they produce is of aerodynamic origin due to the flow of air around the blades, which produces harmonics and has very complicated dynamics in general. Helicopters in particular have more complicated acoustical signatures due to blade-vortex interaction, where their blades create turbulence as they pass behind one another in their own wakes.

Perhaps the main challenge of characterizing, or being robust to many types of interference sources is that they emit erratic signatures owing to the lack of a consistent operational mode. For instance, as seen in the example in Fig. 2.4(b), individual bird song calls often appear similar, but the overall order or spacing between successive calls in a song can be unpredictable. This problem is exacerbated by the 1/3 octave data format since bird calls are often short in duration (only 5–10 ms for some species), so the signatures of either many calls or just a few may be present in a single vector. Similar behavior is witnessed for thunder, as shown in Fig. 2.3(b), owing to the lack of predictability of its occurrence and properties. Another example of an erratic natural sound is rain, which often inconsistently changes in intensity despite the fact that its frequency characteristics are often very consistent compared to most other interference types.

Further complications arise from severe between-class similarities, i.e., when events associated with different signal classes appear similar in many respects. Fig. 2.5 demonstrates such a situation by showing propeller plane and helicopter events that share many of the same features, e.g., the consistent high energy in the eighth and ninth 1/3 octave bands, and the weaker broadband energy present in approximately the first 2/3 and last 1/2 of the plane and helicopter events, respectively. While the low frequency components present in the latter are enough to classify this event as helicopter, the spectral signatures of both events overlap in many other bands, meaning there are limited features that may be used to distinguish them. As discussed in more detail below, the problem becomes even more challenging when these limited discriminatory features are obscured by competing interference. In any case, given that the signatures of two different types of sources

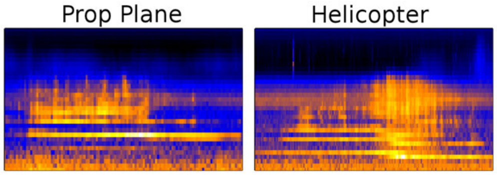
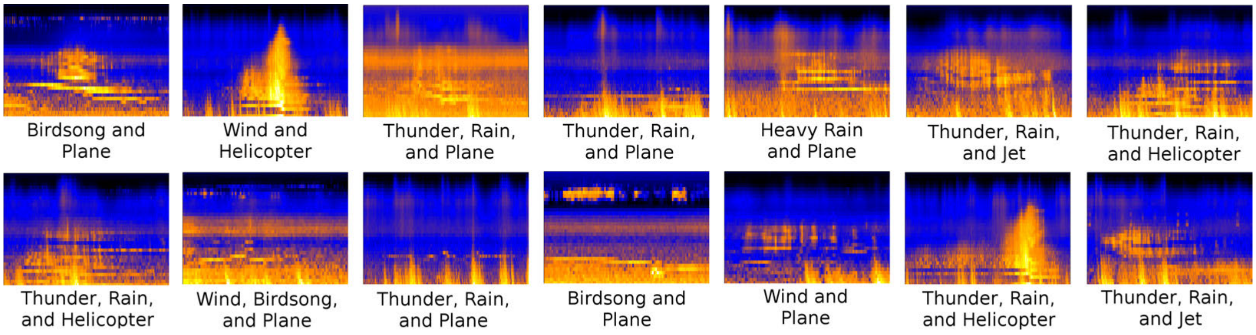


Figure 2.5. Example illustrating potential similarities between propeller plane and helicopter signatures in the 1/3 octave domain.

can be identical for a subset of the vectors that are contained within their respective events, it is often the case that successful source characterization cannot be accomplished by analyzing each 1/3 octave vector independently. Instead, the composite temporal structure of each 1/3 octave event must be considered.

2.2.3. Soundscape Compositions

Beyond the properties of specific source signatures, the unique aspect of NPS soundscape data is the complex interactions between sources and the acoustical environments they are present in. First off, it is clear that most if not all sources have unpredictable times of arrival (ToA), hence the utility of sequential detection methods. Additionally, multiple acoustical events are frequently concurrent within a soundscape. The simultaneous presence of a signal and some type(s) of interference is the most common scenario, as the latter source types are part of the natural environment and therefore ubiquitous. This is especially true for acoustical signatures resulting from weather effects and bird/insect calls, which persist regardless of the presence of extrinsic sources. Examples of 1/3 octave sequences containing the superimposed signatures of a signal and interference source are shown in Fig. 2.6. As can be seen, sometimes the presence of interference has little impact on detection and classification of signals, such as the simultaneous presence of birdsong and a plane with little associated Doppler shift (center bottom of Fig. 2.6). On the other hand, some circumstances lead to a very difficult source characterization problem, such as when the signatures of a jet and thunder are superimposed, as they both tend to occupy the low-mid frequency bands. Clearly, a successful approach to this problem must be capable of characterizing a given signal both


Birdsong and Plane

Wind and Helicopter

Thunder, Rain, and Plane

Thunder, Rain, and Plane

Heavy Rain and Plane

Thunder, Rain, and Jet

Thunder, Rain, and Helicopter

Thunder, Rain, and Helicopter

Wind, Birdsong, and Plane

Thunder, Rain, and Plane

Birdsong and Plane

Wind and Plane

Thunder, Rain, and Helicopter

Thunder, Rain, and Jet

FIGURE 2.6. Examples of superimposed signal and interference.

in the presence *and* absence of different types of interference that might reasonably be present at the same time. Fortunately, for many data collection sites it is highly unlikely for certain pairs of signals to be present at the same time due to the rarity of one or both types, e.g., the simultaneous presence of a jet and helicopter in Kenai Fjords, Alaska. Soundscape characterization can be simplified by excluding such possibilities, since detecting the simultaneous presence of two signals of interest amidst the other mentioned complications is a very difficult problem.

Also of significance are changes in acoustical conditions owing to operational and environmental variations between different data collection sites, as well as within data collected at a single site. Ambient background noise (see Appendix B), defined as the components of an observation vector that are not associated with signal or interference sources, is caused by light wind, water flow, sensor noise, or any phenomenon that is continually present and mostly random, and as such has a continuous impact on the data. Such noise not only inhibits proper source characterization, but is also subject to statistical variations. As a rule, it can be safely assumed that little variation in the ambient noise exists locally at a given site within a time span of say, a few hours, but slightly more variations can be expected across a larger span of time at a given site. Naturally, one can expect larger noise variation between different sites at the same park, and a huge amount of variation between sites associated with different parks.

There are also different source types to consider at different sites and especially different parks. For instance, there may be two acoustical monitoring stations located within different zones in a single park, e.g., one in a forest, and the other in an alpine setting. Not only will different types of wildlife be present in these two environments, but also the signatures of the same source may be different owing to factors such as varying elevation and dampening medium such as a forest canopy, or different types of soil. Interference caused by strong winds blowing on the microphone baffle is also far more common in an alpine setting due to more frequent inclement weather and a lack of objects that obstruct airflow.

For sites from different parks, entirely different source types often need to be considered. For instance, in Everglades National Park, FL there are many types of insects and other wildlife that are not typically present at other parks. At Kenai Fjords National Park, Alaska certain types of propeller planes can often be heard that are otherwise rare, since they offer utilities (e.g., supply drops) that are typically not needed in other places. If every source type from every park (especially interference) was a candidate for each witnessed event then the classification task would be difficult if not impossible, since there would be an increasingly diminished set of features that could be used to distinguish between the signatures of different source types. Thus, an enlightened approach to this problem considers the fact that only certain source types may be observed within a given park and the system used for analysis designed accordingly.

Finally, it is possible that new sources not accounted for when training the system could be encountered. Such sources may include those that are present infrequently in the vicinity of a particular site, or those that are obscure or erratic enough that their consideration is not worth the added complications. An example of the latter type of source is human speech for very remote backcountry sites which, apart from the frequency range of its signatures, is highly erratic in most respects. Discretion concerning which sources are worthwhile to model is key here, by considering factors such as how frequently a source is expected to appear, based on historical data and/or intuition about a particular acoustical zone, and the potential for an increase in classification errors (and computational complexity) due to the inclusion of an additional source model.

2.3. Experimental Setup

In this thesis, two separate data sets are used for the development, testing, and analysis of the proposed source characterization methods. Although identical acoustical monitoring stations were used to collect both data sets, the noise variations between parks, as well as the differences in source types that are encountered, implies that the two data sets need to be analyzed separately using systems that are trained to operate in the appropriate local environment. The specifics of these two data sets are discussed below.

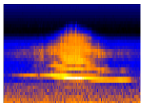
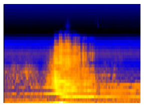
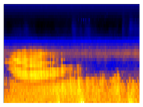
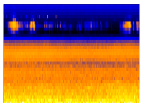
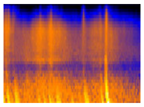
2.3.1. Kenai Fjords Site 4 Data Set and Properties

The first data set contains 1/3 octave vector sequences representing recordings of a soundscape associated with a relatively remote site within Kenai Fjords National Park, Alaska. The NPS refers to this data set as "KEFJ004", where "004" corresponds to the site number (out of four total), which is the label that will be used throughout the remainder of this thesis. A photo of this data collection site is displayed in Fig. 2.1(a), which shows that the acoustical monitoring station was deployed in a forested area. What this photo does not show is that site four is located relatively close to a river, but very far from roads or other infrastructure. In total, the soundscape was recorded for approximately 19 full days from July 22nd – August 15th, 2008, where no data was collected from August 1st – 6th due to the monitoring station being damaged in a bear attack. As a reminder, a single 1/3 octave vector was recorded every second, meaning a day of data consists of 86,400 observation vectors.

The types of signal and interference sources that were frequently captured in KEFJ004 are listed in Table 2.1, along with brief descriptions of the general structures of their corresponding 1/3 octave signatures, durations of typical associated events, and example events for each source type. By far the most frequently occurring type of signal events were caused by propeller planes, which commonly operate in the Alaskan wilderness for transportation and to drop off supplies for backcountry travelers with extended itineraries. The signatures of a stationary plane are normally confined to one or two 1/3 octave frequency bands, but are heavily influenced by Doppler when the plane is in motion, leading to the shift in frequency as the event progresses as well as the prominent broadband signatures (see Appendix C). Helicopters are also fairly common in the KEFJ004 data and have perhaps the most erratic signatures due to complicated mechanics that cause, e.g., the blade-vortex interactions described in Section 2.2.2. Helicopter signatures are also highly directional, leading to asymmetric 1/3 octave signatures, that can be exploited by certain source characterization methods. Unfortunately, apart from the broadband signatures present in the middle of a helicopter event, most signatures of this signal type are low-frequency, meaning they are susceptible to overlap from the most prominent types of interference whose signatures occupy the same frequency bands. The last type of signal considered for the KEFJ004 data is jet (of unknown type), which have a fairly low occurrence rate relative to the other signal types. Similar to helicopters, the signatures of jets are highly directional and predominantly low frequency, especially for the latter half of an event as the jet moves away from the sensors.

Only two types of interference are common enough to consider for the KEFJ004 data, namely birdsong and rain/thunder, where the latter also includes light wind since it appears similar to weak thunder in the 1/3 octave domain. The former typically introduces few complications to signal classification, mainly because signatures are exclusively high frequency, and no signals have signatures that lie only within these associated bands. On the other hand, rain/thunder presents significant challenges to the detection and classification of signals. Rain can resemble the broadband signatures within plane and helicopter events, while thunder tends to be similar to the low frequency signatures present in helicopter and jet events. This can lead to either false detection of signals when interference alone is present or missed signals when both signal and interference are present. Moreover, due to the climate of Kenai Fjords, rain persists for a large percentage of the soundscape recordings. Although the monitoring station is close to a river, the resulting acoustical signatures

Table 2.1. Characteristics of different source types in the KEFJ004 data set.

	Source	Typical Event Description	Typical Duration	Example
Signal	Propeller Plane	Signatures evolve from mid-frequency narrowband to wideband, and revert back to narrowband.	30–240 s	
	Helicopter	Signatures evolve from low-frequency narrowband to very wideband, and revert back to narrowband.	40–400 s	
	Jet	Starts with low-to-mid frequency signatures, with the mid-frequency signatures slowly fading approximately half-way through the event.	80–260 s	
Interference	Birdsong	Signatures are restricted to high frequency bands and have erratic temporal patterns.	1 s – several hours	
	Rain/ Thunder	Rain has mid-to-high frequency broadband signatures, while thunder is often superimposed with rain and adds impulsive, low-to-mid frequency signatures.	few seconds – several days	

are consistently present, and resemble elevated ambient background noise levels rather than a unique type of interference.

Very rarely, there were novel source types that contributed to the KEFJ004 soundscape, e.g., wildlife activity or human speech, that are generally not of interest due to their rarity and consequent low impact on the soundscape. Such source types are not highlighted in Table 2.1 or considered when training a system for source characterization since, as mentioned above, designing a system around sources that are possible but extremely rare adds significant complications to the source characterization tasks, with very few benefits.

Due to the complexity of the KEFJ004 soundscape, manual annotation of the data was previously the only available approach for locating and labeling sources. Therefore, such annotations existed before the development of the methods proposed in this thesis, and serve as the truth that is used to generate results in Chapter 6. In particular, two well-trained operators visually inspected the data to identify acoustical events associated with signals of interest, which are those listed in

Table 2.1, as they occur most frequently and prominently in this particular site. Only the presence of interference (not its type) was annotated since it was present a large portion of the time and such sources are viewed as a nuisance for the present application. The KEFJ004 data set also has corresponding compressed audio data (MP3 format) available for the same time period that 1/3 octave data was recorded. For reasons mentioned in Section 1.1, the 1/3 octave format is still used exclusively for source characterization in this thesis, though the raw audio was used to aid the annotation process by allowing the operators to hear the actual acoustical events, when needed.

In summary, above and beyond the mentioned challenges inherent with characterizing national park soundscapes using 1/3 octave data, there are several properties of the KEFJ004 data set that further complicate this task. These issues, in order of importance, are as follows.

- A large number of weak signal events are present throughout the data, possibly due to wildly varying trajectories of different aircraft that are present, and an apparent lack of obstacles to inhibit propagation of sounds from distant sources. Proper detection of such signals while maintaining a relatively low false alarm rate presents a significant challenge.
- A large variety of plane types, each with unique mechanics, leads to greater within-class diversity for plane events than would normally be encountered for any one signal source in a national park soundscape. This amplifies problems caused by within-class diversity and between-class similarities, that were mentioned above.
- Rain and thunder of varying intensity are present throughout a large portion of the recordings, leading to often severe overlap with signal events. These complications are exacerbated by the fact that most signals have 1/3 octave signatures that appear similar to rain/thunder for some portions of their associated events.
- Related to the above, the relative rarity of witnessing a signal event that is not superimposed with interference means there are fewer events to use for training, where the use of such "clean" events is often necessary.

- The forest canopy in the vicinity of the monitoring station leads to increased noise through interaction with falling rain.
- There are constant water flow sounds due to the monitoring station's close proximity to a river, which leads to a fairly high ambient noise level.

Clearly, these challenges indicate that successful automated analysis of the KEFJ004 data set requires development of robust methods that account for extreme variations in environmental and operating conditions. These complications are representative of those that are encountered for many data sets collected in different sites and parks, hence the reason KEFJ004 was selected for performance evaluations. On the other hand, KEFJ004 represents a relatively low-traffic park, where an average of only two or three signal events per hour is common. Nonetheless, the second data set introduced below contains some additional challenges, such as extremely frequent signal events, that are common for some other NPS soundscape data sets, e.g., Grand Canyon and Yosemite National Parks.

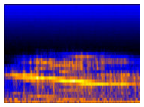
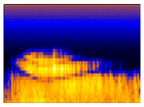
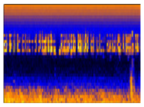
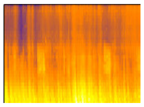
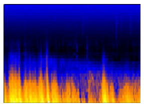
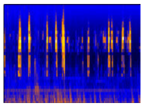
2.3.2. Great Sand Dunes Site 1 Data Set and Properties

The second data set contains 1/3 octave vector sequences representing recordings of a soundscape within Great Sand Dunes National Park, Colorado. Only one site (and one monitoring station) was used to collect acoustical data within this park, and hence, the NPS refers to this data set as "GRSA001", which is the label that will be used throughout the remainder of this thesis. A photo of this data collection site is displayed in Fig. 2.1(b), which shows that the acoustical monitoring station was deployed in a rather open and arid area. As with KEFJ004, the GRSA001 monitoring site is also far from infrastructure, thereby reducing susceptibility to, e.g., ground vehicle traffic, though the ambient noise level is generally lower in the GRSA001 data since no river is near the site. In total, the soundscape was recorded for approximately 17 full days from September 24th – October 10th, 2008.

The types of signal and interference sources that were frequently captured by this monitoring station are listed in Table 2.2, along with the same characteristic information that was provided for sources in the KEFJ004 data set. The two signal sources that are of interest for GRSA001, due to their frequent occurrence, are propeller planes and jets (of unknown type but likely commercial airline). Since these signal types were also of interest for the KEFJ004 data set, their properties will not be elaborated on here, though it is important to note that a typical plane event in the GRSA001 data has a slightly different structure from the typical KEFJ004 plane event, most probably because different models/makes are common in these two areas. The reason being that planes in KEFJ004 are generally used for utilitarian purposes (e.g., supply drops), whereas planes in GRSA001 are most probably used for hobbyist purposes (e.g., air tours). Also of importance is the fact that jet events are extremely common in the GRSA001 data; more so than the combined occurrences of all signal events in the KEFJ004 data set, though jet signatures present in both data sets appear rather similar. Conversely, plane events are rather rare in the GRSA001 data set.

As seen from Table 2.2, the types and behavior of interference sources in the GRSA001 data set represents the largest difference from the KEFJ004 data. Birdsong and rain are also present in the GRSA001 data, but are less common due to the drastic differences in climate and terrain (lack of trees). However, these interference types still occur frequently enough to warrant consideration by any source characterization method. Elk calls are unique to the GRSA001 data set and occur frequently since acoustical monitoring was performed during mating season. Such interference is similar in character to a broadband version of birdsong signatures, but is more disruptive to the detection and classification of signals since it is typically higher magnitude and broadband. Though strong wind was occasionally present in the KEFJ004 data, in short bursts it appears similar to thunder, and hence, lumping it into that interference category makes sense for that data set. Conversely, wind is extremely loud and common in the GRSA001 data, mostly due to the fact that the flat and open area where the monitoring station was placed is conducive to rapid airflow.

Table 2.2. Characteristics of different source types in the GRSA001 data set.

	Source	Typical Event Description	Typical Duration	Example
Signal	Propeller Plane	Signatures evolve from mid-frequency narrowband to wideband, and revert back to narrowband.	30–240 s	
	Jet	Starts with low-to-mid frequency signatures, with the mid-frequency signatures slowly fading approximately half-way through the event.	80–260 s	
Interference	Birdsong	Signatures are restricted to high frequency bands and have erratic temporal patterns.	1 s – several hours	
	Rain/ Thunder	Rain has mid-to-high frequency broadband signatures, while thunder is often superimposed with rain and adds impulsive, low-to-mid frequency signatures.	few seconds – several hours	
	Strong Wind	Impulsive low-to-mid frequency signatures appearing similar to thunder, but typically more relentless.	few seconds – several days	
	Elk Calls	Similar to birdsong, but has broadband signatures in the mid-to-high frequency region and is typically higher magnitude.	few seconds – several hours	

Strong wind causes high energy low-to-mid frequency signatures that have high variability and can extend for long periods of time, making it the most disruptive type of interference for proper detection and classification of signals (jets in particular).

As with the KEFJ004 data, manual annotation existed before the development of the methods proposed in this thesis, and serves as the truth that is used to generate results in Chapter 6. The same annotation process was used for both data sets presented in this section. The GRSA001 data set presents many of the same challenges as the KEFJ004 data set, while offering some new ones. In order of importance, the main challenges associated with detecting and classifying signals in the GRSA001 data set are as follows.

- Strong wind is dominant throughout a large percentage of the recordings, and its signatures commonly overlap with those of signals. At times, wind is violent enough to saturate the microphone and cause a “skipping” effect that can be heard when listening to the raw audio.

- There are many different types of interference to contend with, making it difficult to incorporate robustness to each of them simultaneously, especially since each type is unique in the frequency bands it occupies and temporal patterns it exhibits.
- As a consequence of prominent interference, it is rare to witness a signal event in isolation that may be used to properly train a system to detect such a signal. This is especially a problem for planes, which were not frequently present.
- Jets pass over the monitoring site so frequently that it is common to have several hours of data where jet signatures are present a higher percentage of time than absent, i.e., this is a very high-traffic monitoring site. In some cases the signatures of multiple jets may even overlap in time and frequency.

As can be seen, the GRSA001 data truly represents a complicated soundscape that would test the limits of any source characterization framework. Together with the KEFJ004 data, very diverse acoustical conditions are represented that provide excellent opportunities to determine whether or not a given approach can provide acceptable detection and classification performance for real soundscape monitoring applications. Owing to the challenges presented by this data, it is clear that developed methods should exploit the structure of the data as much as possible.

2.4. Conclusions

This chapter discusses the details pertaining to the national park soundscape data that motivated the development of the source characterization methods introduced in this thesis. An overview of the acoustical monitoring process was first given, including details on the data collection procedure using monitoring stations, general properties of extrinsic and intrinsic sources that are typically encountered in soundscapes, and the interactions of sources with each other and the natural environment. It was shown that this monitoring process yields data that presents unique source characterization challenges including highly nonstationary signals that leads to extreme within-class diversity and between-class similarities, complicated and strong structured interference that is often simultaneously present with signals, and a data format that is sufficient, but not optimal for performing the required tasks.

Two data sets were then introduced, that are used to conduct performance evaluations in this thesis. The challenges presented by these data sets are realizations of the more general complications discussed before, and offer an abundance of scenarios to appropriately stress test source characterization algorithms. For instance, the KEFJ004 data set contains many weak signal events owing to the varying trajectories of associated sources, making them difficult to detect and classify. Moreover, ambient noise levels are fairly high in this data set, and rain and thunder that obstruct the signals to be detected are present a high percentage of the time. The GRSA001 data set, on the other hand, presents a somewhat different (though not disjoint) set of challenges, including extremely strong and persistent wind noise, an abundance of different interference types, and very frequent occurrence of jet events.

The discussion in this chapter underlines the need for robust methods to simultaneously detect and classify highly erratic and nonstationary signals in the presence of prominent competing interference, using sequential multivariate data, which is a problem that has not been addressed before in a comprehensive fashion. In the next chapter, some of the more fundamental approaches to these individual tasks will be introduced both to provide helpful preliminary information, and because they serve as building blocks for the complete source characterization systems introduced in Chapters 4 and 5.

CHAPTER 3

An Introduction to Detection and Classification of Transient Sources

3.1. Introduction

Detection and classification of transient sources is a broad area of research that has resulted in a variety of effective solutions, many of which are tailored to specific applications or assumptions about the data. As indicated in Section 1.2, many common approaches to this task [23–27, 29–31, 49] are not applicable to the soundscape characterization problem considered in this thesis, mainly due to incompatible assumptions about the data and/or sources. Here, a transient source is one whose signatures are not continually present in the data, but appear at some unknown time k_1, and cease to be extant at some other unknown time $k_0 > k_1$. Therefore, this process involves estimating the onset time k_1, duration $k_0 - k_1 - 1$, and class of each new acoustical event as new multivariate observations arrive. Additionally, due to the properties of acoustical sources introduced in Chapter 2, their signatures must be modeled as random. Therefore, in this thesis transient detection is generally performed by looking for a change in the model parameters that best fit the data [33, 35] from those associated with a quiescent period (absence of signal) to those associated with a signal of interest. Classification can typically be performed using a direct extension of the detection approach, namely by determining which class-specific model parameters are most likely given the data.

The main goal of this chapter is to provide an overview of existing techniques that are useful for detecting, and sometimes simultaneously classifying, the aforementioned types of transient sources. The utility of this information is two-fold. First, it provides an introduction to the underlying mechanics of typical transient source characterization methods, so that fundamental concepts may be discussed before considering more complicated scenarios in subsequent chapters, e.g., observations containing superimposed signal and interference source signatures. Second, many of the concepts discussed in this chapter are extended or used as building blocks to form the more complete source characterization systems developed in Chapters 4 and 5. Introducing the established concepts in this chapter, therefore, simplifies the descriptions of these new systems. Consequently, none of the ideas discussed in this chapter offer a complete solution to the soundscape characterization problem at the core of this thesis, as each fails to address certain facets of this problem. In an attempt to keep the concepts general, methods that assume the data follows a very specific structure, e.g., piecewise autoregressive (AR) [67], are not covered. Focusing on generalized concepts ensures the covered materials can be extended or repurposed for use in more advanced algorithms that remain flexible, as intended.

This chapter is organized as follows. Section 3.2 discusses the fundamental transient detection approaches that are relevant to the problem considered in this thesis, namely the sequential probability ratio test (SPRT), cumulative sum (CUSUM) procedure, and generalized likelihood ratio test (GLRT) for detection (and classification) from individual observations. Section 3.3 introduces a new method for fusing a sequential stream of classification decisions using Hidden Markov Models (HMM) [36] so that entire detected acoustical events are assigned a unified class label. This decision fusion is often necessary when applying methods that make separate decisions on individual observations. Conclusions are then drawn in Section 3.4.

3.2. Detection Using Sequential-Based Methods

This section introduces some fundamental concepts for detecting a change in the parameters of the data. When appropriate, assigning class labels to detected events is also discussed, which involves determining the parameterization of the data that is most likely. The main detection statistic of interest, which is the log-likelihood ratio (LLR), and the SPRT that uses this statistic, are described first. It is then shown how repeated SPRTs can be used to implement a CUSUM procedure for detecting the onset of a transient event. Finally, use of the LLR for detecting and classifying specific combinations of source signatures from individual observations is discussed. Note that many of the concepts and terminologies introduced in this section are drawn from [33, 35].

3.2.1. Sequential Probability Ratio Test

Denote $\mathbf{y}_k \in \mathbb{R}^N$ as an observation vector at time k, and consider the vector sequence $\mathbf{Y}_1^n = \{\mathbf{y}_k\}_{k=1}^n$, where $n = 1, 2, \ldots$ increases as new data arrives. A SPRT, also called Wald's test [68], continually updates a LLR test statistic for each incoming $\mathbf{y}_n$ in order to test between two hypotheses about the arbitrary model parameter set Θ, given the data $\mathbf{Y}_1^n$. Such a hypothesis test can be written as

$$\begin{aligned} \mathcal{H}_0 &: \Theta = \Theta_0 \\ \mathcal{H}_1 &: \Theta = \Theta_1 \end{aligned} \tag{3.1}$$

where Θ_0 and Θ_1 are the model parameter sets under the null and alternative hypotheses, respectively. Note that this definition of Θ is general, e.g., it may contain parameters of a HMM (as in Section 3.3) or parameterize a distribution for the data, so long as it may be used to generate a probability measure based on $\mathbf{Y}_1^n$. In detection literature the null and alternative hypotheses are typically associated with the absence and presence of a signal in the observation, respectively. For instance, the matched subspace detector [44] in Appendix D defines Θ as a set of subspace coordinate vectors for signal and interference components. The alternative hypothesis uses estimates of these coordinates that are the projection of the observation onto the associated (known) subspace, whereas the null hypothesis assumes the signal coordinate vector is equal to zero. Clearly, a binary hypothesis test such as (3.1) is insufficient for the data in Chapter 2, but is helpful to introduce the fundamental concepts.

Define $\ell\left(\Theta; \mathbf{Y}_1^n\right)$ as the likelihood of $\Theta \in \{\Theta_0, \Theta_1\}$ given the data $\mathbf{Y}_1^n$, and $f_\Theta(\cdot)$ as the parameterized distribution function used to evaluate $\ell\left(\Theta; \mathbf{Y}_1^n\right)$. A SPRT implements (3.1) using the LLR

$$L(n) = \ln\left(\frac{\ell\left(\Theta_1; \mathbf{Y}_1^n\right)}{\ell\left(\Theta_0; \mathbf{Y}_1^n\right)}\right) = \ln\left(\frac{f_{\Theta_1}\left(\mathbf{Y}_1^n\right)}{f_{\Theta_0}\left(\mathbf{Y}_1^n\right)}\right) \tag{3.2}$$

$$= \ln\left(\frac{f_{\Theta_1}\left(\mathbf{y}_1\right)}{f_{\Theta_0}\left(\mathbf{y}_1\right)}\right) + \sum_{k=2}^{n} \ln\left(\frac{f_{\Theta_1}\left(\mathbf{y}_k | \mathbf{y}_{k-1}, \ldots, \mathbf{y}_1\right)}{f_{\Theta_0}\left(\mathbf{y}_k | \mathbf{y}_{k-1}, \ldots, \mathbf{y}_1\right)}\right) \tag{3.3}$$

where (3.2) can be decomposed to yield (3.3) owing to the probability chain rule and possible dependence between $\mathbf{y}_k$'s. The conditional distribution $f_{\Theta}\left(\mathbf{y}_k | \mathbf{y}_{k-1}, \ldots, \mathbf{y}_1\right)$ can be difficult to generate in practice unless certain assumptions about the data are made, since the number of previous observations used continually increases until the SPRT concludes. One common approach to simplifying (3.3) is to assume $\mathbf{y}_k$'s are statistically independent and identically distributed (IID), meaning (3.3) becomes

$$L(n) = \sum_{k=1}^{n} \ln\left(\frac{f_{\Theta_1}\left(\mathbf{y}_k\right)}{f_{\Theta_0}\left(\mathbf{y}_k\right)}\right).$$

This IID assumption will be used throughout the remainder of this section for simplicity, though one technique for maintaining observation dependence is discussed in Section 3.3. The term

$$s_k = \ln\left(\frac{f_{\Theta_1}\left(\mathbf{y}_k\right)}{f_{\Theta_0}\left(\mathbf{y}_k\right)}\right) \tag{3.4}$$

is the sufficient statistic for the test in (3.1), i.e., information about the unknown parameter Θ that is contained in $\mathbf{y}_k$ is concentrated in s_k [33]. In other words, basing the SPRT on s_k ensures all relevant evidence is being used to discriminate between the different hypotheses in (3.1). Note that even the marginal distribution $f_{\Theta}\left(\mathbf{y}_k\right)$ can be difficult to form when using vector observations without making further assumptions due to the complexity of observation compositions associated with certain hypotheses. Therefore, the SPRT is typically presented in terms of scalar observations [33]. Specific approaches for handling vector observations are discussed in Chapters 4 and 5.

A SPRT concludes whenever $L(n)$ crosses one of the thresholds given by the real numbers A and B, with $-\infty < B < 0 < A < \infty$, which are chosen based on acceptable error probabilities as explained below. The stopping time at which a final decision is made is given by

$$n^* = \min \{n \geq 1 : L(n) \geq A \text{ or } L(n) \leq B\}$$

where $L(n^*) \geq A$ and $L(n^*) \leq B$ correspond to accepting $\mathcal{H}_1$ and $\mathcal{H}_0$, respectively. As in binary hypothesis tests with fixed n, the parameters used to control performance of a SPRT are the false alarm rate α and probability of a missed detection β, given by

$$\alpha = \Pr \left(L(n^*) \geq A | \mathcal{H}_0 \right)$$

$$\beta = \Pr \left(L(n^*) \leq B | \mathcal{H}_1 \right).$$

In [68], relationships between the LLR thresholds and the error probabilities were established as the following approximations

$$A \approx \frac{1-\beta}{\alpha}$$

$$B \approx \frac{\beta}{1-\alpha}$$

with an underlying assumption that $L(n^*)$ will be exactly equal to one of the two thresholds when the test concludes.

Since the SPRT is sequential, the performance measures that are of primary interest are the average run length under each hypothesis, given by

$$T^* \approx E_{\mathcal{H}_0} \left[n^* \right]$$

$$D^* \approx E_{\mathcal{H}_1} \left[n^* \right]$$

where $E_{\mathcal{H}_i}$ denotes the expectation over an ensemble of observation sequences belonging to $\mathcal{H}_i,\ i \in \{0, 1\}$. The SPRT is important for sequential detection since it is optimal in terms of average run length under each hypothesis, given fixed values of error probabilities [68], assuming observations are independent. That is to say, the SPRT decides between $\mathcal{H}_0$ and $\mathcal{H}_1$ using the fewest number of samples possible for given values of α and β. In practice, the SPRT also work well when reformulated to consider dependent observations [35], as in (3.3) (assuming the conditional distributions can be properly formed), though currently no rigorous mathematical proof exists showing optimality in this case. Note that a SPRT is closed [35], meaning $\Pr(n^* < \infty) = 1$ due to the antipodality condition

$$E_{\mathcal{H}_0}\left[s_k\right] < 0$$

$$E_{\mathcal{H}_1}\left[s_k\right] > 0.$$

Since the SPRT assumes all of the observations in $\mathbf{Y}_1^n$ belong to one of two hypotheses, it cannot be used by itself to perform transient detection, where the parameters of the data change at some unknown time. However, the concepts introduced by the SPRT can be used as building blocks to construct appropriate transient detection schemes, as discussed below.

3.2.2. Change Detection using CUSUM

The CUSUM procedure [69], also known as Page's test, is an efficient method for detecting a change in the parameters of a model/distribution governing a data sequence. CUSUM implements the following binary hypothesis test

$$\mathcal{H}_0 : \mathbf{y}_k = \mathbf{v}_k, \ 1 \le k \le n$$

$$\mathcal{H}_1 : \mathbf{y}_k = \begin{cases} \mathbf{v}_k, & 1 \le k < k_1 \\ \mathbf{z}_k, & k_1 \le k \le n \end{cases} \tag{3.5}$$

where k_1 is the unknown change time, and $\mathbf{v}_k$ and $\mathbf{z}_k$ are independent vectors from separate IID vector sequences with associated probability measures $f_{\Theta_0}(\cdot)$ and $f_{\Theta_1}(\cdot)$, respectively. In other words, (3.5) is applicable when the model (or distribution) of observations $\mathbf{y}_k$'s before and after some unknown time k_1 is different, and the goal is to estimate k_1 as quickly as possible. As with the SPRT, there is no inherent restriction on the meaning of the parameters Θ_0 and Θ_1. The CUSUM procedure is relevant to simplified scenarios considered in this thesis (see Section 1.3), as it can be 1) used to process sequential data "on-line", 2) used to detect random sources, 3) extended to handle dependent observations and an unknown parameter after the change.

Denote L_k^n as the LLR given $\mathbf{Y}_k^n$, i.e.,

$$L_k^n = \ln\left(\frac{f_{\Theta_1}\left(\mathbf{Y}_k^n\right)}{f_{\Theta_0}\left(\mathbf{Y}_k^n\right)}\right).$$

Due to a change in distribution at the unknown time k_1 in (3.5), it is easy to see that

$$\begin{aligned} L_1^n &= \ln\left(\frac{f_{\Theta_0}\left(\mathbf{Y}_1^{k_1-1}\right) f_{\Theta_1}\left(\mathbf{Y}_{k_1}^n\right)}{f_{\Theta_0}\left(\mathbf{Y}_1^{k_1-1}\right) f_{\Theta_0}\left(\mathbf{Y}_{k_1}^n\right)}\right) \\ &= \sum_{k=k_1}^{n} \ln\left(\frac{f_{\Theta_1}\left(\mathbf{y}_k\right)}{f_{\Theta_0}\left(\mathbf{y}_k\right)}\right) = L_{k_1}^n. \end{aligned} \tag{3.6}$$

This leads to Page's decision rule [69], which can be derived from the GLRT, used to find the stopping time (i.e., estimated change time)

$$k_1^* = \arg\min_n \left\{ \left(\max_{1\leq k\leq n} L_k^n \right) \geq \eta \right\} \tag{3.7}$$

where η is a predetermined detection threshold. This decision rule essentially states that a change in the model parameters should be declared (accept $\mathcal{H}_1$) whenever any segment of L_1^n increases by at least η, and the estimated change time should be the earliest sample where this level of increase is observed. This concept is demonstrated by the solid line in Fig. 3.1, where the presence of a signal in the 1/3 octave data sequence (see Chapter 2) causes the LLR in (3.6) to increase. The

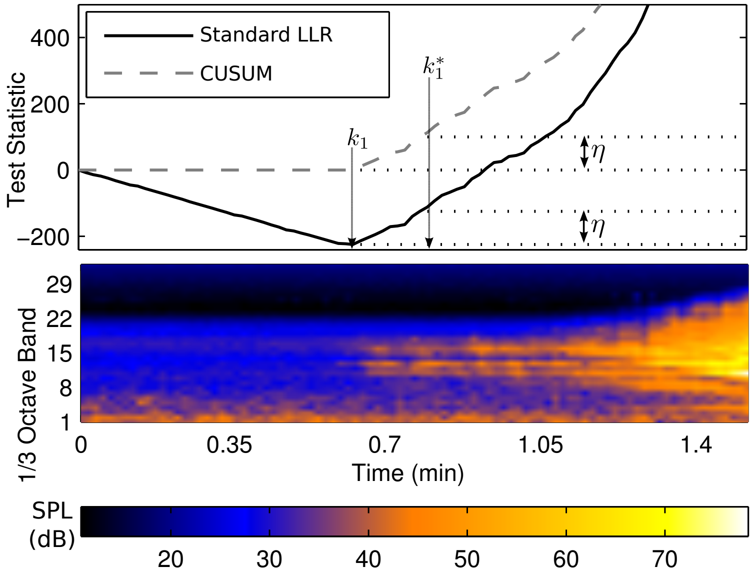


FIGURE 3.1. Examples of the standard LLR and CUSUM test statistics used in (3.6) and (3.9), respectively, along with the corresponding conditions for acceptance of $\mathcal{H}_1$ when $\eta = 100$.

signal is detected at time k_1^* when the LLR increases by at least η as measured from the minimum value it achieves.

Given that observations are assumed to be IID (for now), it can be seen that

$$\max_{1 \leq k \leq n} L_k^n = L_1^n - \min_{1 \leq k \leq n} L_1^{k-1} \tag{3.8}$$

meaning an equivalent detection statistic can be derived by ignoring the first $0 \leq k \leq n-1$ samples prior to the minimum value of L_1^n. Consequently, the standard recursion that implements the CUSUM procedure can be written as [35]

$$S(n) = \max\{0,\ S(n-1) + s_n\},\ n = 1, 2, \ldots \tag{3.9}$$

where $S(0) = 0$ and the update nonlinearity s_n is given in (3.4). The stopping time is then

$$k_1^* = \arg\min_n \{S(n) \geq \eta\}\,. \tag{3.10}$$

Equation (3.9) specifies that the statistic $S(n)$ should be reset whenever it falls below zero, and implements the concept shown in (3.8) where the segment of the LLR leading to its minimum value is ignored, i.e., the associated observations are said to belong to $\mathcal{H}_0$.

The dashed line in Fig. 3.1 shows an example of the test statistic $S(n)$ and the equivalence of using (3.7) and (3.10) for estimating k_1. The CUSUM statistic is clamped at zero while L_1^n decreases below zero, and starts to increase after L_1^n reaches its minimum value. A change is detected whenever $S(n)$ exceeds η, making CUSUM much simpler to interpret than the decision rule in (3.7), despite their equivalence [35]. Operationally, the CUSUM procedure is equivalent to a series of SPRTs with upper and lower thresholds $A = \eta$ and $B = 0$, respectively. That is to say, CUSUM functions as if repeated SPRTs are run that all end in accepting $\mathcal{H}_0$, with the exception of the final test that ends in accepting $\mathcal{H}_1$.

The standard implementation of CUSUM seen here also differs from the SPRT and traditional fixed sample detectors since, assuming the test is closed ($\Pr(k_1 < \infty) = 1$), there is no probability of detection, i.e., $\mathcal{H}_1$ is always accepted eventually [35]. Therefore, the performance of CUSUM is measured in terms of the mean number of samples between false alarms T, and the delay to detection D. The goal is then to achieve large T and small D, which is managed by the choice of threshold η. The merit of CUSUM is that, when using the LLR as the update nonlinearity (as in (3.9)), it has min-max optimality in terms of average run length, i.e., it minimizes the worst case D for a given level of T [34]. This property follows from the minimum average run length property of the SPRT, and the equivalence of CUSUM to a series of SPRTs. This means that the delay to detection will be the same for these procedures, given the same data and threshold. Deriving an explicit expression for these average run length measures in terms of η is difficult even in the IID case, and it is thought that no feasible solution exists in the case of dependent observations [33]. Instead, approximations to these measures are often made, as in [35]. Nonetheless, a small D is important for the soundscape characterization problem considered in this thesis, as it determines the extent to which event durations are underestimated. Furthermore, if events are short enough, they may be missed entirely if D is too large. These risks must be balanced with the occurrence of false signal detections, which are controlled by T.

Direct application of the original CUSUM method to the data in Chapter 2 is infeasible unless it is extended. This is because the CUSUM procedure can only detect the onset of events of interest, i.e., it does not inherently perform classification or estimate the duration of events. As discussed in Section 3.3 and Chapter 5, classification using the CUSUM procedure can be accomplished by allowing the parameters of the data model after the change to vary, and using their maximum likelihood (ML) estimates to assign labels.

3.2.3. Detection of Transient Events Using Decisions on Individual Observations

While the CUSUM procedure is effective for detecting a change in the parameters of the data, Chapter 2 showed that the data considered in this thesis contains observations whose parameters change frequently owing to different combinations of sources (signals and interference of different types) being present at different times. If interference sources are not suppressed in particular (see Chapter 5), then changes are often so rapid that the CUSUM statistic rarely has time to accumulate. Moreover, a LLR must be found given the same set of observations for each hypothesis [33], but the set of observations that can reasonably occur under each hypothesis is often not identical. One approach to addressing these problems is to assign class labels to each $\mathbf{y}_n$ separately, as opposed to accumulating evidence to make a decision based on the likelihood of observing the entire sequence $\mathbf{Y}_1^n$. Detection using individual observations can be implemented by using a modified version of the log-likelihood ratio test (LLRT) statistic in (3.4). Use of the LLRT can be justified by the Neyman-Pearson lemma, which demonstrates that it has the highest power among all competitors [70]. The unique aspect to this work is determining which sources are of interest for detection (signal and/or interference), and what combinations of them can realistically occur simultaneously (see Chapter 2).

Denote $\mathcal{T}$ as the set containing the general parameters Θ_1's corresponding to the alternative hypothesis, i.e., those parameters for observation models containing specific types of sources that we wish to detect. For example, a given $\Theta_1 \in \mathcal{T}$ may model observations containing one type p signal source and one type q interference source, as in Chapter 4. Since the parameter of $\mathbf{y}_n$ under the alternative hypothesis is unknown, the problem now involves the composite hypothesis test

$$\begin{aligned} \mathcal{H}_0 &: \Theta = \Theta_0 \notin \mathcal{T} \\ \mathcal{H}_1 &: \Theta \in \mathcal{T}. \end{aligned} \tag{3.11}$$

Here, Θ_0 is the known parameter under $\mathcal{H}_0$ and models $\mathbf{y}_n$'s that do not contain source signatures of interest, e.g., it may be noise alone or interference plus noise if we are only interested in detecting signals. Detection can then be performed via the well-known GLRT, which uses the maximum likelihood estimates of unknown parameters [33]

$$\max_{\Theta_1 \in \mathcal{T}} \ln \left(\frac{\ell(\Theta_1; \mathbf{y}_n)}{\ell(\Theta_0; \mathbf{y}_n)} \right) \quad \begin{matrix} \mathbf{y}_n \in \mathcal{H}_1 \\ \geq \\ < \\ \mathbf{y}_n \in \mathcal{H}_0 \end{matrix} \quad \eta \tag{3.12}$$

where η is a predetermined detection threshold. A specific example of a detector based on the GLRT is the matched subspace detector [44] in Appendix D, which assumes the signal lies in some known subspace, and the unknown (but deterministic) parameters are the coordinates of the observation relative to this subspace.

A typical and simple classification strategy to use under the detection framework in (3.12) is to estimate which sources are present based on the ML parameter, i.e. find

$$\hat{\Theta}_1 = \arg\max_{\Theta_1 \in \mathcal{T}} \ln \left(\frac{\ell(\Theta_1; \mathbf{y}_n)}{\ell(\Theta_0; \mathbf{y}_n)} \right). \tag{3.13}$$

The class labels assigned to $\mathbf{y}_n$ would be those corresponding to source types that are assumed to always be present in observations parameterized by $\hat{\Theta}_1$. In essence, (3.13) implements a multiple hypothesis test. A caveat to this ML classification strategy is that it may only be used when the complexity of the observation model (number of parameters) does not vary between the hypotheses that are considered, as this leads to a bias in favor of more complex hypotheses [39]. For instance, matched subspace classifiers [40], which use this ML principle by applying a separate matched subspace detector for each class, guarantees the same number of parameters for each signal class by assuming each signal subspace has the same dimensionality. This can be inferred from Appendix D, since an increase in the dimensionality of the signal subspace can only increase the amount of observation energy that lies in this subspace, which leads to a larger detection test statistic. However, when the number of sources that may be present changes frequently, assuming that each Θ_1 has the same number of parameters is often unrealistic. Circumventing these issues associated with varying complexities of hypotheses is outside the scope of the preliminary material discussed in this chapter, and is instead covered in Chapter 4, where a hierarchical testing scheme is introduced.

Note that the framework in this subsection implies that detection and classification are performed on individual observations separately, but it does not preclude the use of past information/observations for making such decisions. This can be accomplished, e.g., by modifying the parameter Θ_1 based on such prior information, as in Chapter 4. The approach in this subsection is most appropriate in situations where the goal is to discover the exact composition of $\mathbf{y}_n$ in terms of the types of signal and interference sources that may be present, and further there is a known parameterization of each possible combination. Unfortunately, encountering observation compositions whose parameterization is not defined by a member of $\mathcal{T}$ may lead to errors, and the cardinality of this set can increase dramatically with the number of possible signal and interference types. Additionally, since many signals of interest have signatures that always span a cluster of adjacent observations, a sequential decision fusion scheme is often needed to improve detection and classification accuracy; a topic that is discussed next.

3.3. HMM-Based Sequential Decision Fusion

Since the soundscape characterization problem addressed in this thesis primary involves detection and classification of signal sources that are extant for approximately 20–240 observations (with one observation per second), it is critical to assign consistent class labels to entire acoustical events produced by such sources. However, most soundscapes have consistently evolving compositions due to intermittently presents sources, meaning certain approaches must make decisions on and assign labels to individual observations separately, e.g., methods related to the general framework discussed in Section 3.2.3. In this section, such approaches are said to produce a sequence of preliminary signal class labels $\{\tilde{c}_k\}_{k=1}^{n}$, with $\tilde{c}_k \in [0, P]$, for the corresponding set of observations. In particular, $\tilde{c}_k = p$ and $\tilde{c}_k = 0$ are labels associated with a type p signal and no signal, respectively, where P is the total number of signal types. In order to reduce inaccuracies in a given decision sequence $\{\tilde{c}_k\}_{k=1}^{n}$, as well as ensure assigned labels are temporally dependent regardless of the classifier used, this section introduces a new method for sequential decision fusion based on HMMs [35, 36]. More specifically, given some $\{\tilde{c}_k\}_{k=1}^{n}$, the idea is to generate a final sequence of class labels $\{c_k\}_{k=1}^{n}$ by aggregating the information of $\tilde{c}_k$'s over time. This results in one c_k per $\mathbf{y}_k$, that denotes the signal type estimated to be present at time k. The approach discussed below will be the sole method of decision fusion used to produce the results in Chapter 6 for any method that assigns class labels separately to individual observations.

Note that the particularly unique aspect to this work is that decisions for multiple transient signals must be sequential fused, meaning we must "detect" the onset time of a cluster of final decisions corresponding to a particular class. Therefore, many existing fusion strategies [71, 72] that use fixed-length data, and/or assign a unified label to all observations, cannot be used for the present problem. The decision fusion considered here involves a sequential multi-hypothesis test, for which an optimal solution exists [73] (albeit impractical) in the case of independent observations, and when all of the data is assumed to belong to a single hypothesis, as in the SPRT in Section 3.2.1.

In contrast, since the types of signals in the observation sequence change over time, the adopted approach uses an extension of the CUSUM-based procedure in Section 3.2.2, since it is optimal in terms of average run length. The main difference here is that observations (i.e., decisions) are no longer considered independent, and hence, HMMs are used to generate likelihoods [35], as they are well-suited for calculating conditional probabilities using sequential data. A brief review of the HMM is first presented in the context of decision fusion, since familiarity with these fundamentals is important for understanding the proposed framework. An excellent and thorough tutorial on HMMs can be found in [36].

3.3.1. HMM Review

A HMM is a type of stochastic model for a data sequence whose distribution at a given time is dependent on the value of an associated hidden state, which itself is part of a Markov chain. HMMs are frequently used to model speech or other types of acoustical signatures as well as other data sequences with high variability [1, 8, 12, 56] due to their flexibility and exploitation of observation dependencies. Unfortunately, HMMs are difficult to directly apply to the soundscape data considered in this thesis due to the intermittent presence of multiple types of interference, as frequent switching between an abundance of HMMs (possibly with significant overlap) would be required to model the variations in the data. On the other hand, they are well-suited for processing a discrete sequence such as the preliminary decisions $\{\tilde{c}_k\}_{k=1}^{n}$ to yield a more accurate estimate of the type of signal that was present in the data when these decisions were made.

A HMM for the pth signal type models the preliminary decisions $\{\tilde{c}_k\}_{k=1}^{n}$ generated by a given classification method when the signatures of this signal are actually present. The states of the pth HMM are denoted by $\{z_k^{(p)}\}_{k=1}^{n}$, with $z_k^{(p)} \in [1, L]$, and are latent variables (not observable) that permit modeling sequences whose distribution changes over time. For decision fusion, these HMM states do not have any specific physical interpretation, but correspond to different preliminary decision distributions resulting from changes in the behavior of a given classification method when applied to highly variable observations. This means that state transitions typically occur when the structure of an acoustical event being evaluated changes (see examples in Section 2.3), as the preliminary decisions may also change. Ultimately, the idea is to perform decision fusion by finding the likelihood of each HMM given $\{\tilde{c}_k\}_{k=1}^{n}$.

A HMM for discrete decision sequences associated with the pth signal type is specified by the parameter set

$$\Theta_p = \{\mathbf{A}^{(p)}, \mathbf{B}^{(p)}, \boldsymbol{\pi}^{(p)}\} \tag{3.14}$$

where

$$\mathbf{A}^{(p)} = \left[a_{ij}^{(p)}\right] = \left[\Pr(z_{k+1}^{(p)} = j | z_k^{(p)} = i)\right],\ i, j = 1, \ldots, L$$

is the state transition matrix whose (ij)th element denotes the probability that the data is in the jth state at time $k+1$ given it was in the ith state at time k. The emission probability matrix is given by

$$\mathbf{B}^{(p)} = \left[b_i^{(p)}(\tilde{c}_k)\right], \quad i = 1, \ldots, L \tag{3.15}$$

where $b_i^{(p)}(\tilde{c}_k) = \Pr(\tilde{c}_k = p | z_k^{(p)} = i)$, i.e., the probability that the pth class label was preliminarily assigned at time k, given the data was in the ith state. As can be seen, for decision fusion applications, $\tilde{c}_k$'s are used in place of what are normally referred to as observations [36, 72]. Finally, the prior probability the data was initially (at time $k = 1$) in each state is encoded by the vector

$$\boldsymbol{\pi}^{(p)} = \left[\pi_i^{(p)}\right] = \left[\Pr(z_1^{(p)} = i)\right],\ i = 1, \ldots, L.$$

Given a preliminary decision sequence, the elements of a HMM can be used to find the value of the forward variable [36] at time k and for state i, i.e.

$$\alpha_k^{(p)}(i) = f_{\Theta_p}(\tilde{c}_k, \ldots, \tilde{c}_1, z_k = i)$$

where $f_{\Theta_p}(\cdot)$ is a probability density function for the HMM Θ_p. The name forward variable comes from the fact that they can be computed recursively over time using

$$\alpha_{k+1}^{(p)}(j) = \left[\sum_{i=1}^{L} \alpha_k^{(p)}(i) a_{ij}^{(p)}\right] b_j^{(p)}(\tilde{c}_{k+1})$$

and initialized as $\alpha_1^{(p)}(j) = \pi_j^{(p)} b_j^{(p)}(\tilde{c}_1)$, which uses all elements of the HMM defined in (3.14). An important property of the forward variable is the ability to find the likelihood of Θ_p given a decision sequence as

$$\begin{aligned} \ell(\Theta_p; \tilde{c}_k, \ldots, \tilde{c}_1) &= f_{\Theta_p}(\tilde{c}_k, \ldots, \tilde{c}_1) \\ &= \sum_{i=1}^{L} \alpha_k^{(p)}(i) \end{aligned} \tag{3.16}$$

meaning we can easily compute conditional probabilities as

$$f_{\Theta_p}(\tilde{c}_k | \tilde{c}_{k-1}, \ldots, \tilde{c}_1) = \frac{\displaystyle\sum_{i=1}^{L} \alpha_k^{(p)}(i)}{\displaystyle\sum_{i=1}^{L} \alpha_{k-1}^{(p)}(i)}. \tag{3.17}$$

However, since the likelihood in (3.16) decreases monotonically as additional $\tilde{c}_n$ arrive, the following scaled forward is typically used in its place

$$\xi_{k+1}^{(p)}(j) = \frac{\left[\displaystyle\sum_{i=1}^{L} \xi_k^{(p)}(i) a_{ij}^{(p)}\right] b_j^{(p)}(\tilde{c}_{k+1})}{\displaystyle\sum_{i=1}^{L} \xi_k^{(p)}(i)}$$

with $\xi_1^{(p)}(i) = \alpha_1^{(p)}(i)$, $\forall i$. Use of $\xi_k^{(p)}(i)$'s still allows for sequential updating of the conditional likelihood since

$$f_{\Theta_p}(\tilde{c}_k | \tilde{c}_{k-1}, \ldots, \tilde{c}_1) = \sum_{i=1}^{L} \xi_k^{(p)}(i) \tag{3.18}$$

just as in (3.17), but without the numerical underflow issues associated with using $\alpha_k^{(p)}(i)$'s. Now that it has been shown how to find the conditional likelihood of a HMM, the decision fusion framework can be presented, which detects changes from one HMM to another.

3.3.2. HMMs for Decision Fusion

The sequential decision fusion method introduced here is based on the generalized likelihood ratio (GLR) for change detection [33], with an unknown HMM Θ_p after the change, which can be implemented by extending the CUSUM procedure in Section 3.2.2. For sequential decision fusion, detecting a single HMM change and stopping, as with the standard CUSUM procedure [35], is insufficient since multiple signals are assumed to be present in the data. This can be remedied by adopting a two phase approach for processing preliminary decisions: 1) detect the time when a signal becomes extant (i.e., $c_k \neq 0$) while it is assumed that none are present (i.e., $c_k = 0$), and 2) detect the time when the signal is no longer extant while it is assumed that one is present. These are referred to as signal and quiescent detection phases, respectively. Estimating the onset time of $c_k = 0$ decisions effectively estimates the duration of the most recent sequence of final class labels corresponding to $c_k \neq 0$, and the process can revert back to looking for a new sequence of signal class labels. The actual values of the labels are only determined after the event duration is estimated for reasons explained below.

Since segments of final decisions corresponding to $c_k \neq 0$ are being continually detected, it is helpful to adopt notation for the various stopping time estimates relative to the current time n. Let k_0 and k_1 denote the unknown onset times of $c_k = 0$ and $c_k \neq 0$ labels, respectively, and let $\hat{k}_0$ and $\hat{k}_1$ denote the estimated onset times for the most recently detected periods where $c_k = 0$ and $c_k \neq 0$, respectively. Fig. 3.2 demonstrates the two-phase concept by showing the circumstances under which each phase is implemented, as well as the previously estimated stopping times relative to the current time n. This figure also provides example preliminary and final decision sequences that are the input and output to the proposed sequential decision fusion process, respectively, and

Table 3.1. Correspondence between colors used in decision strips and each signal type.

Signal Type	None	Plane	Helicopter	Jet
Color Code		🟥	🟩	🟦

were generated by the method in Chapter 4 for the data sequence shown at the bottom. The color of the "decision strip" at a given point in time indicates the assigned preliminary/final class label, according to the key in Table 3.1. As can be seen, several errors in the preliminary decision sequence are corrected to yield a unified class label for each signal event, including some misclassifications of observations at the beginning and end of the first and second signal events, respectively, and some missed detections at the beginning and end of the second signal event.

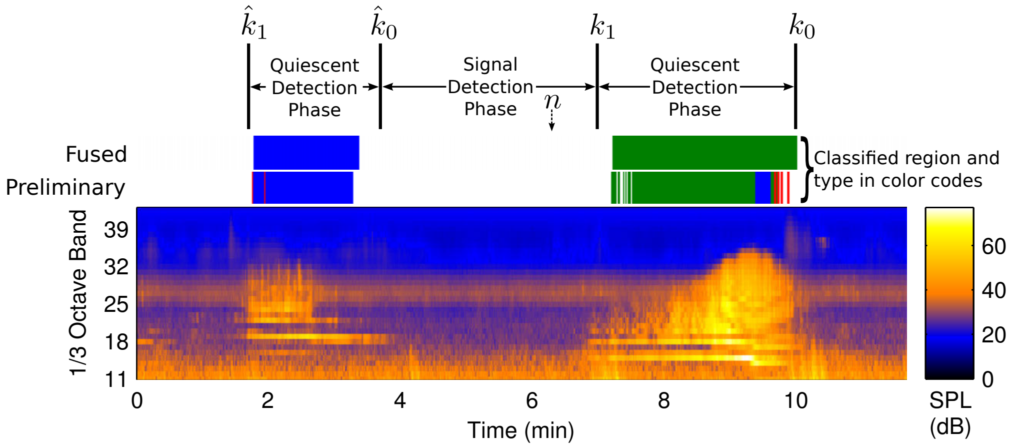


Figure 3.2. Illustration of the two phase decision fusion approach, where the durations of several phases are shown above part of a 1/3 octave observation sequence and corresponding preliminary and final decision sequences. The unknown and estimated onset times for each hypothesis are shown relative to the current time n.

The first phase of the proposed decision fusion process is characterized by the following hypothesis test concerning the final decision sequence

$$\mathcal{H}_0 : c_k = 0, \ \hat{k}_0 \leq k \leq n \tag{3.19}$$

$$\mathcal{H}_1^{(p)} : c_k = \begin{cases} 0, & \hat{k}_0 \leq k < k_1 \\ p, & k_1 \leq k \leq n \end{cases} .$$

As can be seen, under $\mathcal{H}_1^{(p)}$ the onset of final decisions indicating the presence of a signal $c_k = p,\ k \in [k_1, n]$ occurs at the unknown time k_1, and the goal is to find the new estimate $\hat{k}_1$. Therefore, an underlying assumption is that only one signal source may be present at a time. Additionally, to accommodate the two phase approach mentioned above, it is assumed that temporally adjacent observations do not have class labels corresponding to two different signals, but rather, final class labels can only switch from $c_k = p$ to $c_k = 0$ or vice versa.

As before, (3.19) may be implemented using the CUSUM procedure and a set of test statistics based on the LLR, that correspond to the different signal types. The main difference here is that the temporal pattern of preliminary decisions is considered for fusion, meaning $\tilde{c}_k$'s are not considered independent. The LLR is therefore [35]

$$\begin{aligned}
L^n_{\hat{k}_0}(\Theta_p, k_1) &= \ln\left(\frac{\ell(\Theta_p; \tilde{c}_n, \ldots, \tilde{c}_{\hat{k}_0})}{\ell(\Theta_0; \tilde{c}_n, \ldots, \tilde{c}_{\hat{k}_0})}\right) \\
&= \ln\left(\frac{f_{\Theta_0}(\tilde{c}_{k_1-1}, \ldots, \tilde{c}_{\hat{k}_0}) f_{\Theta_p}(\tilde{c}_n, \ldots, \tilde{c}_{k_1})}{f_{\Theta_0}(\tilde{c}_{k_1-1}, \ldots, \tilde{c}_{\hat{k}_0}) f_{\Theta_0}(\tilde{c}_n, \ldots, \tilde{c}_{k_1} | \tilde{c}_{k_1-1}, \ldots, \tilde{c}_{\hat{k}_0})}\right) \\
&= \ln\left(\frac{f_{\Theta_p}(\tilde{c}_{k_1})}{f_{\Theta_0}(\tilde{c}_{k_1}, | \tilde{c}_{k_1-1}, \ldots, \tilde{c}_{\hat{k}_0})}\right) + \sum_{k=k_1+1}^{n} \ln\left(\frac{f_{\Theta_p}(\tilde{c}_k | \tilde{c}_{k-1}, \ldots, \tilde{c}_{k_1})}{f_{\Theta_0}(\tilde{c}_k, | \tilde{c}_{k-1}, \ldots, \tilde{c}_{\hat{k}_0})}\right) \qquad (3.20)
\end{aligned}$$

where Θ_0 is the HMM under $\mathcal{H}_0$. The second equality in (3.20) comes from assuming that, under $\mathcal{H}_1^{(p)}$, the preliminary decisions before and after the change (onset of the signal) are independent of each other. The LLR in (3.20) is a function of the HMM Θ_p as well as the unknown change time k_1 since observations are dependent, meaning the subscript $\hat{k}_0$ is used to keep track of earliest observation under consideration.

As in Section 3.2.2, a recursive test statistic for the pth signal can be used [35] in place of (3.20), that is given by

$$B_p(n) = \max\{0,\ B_p(n-1) + b_p(n, k_r)\},\ n = \hat{k}_0, \hat{k}_0 + 1, \ldots$$

with $B_p(\hat{k}_0-1)=0,\ \forall p$ and where k_r is the time step immediately after the last reset of $B_p(\hat{k}_0-1)$ to zero. For decision fusion, the update nonlinearity uses conditional probabilities, calculated using the HMM Θ_p, i.e. [35]

$$b_p(n,k_r)=\ln\left(\frac{f_{\Theta_p}(\tilde{c}_n|\tilde{c}_{n-1},\ldots,\tilde{c}_{k_r})}{f_{\Theta_0}(\tilde{c}_n|\tilde{c}_{n-1},\ldots,\tilde{c}_{k_r})}\right)=\ln\left(\frac{\sum\limits_{i=1}^{L}\xi_n^{(p)}(i)}{\sum\limits_{i=1}^{L}\xi_n^{(0)}(i)}\right) \tag{3.21}$$

where the second equality is due to (3.18). Whenever $B_p(n)<0$ we set $k_r=n$ and reinitialize the scaled forward variables as $\xi_{k_r}^{(p)}(j)=\pi_j^{(p)}b_j^{(p)}(\tilde{c}_{k_r})$. There is a notable difference between the conditional likelihoods in (3.20) and $b_p(n,k_r)$ in (3.21), namely that in the latter both the numerator and the denominator are conditioned on the same set of preliminary decisions $\tilde{c}_{n-1},\ldots,\tilde{c}_{k_r}$. This is made possible by assuming that the HMM Θ_0 is stationary [35], meaning the distribution of the initial state in the hidden sequence follows the stationary distribution of the HMM states. Detection of the time when class labels no longer follow $\mathcal{H}_0$ is then performed using

$$\max_p B_p(n) \underset{c_n=0}{\overset{c_n\neq 0}{\gtrless}} \eta$$

where η is a predetermined threshold. As can be seen, since the update nonlinearity in (3.21) is based on a conditional likelihood, $B_p(n)$ represents the likelihood of Θ_p given the entire preliminary decision sequence since the last reset.

Continuing with the two-phase approach, once a sequence of signal class labels is detected at time $\hat{k}_1$, the process reverts to detecting a sequence of $c_k=0$ final decisions according to

$$\mathcal{H}_1^{(p)}: c_k=p,\ \hat{k}_1\le k\le n \tag{3.22}$$

$$\mathcal{H}_0: c_k=\begin{cases}p, & \hat{k}_1\le k<k_0\\ 0, & k_0\le k\le n\end{cases}.$$

Using the same principles as in (3.20), the LLR for implementing this test is a function of the unknown change time k_0 given by

$$\begin{aligned} F_{\hat{k}_1}^{n}(\Theta_{p^*}, k_0) &= \ln\left(\frac{\ell(\Theta_0; \tilde{c}_n, \ldots, \tilde{c}_{\hat{k}_1})}{\ell(\Theta_{p^*}; \tilde{c}_n, \ldots, \tilde{c}_{\hat{k}_1})}\right) \\ &= \ln\left(\frac{f_{\Theta_0}(\tilde{c}_{k_0})}{f_{\Theta_{p^*}}(\tilde{c}_{k_0}|\tilde{c}_{k_0-1}, \ldots, \tilde{c}_{\hat{k}_1})}\right) + \sum_{k=k_0+1}^{n} \ln\left(\frac{f_{\Theta_0}(\tilde{c}_k|\tilde{c}_{k-1}, \ldots, \tilde{c}_{k_0})}{f_{\Theta_{p^*}}(\tilde{c}_k|\tilde{c}_{k-1}, \ldots, \tilde{c}_{\hat{k}_1})}\right) \end{aligned} \tag{3.23}$$

where

$$p^* = \arg\max_{p} \, B_p(n)$$

is the ML signal type at time n based on the preliminary decision sequence $\{\tilde{c}_k\}_{k=1}^{n}$. The LLR in (3.23) compares the quiescent HMM Θ_0 with the ML signal HMM Θ_{p^*} so that an increase in this test statistic corresponds to the renewed dominance of $\mathcal{H}_0$. Therefore, the likelihood of each Θ_p is tracked even during quiescent detection phases so that the ML signal type is known at each discrete time instant. The equivalent recursive statistic [35] used to implement this test is

$$T_p(n) = \max\{0, \; T_p(n-1) + t_p(n, k_r)\}, \; n = \hat{k}_1, \hat{k}_1 + 1, \ldots$$

with $T_p(\hat{k}_1 - 1) = 0, \; \forall p$, where k_r denotes the time $T_p(n)$ was last reset and

$$t_p(n, k_r) = \ln\left(\frac{\displaystyle\sum_{i=1}^{L} \xi_n^{(0)}(i)}{\displaystyle\sum_{i=1}^{L} \xi_n^{(p)}(i)}\right)$$

is the update nonlinearity. Final decisions corresponding to $\mathcal{H}_0$ are again accepted when

$$T_{p^*}(n) \underset{c_n \neq 0}{\overset{c_n = 0}{\gtrless}} \gamma. \tag{3.24}$$

As mentioned before, to exploit all available evidence for fusing decisions, final labels are only assigned to a set of observations after $\mathcal{H}_0$ is again accepted according to the test in (3.24). The assigned final class label corresponds to the ML HMM Θ_{p^*} at the time step immediately preceding that where $\mathcal{H}_0$ was accepted. More formally, $\{c_k\}_{k=\hat{k}_1}^{\hat{k}_0-1} = p^*$, where $\hat{k}_0$ is the new estimated time denoting the onset of final decisions $c_k = 0$. As can be seen, this fusion approach assigns a unified signal class label to entire acoustical events, as desired.

As a final note, it is important to consider the increase in computational complexity for the overall transient detection and classification process as a result of using the proposed HMM-based sequential decision fusion, when compared to simply accepting the preliminary decision sequence generated by a given method. In particular, this decision fusion requires $O(L^2P)$ additional operations, since determining the likelihood of each HMM using (3.18) requires $O(L^2)$ operations [36]. Clearly, the number of states L should be small for resource intensive applications, though this is generally advisable anyway when the number of signal types P is small since, in this case, there is often fewer variations in preliminary decision sequences modeled by a given HMM.

3.4. Conclusions

This chapter introduced some of the fundamental concepts used to detect and possibly classify multiple transient sources from sequential multivariate data (e.g., 1/3 octave). These baseline approaches can often be used in simpler source characterization problems, e.g., when observations are independent regardless of the hypothesis they are associated with, and/or when the presence of interference is not considered. This introduction provides several key benefits for understanding the remaining material in this thesis. First, it specifies what constitutes a change of interest for the soundscape characterization application, and supplies the terminology used to define such changes in the context of relatively simple problems. Additionally, the materials in this chapter motivate the use of more sophisticated soundscape characterization frameworks discussed in subsequent chapters, and simplifies their descriptions since the fundamental concepts are often used as building blocks, due to their optimality in certain conditions.

Existing approaches for detection of transient sources were first discussed. The SPRT [68] was introduced for problems where the data in a sequence belongs to one of two hypotheses, and the goal is to use incoming observations to determine which of these hypotheses to accept as quickly as possible. While the formulation used by the SPRT is not representative of the natural soundscape characterization problem, due to constant hypothesis changes in the associated data, this test can be used as a building block for the CUSUM method [69] for detecting a change in the parameters of the data. The standard CUSUM procedure for detecting a hypothesis switch given IID observations was introduced, and shown to be equivalent to repeating SPRTs. Classification in the CUSUM framework typically involves using a generalized likelihood ratio, where the parameter under the alternative hypothesis is allowed to vary, and the assigned class label corresponds to the parameter that maximizes the CUSUM test statistic after a change is detected. The CUSUM method is important for the problem considered in this thesis since it is optimal in terms of average run length, is well-suited for detecting and classifying entire events as it uses all evidence to assign class labels (see Chapter 5), and it is very general, i.e., a variety of models can be used to form the test statistics. On the other hand, there is an inherent delay to detection of transient sources when using the CUSUM method, and it does not perform well when the parameters of the data frequently change, since the test statistics do not have ample time to accumulate in these cases. The latter means that it cannot inherently handle superimposed signal and interference sources that are typically encountered in our problem.

A simple detection method involving a composite hypothesis test for individual observations was also discussed for cases where the goal is to detect observations with specific compositions in terms of signal and interference source signatures. Classification in this framework is also based on the parameter that maximizes the LLR used for detection. This approach performs well in cases where the data composition (hypothesis) frequently switches, as it is based on a uniformly most powerful detector. It can also incorporate an arbitrary number of alternative hypotheses, which is a significant benefit for the present soundscape characterization application. Unfortunately, this approach can only be directly applied in cases where each hypothesis has the same number of associated parameters, and the possible combinations of sources that may be simultaneously present must be known *a priori.* Furthermore, decision fusion must often be used in conjunction with this approach, especially in cases where events contain novel signatures, leading to a potential mixture of preliminary class labels assigned to a single acoustical event.

Finally, a HMM-based sequential decision fusion scheme was developed to address the aforementioned problems associated with assigning class labels separately to individual observations. This fusion is often necessary for the soundscape characterization problem considered in this thesis since signal sources of interest generate acoustical events that generally span 30–240 observations, and hence, a unified class label should be assigned to such events. Since preliminary decision sequences must be used to constantly detect periods of time when signals are present, this framework uses a two-phase CUSUM approach to look for periods of $c_k \neq 0$ labels while it is assumed that $c_k = 0$, and vice versa.

In Chapters 4 and 5, comprehensive solutions to the soundscape characterization problem are introduced, that are capable of detecting and classifying multiple transient events of different types. These approaches were specifically designed to address the main issues with applying the methods discussed in Section 3.2 to the data in Chapter 2. More specifically, the new approaches are able to process sequential data streams containing multiple signal events, inherently exploit the properties of multivariate observations as well as the temporal dependencies between them, and can handle the intermittent presence of multiple types of competing interference. These capabilities are essential when analyzing most real data that does not follow the somewhat restrictive assumptions used by the methods in Section 3.2, e.g., independence of observations.

CHAPTER 4

A Sequential Random Coefficient Tracking Framework

4.1. Introduction

As mentioned in Chapter 1, an extensive amount of research exists offering solutions to various subsets of the tasks required for the transient source characterization problem addressed in this thesis. For instance, many general methods (i.e., those not specifically designed for natural soundscape analysis) in their original form either 1) are incapable of recognizing any general source type [5, 8], 2) cannot handle significant within class diversity [44, 46], 3) cannot perform classification (only detection) [32, 35], and/or 4) are not robust to the presence of structured interference [5, 8, 32, 35, 46, 60]; all of which are essential for handling the intricacies of natural soundscapes. See Section 1.3 for a comprehensive list of necessary capabilities for achieving acceptable source characterization performance. Combining incomplete solutions is possible is some cases, but a lack of cohesion between processing steps (e.g., transient detection and classification) may lead to redundant computations and decreased performance, analogous to performing distributed detection without collaboration between decision-making agents [43]. Moreover, many existing approaches [2, 7, 11, 13, 15, 16, 20] that attempt to offer comprehensive solutions to very similar problems are also not appropriate for this problem for various reasons, e.g., they presume access to the raw audio data.

To address the shortcomings of existing methods for characterization of multiple transient sources, this chapter develops a sequential random coefficient tracking (SRCT) framework that applies a hierarchy of log likelihood ratio tests (LLRT) to individual observations, each of which may contain the signatures of a subspace signal and/or a subspace interference source, both of which can be one of multiple types. Since class labels are assigned separately to each observation, the SRCT method can be seen as a specific implementation of the general detection and classification approach in Section 3.2.3. As suggested in Section 2.2.2, source signatures are modeled as random to capture the variability between different events associated with the same source type. A Kalman filter that exploits known source subspace and coefficient dependency models is used to generate the parameters of the conditional densities necessary for calculating the test statistics, where dependence is on previous observations and a specific source model. It is assumed that at most one type of signal and one type of interference are present at a given time, though as required this SRCT method may be continuously applied to streaming data in order to detect and classify new transient sources, possibly of different types. The method developed in this chapter is also capable of performing separation of the signal and interference portions of the measurement to produce estimates of their signatures in isolation.

This chapter is organized as follows. Section 4.2 introduces the observation model used as the basis for developing the SRCT method. Section 4.3 formally introduces the problem and discusses the entire SRCT source characterization framework, including the general form of the LLRT, the procedure for calculating the parameters necessary to construct a LLRT, and the explicit form of each of the LLRTs that are hierarchically applied to individual observations. Finally, Section 4.4 provides concluding remarks. The experimental results of applying the SRCT method to national park soundscape recordings are presented in Chapter 6, together with the results produced by other methods described in this thesis, so that effective comparisons can be made. Note that most of the material presented in this chapter is also reported in [74].

4.2. Observation Model

The kth vector in the data sequence to be evaluated is referred to as an observation vector and is denoted by $\mathbf{y}_k = [y_k[1] \ \cdots \ y_k[N]]^T = G(\mathbf{u}_k) \in \mathbb{R}^N$, where $\mathbf{u}_k \in \mathbb{R}^L$ is the kth time interval (contains samples $[kL+1, \ (k+1)L]$) of the original sampled audio recording, and $G(\cdot)$ represents a mapping function, e.g., the 1/3 octave mapping [62] described in Appendix A. Since data is being constantly recorded, $\mathbf{y}_k$'s are continually arriving, and can be generally represented as

$$\mathbf{y}_k = \alpha_k \mathbf{s}_k + \beta_k \mathbf{h}_k + \mathbf{w}_k$$

where $\mathbf{s}_k \in \mathbb{R}^N$ and $\mathbf{h}_k \in \mathbb{R}^N$ are random vectors defined similar to $\mathbf{y}_k$, but represent the signatures produced by the unknown signal and interference sources to be characterized, respectively; α_k and β_k are binary variables that indicate the presence $(\alpha_k, \beta_k = 1)$ or absence $(\alpha_k, \beta_k = 0)$ of a signal and interference source, respectively; and $\mathbf{w}_k \overset{\text{IID}}{\sim} \mathcal{N}\left(\mathbf{0}, \mathbf{R}_\mathbf{w}\right)$ is an independent and identically distributed (IID) measurement noise vector, where $\mathbf{R}_\mathbf{w} \in \mathbb{R}^{N \times N}$ is a known full-rank noise covariance matrix. Justification for the assumed distribution of $\mathbf{w}_k$ can be found in Appendix B for the data considered in this thesis where $G(\cdot)$ is the 1/3 octave mapping. Here, zero mean noise is assumed without loss of generality, as the noise mean can always be subtracted from each observation prior to the application of the SRCT method. This assumed noise distribution is also valid whenever 1) the noise in $\mathbf{u}_k$ is IID Gaussian with zero mean and variance σ^2 and 2) $G(\mathbf{u}_k) = \mathbf{G}\mathbf{u}_k$, with $\mathbf{G} \in \mathbb{R}^{N \times L}$, i.e., G is a linear mapping, since in this case $\mathbf{w}_k \overset{\text{IID}}{\sim} \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{G}\mathbf{G}^T\right)$.

For reasons discussed in the next section, it is more convenient to operate on transformed observations $\mathbf{z}_k = \mathbf{R}_\mathbf{w}^{-\frac{1}{2}} \mathbf{y}_k$ with white observation noise $\boldsymbol{\omega}_k = \mathbf{R}_\mathbf{w}^{-\frac{1}{2}} \mathbf{w}_k$, i.e. $E\left[\boldsymbol{\omega}_k \boldsymbol{\omega}_{k-j}^T\right] = \mathbf{I}_N \delta(j)$, where $\mathbf{I}_N$ is the $N \times N$ identity matrix. It is assumed that the transformed signal vector $\mathbf{R}_\mathbf{w}^{-\frac{1}{2}} \mathbf{s}_k$ and transformed interference vector $\mathbf{R}_\mathbf{w}^{-\frac{1}{2}} \mathbf{h}_k$ lie in known low-dimensional subspaces, $\langle \mathbf{S}_k \rangle$ and $\langle \mathbf{H}_k \rangle$, respectively, that are spanned by the columns of $\mathbf{S}_k \in \mathbb{R}^{N \times M}$ and $\mathbf{H}_k \in \mathbb{R}^{N \times M}$, respectively, with $M \ll N$. These subspaces are time-dependent since the source types may change. Specifically, $\mathbf{S}_k \in \{\mathbf{S}^{(p)}\}_{p=1}^P$ and $\mathbf{H}_k \in \{\mathbf{H}^{(q)}\}_{q=1}^Q$, meaning there are P possible signal types and Q possible interference types. All source subspaces, $\left\langle \mathbf{S}^{(p)} \right\rangle$'s and $\left\langle \mathbf{H}^{(q)} \right\rangle$'s, are assumed to be linearly independent of each other (not necessarily orthogonal) and exactly M-dimensional, where the latter requirement is necessary for subspace classifiers [40] in order to prevent bias in favor of those source types with larger associated subspace dimensions. The model for the transformed observation can therefore be written as

$$\mathbf{z}_k = \alpha_k \mathbf{S}_k \mathbf{a}_k + \beta_k \mathbf{H}_k \mathbf{b}_k + \boldsymbol{\omega}_k \tag{4.1}$$

where $\mathbf{a}_k \in \mathbb{R}^M$ and $\mathbf{b}_k \in \mathbb{R}^M$ are random signal and interference coefficient vectors. Hereafter in this chapter, $\mathbf{z}_k$ (rather than $\mathbf{y}_k$) is referred to as the observation at time k.

A subspace model is useful for acoustical source classification from sequential multivariate data since different linear combinations of basis vectors may be used to capture various nonstationarities that are often present in the source's signatures. For instance, as discussed in Appendix C, when using the 1/3 octave representation the time-frequency signatures of a Doppler shifted waveform may change from narrowband to broadband during the time the most rapid frequency shifts occur, and different linear combinations of basis vectors can be used to model this time-varying behavior. Thus, to represent variations among different source events, it is assumed that $\mathbf{a}_k$ and $\mathbf{b}_k$ obey the following respective vector linear autoregressive (AR) models

$$\begin{aligned}
\mathbf{a}_k &= \sum_{j=1}^{J} \boldsymbol{\Phi}_{j,k} \mathbf{a}_{k-j} + \boldsymbol{\xi}_k \\
\mathbf{b}_k &= \sum_{j=1}^{J} \boldsymbol{\Psi}_{j,k} \mathbf{b}_{k-j} + \boldsymbol{\nu}_k
\end{aligned} \tag{4.2}$$

where $\boldsymbol{\Phi}_{j,k}$ and $\boldsymbol{\Psi}_{j,k}$ are the jth AR parameter matrices for the signal and interference sources at time k, respectively. The vectors $\boldsymbol{\xi}_k \overset{\text{IID}}{\sim} \mathcal{N}\left(\mathbf{0}, \mathbf{R}_{\boldsymbol{\xi},k}\right)$ and $\boldsymbol{\nu}_k \overset{\text{IID}}{\sim} \mathcal{N}\left(\mathbf{0}, \mathbf{R}_{\boldsymbol{\nu},k}\right)$ are driving processes for the signal and interference, respectively, with known covariance matrices $\mathbf{R}_{\boldsymbol{\xi},k}$ and $\mathbf{R}_{\boldsymbol{\nu},k}$, that are assumed to be independent of each other and with $\boldsymbol{\omega}_k$. The time index k on the AR parameters again indicates that they may change over time when a new source type becomes extant. Although in (4.2) all source coefficient AR models are assumed to be of order J for notational simplicity, the proposed method can still apply for different AR model orders for each source type.

4.3. Source Characterization

The SRCT method introduced in this section performs transient detection, classification, and estimation of multiple source types by sequentially evaluating $\mathbf{z}_k$'s on-line, i.e., $\mathbf{z}_{k+j},\ j > 0$ is not accessible at time k. This is done by applying a hierarchy of LLRTs to each $\mathbf{z}_k$ to test between several hypotheses that account for each possible observation composition in terms of signal and interference according to the model in (4.1). Specifically, these hypotheses are given by

$$\begin{aligned}
\mathcal{H}_0 &: \mathbf{z}_k = \boldsymbol{\omega}_k \text{ or } \alpha_k = \beta_k = 0 \\
\mathcal{H}_1^{(p)} &: \mathbf{z}_k = \mathbf{S}^{(p)}\mathbf{a}_k + \boldsymbol{\omega}_k \text{ or } \alpha_k = 1,\ \beta_k = 0 \\
\mathcal{H}_2^{(q)} &: \mathbf{z}_k = \mathbf{H}^{(q)}\mathbf{b}_k + \boldsymbol{\omega}_k \text{ or } \alpha_k = 0,\ \beta_k = 1 \\
\mathcal{H}_3^{(p,q)} &: \mathbf{z}_k = \mathbf{S}^{(p)}\mathbf{a}_k + \mathbf{H}^{(q)}\mathbf{b}_k + \boldsymbol{\omega}_k \text{ or } \alpha_k = \beta_k = 1
\end{aligned} \tag{4.3}$$

where superscripts $p \in \{1, \ldots, P\}$ and/or $q \in \{1, \ldots, Q\}$ indicate dependencies of hypotheses on source models. Thus, there are a total of $P + Q + PQ$ different possible hypotheses excluding $\mathcal{H}_0$. The models associated with the above hypotheses are specified as

$$\begin{aligned}
\mathcal{H}_0 &: \lambda_0 \\
\mathcal{H}_1^{(p)} &: \lambda_1^{(p)} = \{\mathbf{S}^{(p)}, \boldsymbol{\Phi}_1^{(p)}, \ldots, \boldsymbol{\Phi}_J^{(p)}, \mathbf{R}_{\boldsymbol{\xi}}^{(p)}\} \\
\mathcal{H}_2^{(q)} &: \lambda_2^{(q)} = \{\mathbf{H}^{(q)}, \boldsymbol{\Psi}_1^{(q)}, \ldots, \boldsymbol{\Psi}_J^{(q)}, \mathbf{R}_{\boldsymbol{\nu}}^{(q)}\} \\
\mathcal{H}_3^{(p,q)} &: \lambda_3^{(p,q)} = \lambda_1^{(p)} \cup \lambda_2^{(q)}
\end{aligned} \tag{4.4}$$

which are formed using training data prior to the application of the SRCT method.

The flow of the SRCT method is shown in the block diagram of Fig. 4.1. As can be seen, the inputs to the system are $\mathbf{z}_k$, the models defined in (4.4), and a set of state parameters that allows the system to exploit the dependency structure of each source (see Section 4.3.2). Ultimately, the system produces one signal and/or one interference label for $\mathbf{z}_k$, in addition to estimates of $\mathbf{s}_k$ and/or $\mathbf{h}_k$, if desired. The test statistics that implement the SRCT method are log likelihood ratios (LLR) formed using the conditional probabilities of observing $\mathbf{z}_k$ given relevant previous observations and specific models in (4.4). The associated LLRTs are applied hierarchically to a given $\mathbf{z}_k$ as follows:

(1) **Detection and Classification:** using (4.23), determine whether $\mathbf{z}_k$ consists of noise alone or contains the signatures of one or two sources; if the latter, then reject $\mathcal{H}_0$ and proceed. The source types, p^* and q^*, that are most likely present in $\mathbf{z}_k$ are also estimated by (4.24).

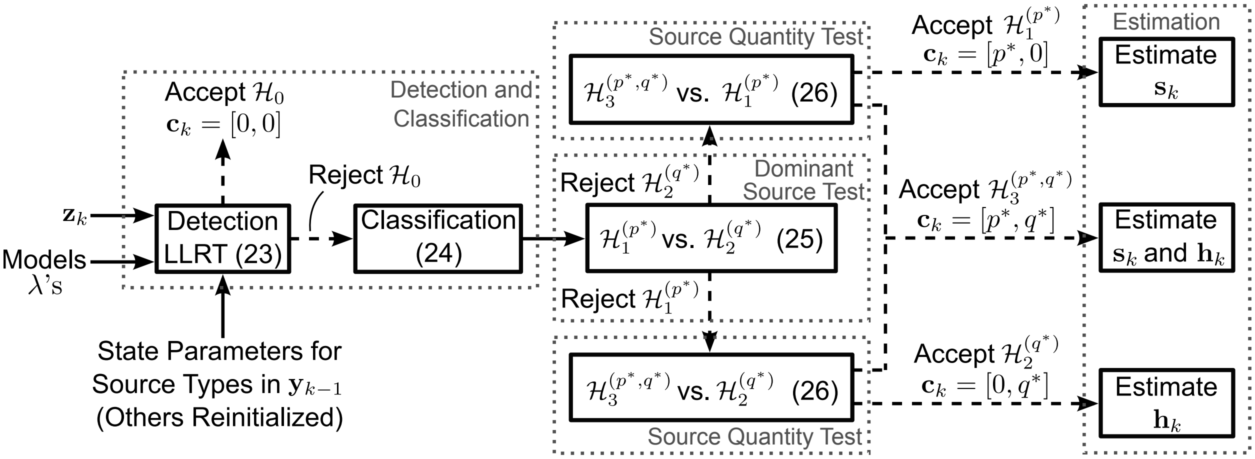
(2) **Dominant Source Test:** use (4.25) to determine whether signal or interference is dominant and, consequently, which remaining single source hypothesis ($\mathcal{H}_1^{(p^*)}$ or $\mathcal{H}_2^{(q^*)}$) is rejected.

(3) **Source Quantity Test:** use (4.26) to test the hypothesis corresponding to the dominant source type against the two source hypothesis $\mathcal{H}_3^{(p^*,q^*)}$ in order to finally accept either $\mathcal{H}_1^{(p^*)}$, $\mathcal{H}_2^{(q^*)}$, or $\mathcal{H}_3^{(p^*,q^*)}$.

(4) **Estimation:** Use estimates of $\mathbf{a}_k$ and $\mathbf{b}_k$, that were generated for the above LLRTs, to form estimates of the actual source signatures. This step is optional and does not involve a LLRT.

A hierarchy of tests is applied here rather than directly finding the most likely hypothesis in (4.4). This is done due to the fact that the complexity of the observation model (number of parameters) varies between hypotheses, which would lead to a bias in LLRTs in favor of more complex hypotheses [39]. A hierarchical test circumvents these issues through the use of different thresholds to compare hypotheses with different relative complexities.

Since each main step in the SRCT process evaluates different hypotheses using the same general form of the LLRT, this form is briefly discussed in the next subsection. Sections 4.3.2 and 4.3.3 discuss calculating distribution parameters for each hypothesis, that are used to form the LLRTs. Explicit forms of the LLRTs are then developed in Section 4.3.4. To conclude this section, the procedure for obtaining estimates of the signatures of detected signal and interference sources is briefly discussed.



Source Quantity Test
Accept $\mathcal{H}_1^{(p^*)}$
$\mathbf{c}_k = [p^*, 0]$
Estimation
Estimate $\mathbf{s}_k$
Accept $\mathcal{H}_0$
$\mathbf{c}_k = [0, 0]$
Detection and Classification
$\mathcal{H}_3^{(p^*,q^*)}$ vs. $\mathcal{H}_1^{(p^*)}$ (26)
Reject $\mathcal{H}_0$
Reject $\mathcal{H}_2^{(q^*)}$
Dominant Source Test
Accept $\mathcal{H}_3^{(p^*,q^*)}$
$\mathbf{c}_k = [p^*, q^*]$
$\mathbf{z}_k$
Detection LLRT (23)
Classification (24)
$\mathcal{H}_1^{(p^*)}$ vs. $\mathcal{H}_2^{(q^*)}$ (25)
Estimate $\mathbf{s}_k$ and $\mathbf{h}_k$
Models λ's
Reject $\mathcal{H}_1^{(p^*)}$
State Parameters for Source Types in $\mathbf{y}_{k-1}$ (Others Reinitialized)
Accept $\mathcal{H}_2^{(q^*)}$
$\mathbf{c}_k = [0, q^*]$
$\mathcal{H}_3^{(p^*,q^*)}$ vs. $\mathcal{H}_2^{(q^*)}$ (26)
Estimate $\mathbf{h}_k$
Source Quantity Test

FIGURE 4.1. Block diagram showing application of the proposed SRCT method to the observation $\mathbf{z}_k$. Dashed lines indicate that a path is followed only when the corresponding decision to accept or reject a given hypothesis is made. The dotted boxes indicate the stage in the hierarchical process.

4.3.1. General form of a LLR

In the SRCT method, a hierarchy of tests is applied to $\mathbf{z}_k$, each of which is based on a likelihood ratio having the general form

$$\begin{aligned} L_k(\boldsymbol{\theta}_k, \boldsymbol{\theta}'_k) &= \frac{\ell\left(\boldsymbol{\theta}_k; \mathbf{z}_k\right)}{\ell\left(\boldsymbol{\theta}'_k; \mathbf{z}_k\right)},\ \boldsymbol{\theta}_k \neq \boldsymbol{\theta}'_k \\ &= \frac{f\left(\mathbf{z}_k | \mathbf{Z}_{k_0}^{k-1}, \lambda\right)}{f\left(\mathbf{z}_k | \mathbf{Z}_{k'_0}^{k-1}, \lambda'\right)},\ \lambda \neq \lambda' \end{aligned} \tag{4.5}$$

where $\ell\left(\boldsymbol{\theta}_k; \mathbf{z}_k\right)$ is the likelihood function of the *distribution parameter* $\boldsymbol{\theta}_k$ (describes conditional PDFs), f is a conditional density function, λ is one of the source models in (4.4), and $\mathbf{Z}_{k_0}^{k-1} = \{\mathbf{z}_{k-1}, \ldots, \mathbf{z}_{k_0}\}$ is the set of observations that $\mathbf{z}_k$ is dependent on under the hypothesis associated with λ. In other words, if a particular source type has been present since time $k_0 \leq k-1$, then $\mathbf{z}_k$ will be dependent on $\mathbf{Z}_{k_0}^{k-1}$ under any hypothesis that assumes the presence of this source, owing to dependence between source coefficient vectors at different times, as shown in (4.2).

The statistic in (4.5) may be used for the test

$$L_k(\boldsymbol{\theta}_k, \boldsymbol{\theta}'_k) \underset{\text{reject } \mathcal{H}}{\overset{\text{reject } \mathcal{H}'}{\gtrless}} \gamma' \tag{4.6}$$

where $\mathcal{H}$ and $\mathcal{H}'$ are hypotheses from (4.3) associated with the models λ and λ', respectively, and γ' is a predetermined threshold. Equations (4.5) and (4.6) show that λ implicitly defines the hypothesis a likelihood function is dependent on, as well as the distribution parameter $\boldsymbol{\theta}_k$. Such tests are used to assign class labels to each $\mathbf{z}_k$ separately, as opposed to accumulating evidence to make a decision based on the likelihood of observing the entire sequence $\mathbf{Z}_{k_0}^{k}$ under a given λ, as in a traditional sequential LLRT [34]. Although the latter approach is possible under the proposed framework, it is avoided since both likelihoods in a LLRT must be found given the same set of observations [33]. On the other hand, since the presence and types of signal and interference change independently as new observations arrive, the set of observations that can reasonably occur under each hypothesis is often not identical. Therefore, using a common $\mathbf{Z}_{k_0}^{k}$, $k_0 < k$ to evaluate all likelihoods can lead to low likelihoods for hypotheses associated with sources that are actually present.

The next subsection shows that $\mathbf{z}_k$ is conditionally multivariate Gaussian with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$, given $\mathbf{Z}_{k_0}^{k-1}$ and λ, owing to the fact that all observations follow the general model introduced in the previous section. Therefore, defining $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ and taking the natural log of (4.5) yields the LLR as

$$\Lambda_k(\boldsymbol{\theta}_k, \boldsymbol{\theta}'_k) = \ln \frac{\ell\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k; \mathbf{z}_k\right)}{\ell\left(\boldsymbol{\mu}'_k, \boldsymbol{\Sigma}'_k; \mathbf{z}_k\right)} = \boldsymbol{\zeta}_k(\boldsymbol{\theta}'_k) - \boldsymbol{\zeta}_k(\boldsymbol{\theta}_k) \tag{4.7}$$

where

$$\boldsymbol{\zeta}_k(\boldsymbol{\theta}_k) = \frac{1}{2} \ln \det\left(\boldsymbol{\Sigma}_k\right) + \frac{1}{2}\left(\mathbf{z}_k - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{z}_k - \boldsymbol{\mu}_k\right) \tag{4.8}$$

TABLE 4.1. Structure of the state variables for each model where $\mathbf{0}_m$ and $\mathbf{0}_{m_1 \times m_2}$ are $m \times m$ and $m_1 \times m_2$ zero matrices, respectively.

Model	$\mathbf{x}_k$	$\mathbf{F}$	$\mathbf{D}$	$\mathbf{C}$	$\mathbf{v}_k$	$\mathbf{R}_\mathbf{v}$
$\lambda_1^{(p)}$	$\mathbf{x}_{1,k} = \begin{bmatrix} \mathbf{a}_k \\ \vdots \\ \mathbf{a}_{k-J+1} \end{bmatrix}$	$\mathbf{F}_1^{(p)} = \begin{bmatrix} \mathbf{\Phi}_1^{(p)} & \cdots & \mathbf{\Phi}_J^{(p)} \\ \mathbf{I}_{(J-1)M} & & \mathbf{0}_{(J-1)M \times M} \end{bmatrix}$	$\begin{bmatrix} \mathbf{I}_M \\ \mathbf{0}_{(J-1)M \times M} \end{bmatrix}$	$\mathbf{S}^{(p)}\mathbf{D}^T$	$\boldsymbol{\xi}_k^{(p)}$	$\mathbf{R}_{\boldsymbol{\xi}}^{(p)}$
$\lambda_2^{(q)}$	$\mathbf{x}_{2,k} = \begin{bmatrix} \mathbf{b}_k \\ \vdots \\ \mathbf{b}_{k-J+1} \end{bmatrix}$	$\mathbf{F}_2^{(q)} = \begin{bmatrix} \mathbf{\Psi}_1^{(q)} & \cdots & \mathbf{\Psi}_J^{(q)} \\ \mathbf{I}_{(J-1)M} & & \mathbf{0}_{(J-1)M \times M} \end{bmatrix}$	$\begin{bmatrix} \mathbf{I}_M \\ \mathbf{0}_{(J-1)M \times M} \end{bmatrix}$	$\mathbf{H}^{(q)}\mathbf{D}^T$	$\boldsymbol{\nu}_k^{(q)}$	$\mathbf{R}_{\boldsymbol{\nu}}^{(q)}$
$\lambda_3^{(p,q)}$	$\begin{bmatrix} \mathbf{x}_{1,k} \\ \mathbf{x}_{2,k} \end{bmatrix}$	$\begin{bmatrix} \mathbf{F}_1^{(p)} & \mathbf{0}_{JM} \\ \mathbf{0}_{JM} & \mathbf{F}_2^{(q)} \end{bmatrix}$	$\begin{bmatrix} \mathbf{I}_M & \mathbf{0}_{JM \times M} \\ & I_M \\ \mathbf{0}_{(2J-1)M \times M} & \mathbf{0}_{(J-1)M \times M} \end{bmatrix}$	$\left[\mathbf{S}^{(p)}, \mathbf{H}^{(q)}\right] \mathbf{D}^T$	$\begin{bmatrix} \boldsymbol{\xi}_k^{(p)} \\ \boldsymbol{\nu}_k^{(q)} \end{bmatrix}$	$\begin{bmatrix} \mathbf{R}_{\boldsymbol{\xi}}^{(p)} & \mathbf{0}_M \\ \mathbf{0}_M & \mathbf{R}_{\boldsymbol{\nu}}^{(q)} \end{bmatrix}$

and similarly for $\boldsymbol{\zeta}_k(\boldsymbol{\theta}'_k)$. Since all tests performed by the SRCT method use a LLR that assumes the general form in (4.7), generating $\boldsymbol{\theta}_k$ under different λ (i.e., different hypotheses) is an integral step of this process, and is discussed next.

4.3.2. Generating Parameter Sets

This subsection discusses calculating $\boldsymbol{\zeta}_k(\boldsymbol{\theta}_k)$ for a given λ, which involves calculating the associated $\boldsymbol{\theta}_k$, i.e., the parameters of the conditional distribution. Beginning with the simple case of $\mathcal{H}_0$ (i.e., $\lambda = \lambda_0$), where $\mathbf{z}_k = \boldsymbol{\omega}_k$, we have

$$\begin{aligned} \mathcal{H}_0 : \boldsymbol{\mu}_k &= E\left[\mathbf{z}_k | \mathbf{Z}_{k_0}^{k-1}, \lambda_0\right] = E\left[\boldsymbol{\omega}_k\right] = \mathbf{0} \\ \mathbf{\Sigma}_k &= E\left[\left(\mathbf{z}_k - \boldsymbol{\mu}_k\right)\left(\mathbf{z}_k - \boldsymbol{\mu}_k\right)^T | \mathbf{Z}_{k_0}^{k-1}, \lambda_0\right] \\ &= E\left[\boldsymbol{\omega}_k \boldsymbol{\omega}_k^T\right] = \mathbf{I}_N \end{aligned}$$

due to the fact that $\boldsymbol{\omega}_k$'s are independent at different times. From (4.8), it follows that $\boldsymbol{\zeta}_k(\boldsymbol{\theta}_k) = \frac{1}{2}\mathbf{z}_k^T \mathbf{z}_k$ under $\mathcal{H}_0$.

Finding the conditional probability of observing $\mathbf{z}_k$ given $\lambda \neq \lambda_0$ is accomplished by using such a model to establish a series of state equations and applying a Kalman filter to obtain the estimates of the time dependent source signatures. More specifically, the idea is to obtain an estimate of the state vector $\mathbf{x}_k$ that represents the relevant past and present source basis coefficient vectors ($\mathbf{a}_{k-j}$'s and/or $\mathbf{b}_{k-j}$'s), under a given hypothesis in (4.3). See Table 4.1 for explicit definitions of the state variables for each model in (4.4). For a specific λ, the general forms of the state equations are given by

$$\mathbf{x}_k = \mathbf{F}\mathbf{x}_{k-1} + \mathbf{D}\mathbf{v}_k \tag{4.9a}$$

$$\mathbf{z}_k = \mathbf{C}\mathbf{x}_k + \boldsymbol{\omega}_k \tag{4.9b}$$

where $\mathbf{F}$, $\mathbf{D}$, and $\mathbf{C}$ are matrices of appropriate forms and dimensions (see Table 4.1), and $\mathbf{v}_k$ is the process noise vector with known covariance matrix $\mathbf{R}_\mathbf{v}$. These equations are obtained by arranging the observation model equations in (4.1) and (4.2) in state space form, and replacing the parameters in the these equations with those from one of the models $\lambda_1^{(p)}$, $\lambda_2^{(q)}$, or $\lambda_3^{(p,q)}$. Using a given set of state equations, a standard Kalman filter can be applied to obtain the *a priori* and *a posteriori* estimates of the source coefficients (states), denoted by $\hat{\mathbf{x}}_{k|k-1}$ and $\hat{\mathbf{x}}_{k|k}$, respectively. The associated error covariance matrices are denoted by

$$\begin{aligned} \mathbf{Q}_{k|k-1} &= E\left[\boldsymbol{\epsilon}_{k|k-1}\boldsymbol{\epsilon}_{k|k-1}^T\right] \\ \mathbf{Q}_{k|k} &= E\left[\boldsymbol{\epsilon}_{k|k}\boldsymbol{\epsilon}_{k|k}^T\right] \end{aligned} \tag{4.10}$$

where $\boldsymbol{\epsilon}_{k|k-1} = \mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1}$ and $\boldsymbol{\epsilon}_{k|k} = \mathbf{x}_k - \hat{\mathbf{x}}_{k|k}$ are the *a priori* and *a posteriori* state error vectors, respectively.

To find $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ for a given λ, the definitions of $\boldsymbol{\epsilon}_{k|k-1}$ and the state variables in Table 4.1 may be used to write (4.9b) as

$$\mathbf{z}_k = \mathbf{C}\left(\hat{\mathbf{x}}_{k|k-1} + \boldsymbol{\epsilon}_{k|k-1}\right) + \boldsymbol{\omega}_k. \tag{4.11}$$

Standard Kalman filter theory [75] dictates that, given the past observations $\mathbf{Z}_{k_0}^{k-1}$, we have $\boldsymbol{\epsilon}_{k|k-1} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{Q}_{k|k-1}\right)$. Since $\boldsymbol{\epsilon}_{k|k-1}$ and $\boldsymbol{\omega}_k$ are both zero mean Gaussian, and together represent the random part of (4.11), $\mathbf{z}_k$ is also conditionally multivariate Gaussian, given $\mathbf{Z}_{k_0}^{k-1}$ and λ. In particular, $f\left(\mathbf{z}_k|\mathbf{Z}_{k_0}^{k-1}, \lambda\right)$ is parameterized by the mean vector [75]

$$\begin{aligned} \boldsymbol{\mu}_k &= \mathbf{C}E\left[\mathbf{x}_k|\mathbf{Z}_{k_0}^{k-1}, \lambda\right] \\ &= \mathbf{C}\hat{\mathbf{x}}_{k|k-1} = \mathbf{C}\mathbf{F}\hat{\mathbf{x}}_{k-1|k-1} \end{aligned} \tag{4.12}$$

and covariance matrix

$$\begin{aligned} \boldsymbol{\Sigma}_k &= E\left[\left(\mathbf{C}\boldsymbol{\epsilon}_{k|k-1} + \boldsymbol{\omega}_k\right)\left(\mathbf{C}\boldsymbol{\epsilon}_{k|k-1} + \boldsymbol{\omega}_k\right)^T|\mathbf{Z}_{k_0}^{k-1}, \lambda\right] \\ &= \mathbf{C}E\left[\boldsymbol{\epsilon}_{k|k-1}\boldsymbol{\epsilon}_{k|k-1}^T|\mathbf{Z}_{k_0}^{k-1}, \lambda\right]\mathbf{C}^T + E\left[\boldsymbol{\omega}_k\boldsymbol{\omega}_k^T\right] \\ &= \mathbf{C}\mathbf{Q}_{k|k-1}\mathbf{C}^T + \mathbf{I}_N \end{aligned} \tag{4.13}$$

where $\hat{\mathbf{x}}_{k-1|k-1}$ is the *a posteriori* state estimate found at time $k-1$, and the expectation in each case is over an ensemble set of source event realizations with a similar structure. Equation (4.12) uses the fact that $\boldsymbol{\omega}_k$ is independent of the source model and previous observations (regardless of their composition), while (4.13) exploits $E\left[\boldsymbol{\epsilon}_{k|k-1}\boldsymbol{\omega}_k^T|\mathbf{Z}_{k_0}^{k-1}, \lambda\right] = \mathbf{0}$ due to conditional independence of $\boldsymbol{\omega}_k$ and source signatures.

It is clear that a separate Kalman filter for each hypothesis in (4.3), except for $\mathcal{H}_0$, is needed to generate a corresponding $\boldsymbol{\zeta}_k(\boldsymbol{\theta}_k)$, as they all assume different observation compositions. Therefore, *a posteriori* values, $\hat{\mathbf{x}}_{k-1|k-1}$ and $\mathbf{Q}_{k-1|k-1}$, should only be used for estimating $\mathbf{x}_{k|k}$ for hypotheses associated with source types that were determined to be present at time $k-1$ (also see Fig. 4.1). The state vector for every other hypothesis should be estimated using $\hat{\mathbf{x}}_{k_0-1|k_0-1}$ and $\mathbf{Q}_{k_0-1|k_0-1}$, that are reinitialized values of the state parameters, and are derived in the next subsection. The reason being such hypotheses essentially assume the onset of a new source type, and hence, there are no valid past states to use. Thus, reinitialization sets k_0 for a given hypothesis, and correspondingly determines the past observations used to compute a likelihood, as in (4.5). This implies that, when there are many source types, most $\boldsymbol{\zeta}_k(\boldsymbol{\theta}_k)$ will be generated using reinitialized state parameters. However, it is shown below that this process is simple, as the initial state error covariance and state vector estimator have a closed form for each hypothesis.

4.3.3. Initializing State Parameters

This subsection derives explicit forms of $\hat{\mathbf{x}}_{k_0-1|k_0-1}$ and $\mathbf{Q}_{k_0-1|k_0-1}$, that are initial values of the state vector estimate and error covariance matrix, respectively. As mentioned before, these initial parameters are needed for generating $\boldsymbol{\zeta}_k(\boldsymbol{\theta}_k)$ for any hypothesis that assumes the presence of a given source type in $\mathbf{z}_k$ that was absent in $\mathbf{z}_{k-1}$, according to the results of applying the SRCT method to this prior observation.

Proposition 1. *The error covariance matrices under the single source hypotheses are initialized as*

$$\begin{aligned} \mathcal{H}_1^{(p)} &: \mathbf{Q}_{k_0-1|k_0-1} = \mathbf{I}_J \otimes \left(\mathbf{S}^{(p)T}\mathbf{S}^{(p)}\right)^{-1} \\ \mathcal{H}_2^{(q)} &: \mathbf{Q}_{k_0-1|k_0-1} = \mathbf{I}_J \otimes \left(\mathbf{H}^{(q)T}\mathbf{H}^{(q)}\right)^{-1} \end{aligned} \tag{4.14}$$

where $\otimes$ denotes the Kronecker product. Corresponding state vectors are initialized using (4.15) *and* (4.16).

Proof. The state vector estimate under a single source hypothesis can be initialized with the linear least squares estimates (or maximum likelihood estimates (MLE) since the observation noise is Gaussian) of the coefficients of $\mathbf{z}_{k_0-j},\ j=1,\ldots,J$ relative to the appropriate basis vectors, i.e.

$$\begin{aligned}
\mathcal{H}_1^{(p)} &: \hat{\mathbf{a}}_{k_0-j} = \left(\mathbf{S}^{(p)T}\mathbf{S}^{(p)}\right)^{-1}\mathbf{S}^{(p)T}\mathbf{z}_{k_0-j} = \mathbf{S}^{(p)\dagger}\mathbf{z}_{k_0-j} \qquad (4.15)\\
\mathcal{H}_2^{(q)} &: \hat{\mathbf{b}}_{k_0-j} = \left(\mathbf{H}^{(q)T}\mathbf{H}^{(q)}\right)^{-1}\mathbf{H}^{(q)T}\mathbf{z}_{k_0-j} = \mathbf{H}^{(q)\dagger}\mathbf{z}_{k_0-j}
\end{aligned}$$

where $\mathbf{S}^{(p)\dagger}$ and $\mathbf{H}^{(q)\dagger}$ denote the Moore-Penrose inverses [76] of $\mathbf{S}^{(p)}$ and $\mathbf{H}^{(q)}$, respectively, which always exist and may be calculated using the given explicit forms owing to the columns of $\mathbf{S}^{(p)}$ and $\mathbf{H}^{(q)}$ being linearly independent, as stated in Section 4.2. Now, the state vector (see Table 4.1) estimates are initialized using the coefficient vector estimates in (4.15) as

$$\begin{aligned}
\mathcal{H}_1^{(p)} &: \hat{\mathbf{x}}_{k_0-1|k_0-1} = \left[\hat{\mathbf{a}}_{k_0-1}^T\ \cdots\ \hat{\mathbf{a}}_{k_0-J}^T\right]^T \qquad (4.16)\\
\mathcal{H}_2^{(q)} &: \hat{\mathbf{x}}_{k_0-1|k_0-1} = \left[\hat{\mathbf{b}}_{k_0-1}^T\ \cdots\ \hat{\mathbf{b}}_{k_0-J}^T\right]^T.
\end{aligned}$$

To derive $\mathbf{Q}_{k_0-1|k_0-1} = E\left[\boldsymbol{\epsilon}_{k_0-1|k_0-1}\boldsymbol{\epsilon}_{k_0-1|k_0-1}^T\right]$ under $\mathcal{H}_1^{(p)}$, first note that in this case

$$\hat{\mathbf{x}}_{k_0-1|k_0-1} - \mathbf{x}_{k_0-1} = \begin{bmatrix} \mathbf{S}^{(p)\dagger}\mathbf{z}_{k_0-1} \\ \vdots \\ \mathbf{S}^{(p)\dagger}\mathbf{z}_{k_0-J} \end{bmatrix} - \begin{bmatrix} \mathbf{a}_{k_0-1} \\ \vdots \\ \mathbf{a}_{k_0-J} \end{bmatrix} = \begin{bmatrix} \mathbf{S}^{(p)\dagger}\boldsymbol{\omega}_{k_0-1} \\ \vdots \\ \mathbf{S}^{(p)\dagger}\boldsymbol{\omega}_{k_0-J} \end{bmatrix}.$$

Now, since $E\left[\boldsymbol{\omega}_k\boldsymbol{\omega}_{k-j}^T\right] = \mathbf{I}_N\delta(j)$, from (4.10) we have

$$\mathcal{H}_1^{(p)} : \mathbf{Q}_{k_0-1|k_0-1} = \mathbf{I}_J \otimes \mathbf{S}^{(p)\dagger}\ \mathbf{I}_N\ \left(\mathbf{S}^{(p)\dagger}\right)^T$$

which can be reduced to (4.14). Similar steps are used to derive $\mathbf{Q}_{k_0-1|k_0-1}$ under $\mathcal{H}_2^{(q)}$. □

Note that (4.15) motivates transforming the original observation to $\mathbf{z}_k = \mathbf{R}_{\mathbf{w}}^{-\frac{1}{2}}\mathbf{y}_k$, as this ensures the elements of the transformed observation noise $\boldsymbol{\omega}_k$ are uncorrelated and the coefficient estimators in (4.15) have minimum variance [77].

Proposition 2. *The error covariance matrix under a dual source hypothesis is initialized as*

$$\mathcal{H}_3^{(p,q)} : \mathbf{Q}_{k_0-1|k_0-1} = \begin{bmatrix} \mathbf{I}_J \otimes \left(\mathbf{S}^{(p)T}\mathbf{P}_{\mathbf{H}}^{(q)\perp}\mathbf{S}^{(p)}\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_J \otimes \left(\mathbf{H}^{(q)T}\mathbf{P}_{\mathbf{S}}^{(p)\perp}\mathbf{H}^{(q)}\right)^{-1} \end{bmatrix} \tag{4.17}$$

where $\mathbf{P}_{\mathbf{S}}^{(p)\perp} = \mathbf{I}_N - \mathbf{S}^{(p)}\left(\mathbf{S}^{(p)T}\mathbf{S}^{(p)}\right)^{-1}\mathbf{S}^{(p)T}$ *and* $\mathbf{P}_{\mathbf{H}}^{(q)\perp} = \mathbf{I}_N - \mathbf{H}^{(q)}\left(\mathbf{H}^{(q)T}\mathbf{H}^{(q)}\right)^{-1}\mathbf{H}^{(q)T}$ *project onto the orthogonal complements of the subspaces* $\langle \mathbf{S}^{(p)} \rangle$ *and* $\langle \mathbf{H}^{(q)} \rangle$*, respectively. The corresponding state vector is initialized using* (4.19) *and* (4.20)*.*

Proof. Ideally, under $\mathcal{H}_3^{(p,q)}$, $\hat{\mathbf{x}}_{k_0-1|k_0-1}$ should contain signal and interference coefficients that are free from the effects of each other. The MLEs of $\mathbf{a}_{k_0-j}$'s and $\mathbf{b}_{k_0-j}$'s under this two source assumption can be obtained using oblique projection [77] matrices

$$\begin{aligned} \mathbf{E}_{\mathbf{S}}^{(p,q)} &= \mathbf{S}^{(p)}\left(\mathbf{S}^{(p)T}\mathbf{P}_{\mathbf{H}}^{(q)\perp}\mathbf{S}^{(p)}\right)^{-1}\mathbf{S}^{(p)T}\mathbf{P}_{\mathbf{H}}^{(q)\perp} \\ \mathbf{E}_{\mathbf{H}}^{(p,q)} &= \mathbf{H}^{(q)}\left(\mathbf{H}^{(q)T}\mathbf{P}_{\mathbf{S}}^{(p)\perp}\mathbf{H}^{(q)}\right)^{-1}\mathbf{H}^{(q)T}\mathbf{P}_{\mathbf{S}}^{(p)\perp}. \end{aligned}$$

The matrices $\mathbf{E}_{\mathbf{S}}^{(p,q)}$ and $\mathbf{E}_{\mathbf{H}}^{(p,q)}$ have respective range spaces $\langle \mathbf{S}^{(p)} \rangle$ and $\langle \mathbf{H}^{(q)} \rangle$ and respective null spaces $\langle \mathbf{H}^{(q)} \rangle$ and $\langle \mathbf{S}^{(p)} \rangle$, and hence, we have

$$\begin{aligned} \mathcal{H}_3^{(p,q)} &: \mathbf{E}_{\mathbf{S}}^{(p,q)}\mathbf{z}_{k_0-j} = \mathbf{S}^{(p)}\mathbf{a}_{k_0-j} + \mathbf{E}_{\mathbf{S}}^{(p,q)}\boldsymbol{\omega}_{k_0-j} \\ \mathcal{H}_3^{(p,q)} &: \mathbf{E}_{\mathbf{H}}^{(p,q)}\mathbf{z}_{k_0-j} = \mathbf{H}^{(q)}\mathbf{b}_{k_0-j} + \mathbf{E}_{\mathbf{H}}^{(p,q)}\boldsymbol{\omega}_{k_0-j}. \end{aligned} \tag{4.18}$$

Now, the MLEs of the signal and interference coefficients under $\mathcal{H}_3^{(p,q)}$ are

$$\begin{aligned} \mathcal{H}_3^{(p,q)} &: \hat{\mathbf{a}}_{k_0-j} = \mathbf{S}^{(p)\dagger}\mathbf{E}_{\mathbf{S}}^{(p,q)}\mathbf{z}_{k_0-j} \\ \mathcal{H}_3^{(p,q)} &: \hat{\mathbf{b}}_{k_0-j} = \mathbf{H}^{(q)\dagger}\mathbf{E}_{\mathbf{H}}^{(p,q)}\mathbf{z}_{k_0-j} \end{aligned} \tag{4.19}$$

and the state vector estimate is initialized using these coefficients as

$$\mathcal{H}_3^{(p,q)} : \hat{\mathbf{x}}_{k_0-1|k_0-1} = \left[\hat{\mathbf{a}}_{k_0-1}^T \cdots \hat{\mathbf{a}}_{k_0-J}^T \; \hat{\mathbf{b}}_{k_0-1}^T \cdots \hat{\mathbf{b}}_{k_0-J}^T\right]^T . \tag{4.20}$$

Since steps to find $\mathbf{Q}_{k_0-1|k_0-1}$ under $\mathcal{H}_3^{(p,q)}$ are similar to those under $\mathcal{H}_1^{(p)}$, we have

$$\hat{\mathbf{x}}_{k_0-1|k_0-1} - \mathbf{x}_{k_0-1} = \begin{bmatrix} \hat{\mathbf{a}}_{k_0-1} \\ \vdots \\ \hat{\mathbf{a}}_{k_0-J} \\ \hat{\mathbf{b}}_{k_0-1} \\ \vdots \\ \hat{\mathbf{b}}_{k_0-J} \end{bmatrix} - \begin{bmatrix} \mathbf{a}_{k_0-1} \\ \vdots \\ \mathbf{a}_{k_0-J} \\ \mathbf{b}_{k_0-1} \\ \vdots \\ \mathbf{b}_{k_0-J} \end{bmatrix} = \begin{bmatrix} \mathbf{S}^{(p)\dagger}\mathbf{E}_{\mathbf{S}}^{(p,q)}\boldsymbol{\omega}_{k_0-1} \\ \vdots \\ \mathbf{S}^{(p)\dagger}\mathbf{E}_{\mathbf{S}}^{(p,q)}\boldsymbol{\omega}_{k_0-J} \\ \mathbf{H}^{(q)\dagger}\mathbf{E}_{\mathbf{H}}^{(p,q)}\boldsymbol{\omega}_{k_0-1} \\ \vdots \\ \mathbf{H}^{(q)\dagger}\mathbf{E}_{\mathbf{H}}^{(p,q)}\boldsymbol{\omega}_{k_0-J} \end{bmatrix} \tag{4.21}$$

which follows from (4.18). Using the fact that

$$\begin{aligned} \mathbf{S}^{(p)\dagger}\mathbf{E}_{\mathbf{S}}^{(p,q)}\mathbf{E}_{\mathbf{S}}^{(p,q)T}\left(\mathbf{S}^{(p)\dagger}\right)^T &= \left(\mathbf{S}^{(p)T}\mathbf{P}_{\mathbf{H}}^{(q)\perp}\mathbf{S}^{(p)}\right)^{-1} \\ \mathbf{H}^{(q)\dagger}\mathbf{E}_{\mathbf{H}}^{(p,q)}\mathbf{E}_{\mathbf{H}}^{(p,q)T}\left(\mathbf{H}^{(q)\dagger}\right)^T &= \left(\mathbf{H}^{(q)T}\mathbf{P}_{\mathbf{S}}^{(p)\perp}\mathbf{H}^{(q)}\right)^{-1} \end{aligned}$$

the covariance of the error term in (4.21) is found to be as shown in (4.17). □

Since $\mathbf{z}_k$'s are processed sequentially, the initial state estimates in (4.16) and (4.20) are formed using the J observations prior to a given $\mathbf{z}_{k_0}$. While these observations might not contain strong signatures of the sources associated with these estimates, these are the best linear estimates of the coefficients that can be obtained.

All components necessary for generating $\boldsymbol{\theta}_k$ for a given λ have now been provided, which may subsequently be used to generate a corresponding $\boldsymbol{\zeta}_k(\boldsymbol{\theta}_k)$ using (4.8). The variables $\boldsymbol{\zeta}_k(\boldsymbol{\theta}_k)$'s may then be used to form test statistics for the hypothesis test in (4.3), as described next.

4.3.4. Determining Observation Composition Using LLRTs

We may now introduce the LLRTs needed to assign a class label vector, $\mathbf{c}_k = \left[\hat{\alpha}_k p,\ \hat{\beta}_k q\right]$, with $p \in \{1, \ldots, P\}$ and $q \in \{1, \ldots, Q\}$, consisting of one signal and one interference label, to $\mathbf{z}_k$. Here " $\hat{}$ " denotes estimate of the indicator variables in (4.1) that determine the presence of signal and/or interference.

Detection and Classification: As indicated by Fig. 4.1, the first steps for processing $\mathbf{z}_k$ are source detection, where it is decided whether to accept or reject $\mathcal{H}_0$, and source classification, where estimates of the most likely signal and interference source types, p^* and q^*, are found. This is accomplished by forming a set of LLR, each of which may be used to test the hypothesis that $\mathbf{z}_k$ consists of noise alone versus the hypothesis that $\mathbf{z}_k$ contains noise plus a type p signal source and/or a type q interference source. Using the general form in (4.7) with $\boldsymbol{\theta}_k$'s generated using the appropriate models, this LLR is given by

$$\Lambda_k\left(\boldsymbol{\theta}_{3,k}^{(p,q)}, \boldsymbol{\theta}_{0,k}\right) = \boldsymbol{\zeta}_k\left(\boldsymbol{\theta}_{0,k}\right) - \boldsymbol{\zeta}_k\left(\boldsymbol{\theta}_{3,k}^{(p,q)}\right) \tag{4.22}$$

where $\boldsymbol{\theta}_{0,k}$ and $\boldsymbol{\theta}_{3,k}^{(p,q)}$ are the distribution parameters corresponding to models λ_0 and $\lambda_3^{(p,q)}$, respectively. Detection of sources in $\mathbf{z}_k$ is then performed using the LLRT

$$\max_{(p,q)} \Lambda_k\left(\boldsymbol{\theta}_{3,k}^{(p,q)}, \boldsymbol{\theta}_{0,k}\right) \underset{\text{accept } \mathcal{H}_0}{\overset{\text{reject } \mathcal{H}_0}{\substack{\geq \\ <}}} \gamma \tag{4.23}$$

where γ is a predetermined threshold that adjusts the sensitivity of the detector. Estimates of the signal and interference source types (classification) are then found using the test statistics already calculated for detection as

$$(p^*, q^*) = \arg\max_{(p,q)} \Lambda_k\left(\boldsymbol{\theta}_{3,k}^{(p,q)}, \boldsymbol{\theta}_{0,k}\right) \geq \gamma. \tag{4.24}$$

Assuming a transient source has been detected at time k, the dominant source and source quantity tests outlined below can subsequently be initiated. In this case, (4.3) is reduced to testing between $\mathcal{H}_1^{(p^*)}$, $\mathcal{H}_2^{(q^*)}$, and $\mathcal{H}_3^{(p^*,q^*)}$. When hypothesis $\mathcal{H}_0$ is accepted, no further processing is required for $\mathbf{z}_k$ and the next observation $\mathbf{z}_{k+1}$ is evaluated. In this case, we have $\hat{\alpha}_k = \hat{\beta}_k = 0$ so that $\mathbf{c}_k = [0,\ 0]$. Justification for using (4.24) to find p^* and q^* is based on the fact that, for source types that were determined to be absent at time $k-1$, the corresponding $\boldsymbol{\zeta}_k\left(\boldsymbol{\theta}_{3,k}^{(p,q)}\right)$'s are found using reinitialized state variables, generated using oblique projections to remove the effects of type q interference from the type p signal estimate and vice versa (see (4.19) in Section 4.3.3). For sources types that were present at time $k-1$, $\boldsymbol{\zeta}_k\left(\boldsymbol{\theta}_{3,k}^{(p,q)}\right)$ is formed using the best (in the mean squared error sense) linear estimates of the source signatures.

Dominant Source Test: When $\mathcal{H}_0$ is rejected for $\mathbf{z}_k$ the next step involves determining whether the dominant source is a type p^* signal or a type q^* interference (see Fig. 4.1). Note that this step does not exclude the possibility that both sources are simultaneously present, but rather it defines the most likely source in the single source case. To this end, the LLRT for determining the dominant source is given by

$$\Lambda_k\left(\boldsymbol{\theta}_{1,k}^{(p^*)},\boldsymbol{\theta}_{2,k}^{(q^*)}\right) = \boldsymbol{\zeta}_k\left(\boldsymbol{\theta}_{2,k}^{(q^*)}\right) - \boldsymbol{\zeta}_k\left(\boldsymbol{\theta}_{1,k}^{(p^*)}\right) \quad \begin{array}{c}\text{reject } \mathcal{H}_2^{(q^*)} \\ \geq \\ < \\ \text{reject } \mathcal{H}_1^{(p^*)}\end{array} \quad \tau \tag{4.25}$$

where $\boldsymbol{\theta}_{1,k}^{(p^*)}$ and $\boldsymbol{\theta}_{2,k}^{(q^*)}$ are the distribution parameters corresponding to models $\lambda_1^{(p^*)}$ and $\lambda_2^{(q^*)}$, respectively, and τ is a predetermined threshold that may be adjusted based on the risk associated with missing signals in a given application.

Source Quantity Test: Following the dominant source test, the remaining single source hypothesis that has not been rejected ($\mathcal{H}_1^{(p^*)}$ or $\mathcal{H}_2^{(q^*)}$) is tested against the remaining dual source

Table 4.2. Class labels and indicator estimates under each hypothesis.

	Accepted Hypothesis			
	$\mathcal{H}_0$	$\mathcal{H}_1^{(p^*)}$	$\mathcal{H}_2^{(q^*)}$	$\mathcal{H}_3^{(p^*,q^*)}$
$\left(\hat{\alpha}_k,\ \hat{\beta}_k\right)$	(0, 0)	(1, 0)	(0, 1)	(1, 1)
$\mathbf{c}_k$	[0, 0]	$[p^*,\ 0]$	$[0,\ q^*]$	$[p^*,\ q^*]$

hypothesis, $\mathcal{H}_3^{(p^*,q^*)}$. The LLRT for making this decision is

$$F_k \quad \underset{\substack{\text{reject } \mathcal{H}_3^{(p^*,q^*)} \\ \left(\text{accept } \mathcal{H}_1^{(p^*)} \text{ or } \mathcal{H}_2^{(q^*)}\right)}}{\overset{\text{accept } \mathcal{H}_3^{(p^*,q^*)}}{\gtrless}} \quad \eta \tag{4.26}$$

where η is a predetermined threshold that should also be set based on the risk associated with missing (weaker) signals and

$$F_k = \begin{cases} \Lambda_k\left(\boldsymbol{\theta}_{3,k}^{(p^*,q^*)}, \boldsymbol{\theta}_{1,k}^{(p^*)}\right), & \text{when } \mathcal{H}_2^{(q^*)} \text{ rejected} \\ \\ \Lambda_k\left(\boldsymbol{\theta}_{3,k}^{(p^*,q^*)}, \boldsymbol{\theta}_{2,k}^{(q^*)}\right), & \text{when } \mathcal{H}_1^{(p^*)} \text{ rejected} \end{cases}. \tag{4.27}$$

This means that the single source log-likelihood value used to form F_k depends on the results of the dominant source test in (4.25) (also see Fig. 4.1). Note that, at this stage, all $\boldsymbol{\zeta}_k(\boldsymbol{\theta}_k)$'s have already been calculated for use in other LLRTs, and hence, these values may be used in (4.27) to find F_k.

Upon completion of the source quantity test (or when no sources are detected in $\mathbf{z}_k$), a class label $\mathbf{c}_k$ may be assigned to this observation that depends on the accepted hypothesis according to Table 4.2, which also provides the values of the estimated indicator variables, $\hat{\alpha}_k$ and $\hat{\beta}_k$, in each case. The entire process in Fig. 4.1 may then be continually applied to subsequent observations $\mathbf{z}_{k+j}$'s indefinitely or until all available observations have been processed.

Source Estimation: Once $\mathbf{c}_k$ has been determined for $\mathbf{z}_k$, estimates of the identified source signatures may then be obtained from the state estimates for the accepted hypothesis, if desired. Specifically, the coefficient estimates, $\hat{\mathbf{a}}_k$ and/or $\hat{\mathbf{b}}_k$, for time k, can be extracted from $\hat{\mathbf{x}}_{k|k}$ and used in conjunction with their respective basis matrices, $\mathbf{S}^{(p^*)}$ and/or $\mathbf{H}^{(q^*)}$, to yield estimates of the vectors representing the signal and interference components, denoted by $\hat{\mathbf{s}}_k$ and $\hat{\mathbf{h}}_k$, respectively. In particular, these estimates are found (see Table 4.1) as

$$\begin{aligned}
\mathcal{H}_1^{(p^*)} &: \hat{\mathbf{s}}_k = \mathbf{R}_{\mathbf{w}}^{\frac{1}{2}}\mathbf{S}^{(p^*)}\mathbf{D}^T\hat{\mathbf{x}}_{k|k} \\
\mathcal{H}_2^{(q^*)} &: \hat{\mathbf{h}}_k = \mathbf{R}_{\mathbf{w}}^{\frac{1}{2}}\mathbf{H}^{(q^*)}\mathbf{D}^T\hat{\mathbf{x}}_{k|k} \\
\mathcal{H}_3^{(p^*,q^*)} &: \hat{\mathbf{s}}_k = \begin{bmatrix}\mathbf{R}_{\mathbf{w}}^{\frac{1}{2}}\mathbf{S}^{(p^*)} & \mathbf{0}_{N\times M}\end{bmatrix}\mathbf{D}^T\hat{\mathbf{x}}_{k|k} \\
&\quad\; \hat{\mathbf{h}}_k = \begin{bmatrix}\mathbf{0}_{N\times M} & \mathbf{R}_{\mathbf{w}}^{\frac{1}{2}}\mathbf{H}^{(q^*)}\end{bmatrix}\mathbf{D}^T\hat{\mathbf{x}}_{k|k}
\end{aligned} \tag{4.28}$$

where each basis matrix is premultiplied by $\mathbf{R}_{\mathbf{w}}^{\frac{1}{2}}$ since their columns are basis vectors for the transformed observation space. In the case $\mathcal{H}_3^{(p^*,q^*)}$ is accepted, a type of source separation is performed by producing estimates, $\hat{\mathbf{s}}_k$ and $\hat{\mathbf{h}}_k$, that represent the separated signal and interference components of the original observation $\mathbf{y}_k$, respectively.

4.4. Conclusions

This chapter introduced the first of two methods intended to address the problem of detecting, classifying, and estimating the signatures of random transient acoustical sources from sequential multivariate data, where a signal and an interference source, both of unknown type, may be simultaneously present. In such problems, it can be difficult to quantify the effects of the presence of one source on the detection of the other since all sources have unknown scaling and times of arrival, and their signatures may overlap to varying degrees. The SRCT method deals with these complications using a series of LLRT, designed specifically to discover the composition of observations that follow the proposed model. Specifically, detection and classification are handled with one LLRT, while two other tests determine the composition of the observation in terms of signal and interference, assuming the detection stage rejected the null hypothesis. Each test uses the log-likelihoods of distribution parameters under specific source composition hypotheses, which incorporate known information about the dependencies between basis coefficients for each source type, thus yielding a cohesive framework that satisfies all of the requirements mentioned in Section 1.3. The SRCT method may also be used for estimating the signatures of sources that are determined to be present, thus performing separation of signal and interference components in the case of superimposed signatures.

The SRCT method is the first of two approaches introduced in this thesis that offer comprehensive solutions to the problem at hand. The advantages and disadvantages of each method are discussed in more detail in Chapter 6. For time time being, we note that SRCT is most useful when priorities are relatively low computational complexity, having a relatively small number of training parameters that must be chosen, and having the ability to assign class labels to both signal and interference sources, rather than just the former.

CHAPTER 5

A Sparse Coefficient State Tracking Framework

5.1. Introduction

Chapter 3 introduced two existing fundamental approaches for detecting and classifying transient sources, with the implication that they were ideal candidates to use as building blocks for constructing more comprehensive solutions to this problem. The previous chapter introduced a particular implementation of one of these fundamental approaches, called the sequential random coefficient tracking (SRCT) method, that detects and classifies sources separately in individual observations. The SRCT method satisfies all of the requirements of successful solutions to the soundscape characterization problem outlined in Section 1.3 and, in the next chapter, is shown to perform well on the data in Chapter 2. On the other hand, the SRCT method assumes that only one type of interference may be present at a time, which is an impractical assumption for some soundscape data containing an abundance of signatures associated with different types of wildlife and weather effects. Additionally, the linear autoregressive basis coefficient model used by this method can fail to capture subtle or novel variations in acoustical events, leading to less accurate estimates.

This chapter introduces a new transient source characterization method based on an extension of the other fundamental approach detailed in Chapter 3, namely the cumulative sum (CUSUM) procedure that implements a sequential version of a log-likelihood ratio (LLR) test. This sparse coefficient state tracking (SCST) method was designed to overcome the main issues of the SRCT approach mentioned above, by drawing from the concepts of classification in a sparse domain [48–51] and modeling of sparse atom coefficients [30, 56–58]. The main advantage of the SCST method is its applicability to data containing signal, interference, and noise components that may not necessarily follow models based on convenient parametric distributions, e.g., multivariate Gaussian. To simplify the data representation, sparse coding and quantization steps are first applied to each incoming observation. This allows for using a Bayesian network (BN) [78] to model the temporal evolution of typical acoustical events. The likelihoods of BNs corresponding to different signal types and noise may then be used to form a set of cumulative test statistics for detection and classification of multiple transient signal events.

This chapter is organized as follows. Section 5.2 describes the problem formulation in the original data space, including the observation model and generalized likelihood ratios tests (GLRT) used for detection and classification of signals. Since implementation of this framework in the original data space is difficult and/or impractical for many applications, Section 5.3 introduces the process for obtaining sparse coefficient state data representations, as well as the associated reformulations of the GLRTs, thus yielding the proposed SCST method. Finally, Section 5.4 provides concluding remarks.

5.2. Detection and Classification of Transient Events - Original Data Space

The goal of this section is to develop the basic framework for performing detection and classification in the original sequential multivariate data space using an extension of the CUSUM framework in Section 3.2.2. The benefits of simplifying the data are then discussed to motivate the development of the SCST method in the next section, which still relies on the primary underlying mechanics discussed below.

5.2.1. Observation Model and Detection and Classification Hypotheses

Let $\mathbf{Y}_1^n = \{\mathbf{y}_k\}_{k=1}^n$, $n = 1, 2, \ldots$ be the observation sequence recorded as of the current time n, where $\mathbf{y}_k \in \mathbb{R}^N$ is the observation at time k. Data arrives continually, meaning n is increasing. Detecting and classifying multiple transient signals requires two distinct phases: 1) signal detection to look for the presence of a signal while it is assumed that none are present, and 2) quiescent detection to look for observations that contain no signal while it is assumed that one is present.

The idea is to alternate between these two phases as new $\mathbf{y}_n$'s arrive, while performing classification by exploiting all available information within a given detected signal event. Note that this phase switching is conceptually similar to the approach used in Section 3.3 for sequential fusion of classification decisions.

Since signals are continually detected and classified, it is helpful to adopt notation associated with the onset of various detection periods, relative to the current time n. Let k_0 and k_1 denote the unknown onset times of the next quiescent and signal periods, respectively, and let $\hat{k}_0$ and $\hat{k}_1$ denote the estimated (known) onset times for the most recently detected quiescent and signal periods, respectively. Fig. 5.1 demonstrates the two-phase concept by showing the circumstances for implementing each phase, as well as the most recent estimated onset times relative to the current time n. This figure also shows the test statistics for each phase, which are discussed in Section 5.2.2.

When the data has been in a quiescent period since time $\hat{k}_0$, signal detection and classification is performed on each $\mathbf{y}_n$ according to the following multiple hypotheses test

$$\mathcal{H}_0 : \mathbf{y}_k = \beta_k \sum_{q=1}^{Q} \mathbf{h}_k^{(q)} + \mathbf{w}_k, \ \hat{k}_0 \leq k \leq n \tag{5.1}$$

$$\mathcal{H}_1^{(p)} : \mathbf{y}_k = \begin{cases} \beta_k \sum_{q=1}^{Q} \mathbf{h}_k^{(q)} + \mathbf{w}_k, & \hat{k}_0 \leq k < k_1 \\ \mathbf{s}_k^{(p)} + \beta_k \sum_{q=1}^{Q} \mathbf{h}_k^{(q)} + \mathbf{w}_k, & k_1 \leq k \leq n \end{cases}$$

where $\mathbf{s}_k^{(p)}$ is a random class $p \in [1, P]$ signal vector, $\mathbf{h}_k^{(q)}$ is a random class $q \in [1, Q]$ interference vector, β_k is a binary variable indicating the presence ($\beta_k = 1$) or absence ($\beta_k = 0$) of interference, and $\mathbf{w}_k$ is an independent and identically distributed (IID) noise vector with $E[\mathbf{w}_k] = \mathbf{0}$. As can be seen, under $\mathcal{H}_1^{(p)}$ the onset of signal components $\mathbf{s}_k^{(p)}$, $k \in [k_1, n]$ occurs at the unknown time k_1, and the goal is to find the new estimate $\hat{k}_1$, as well as the class of this signal. Unlike

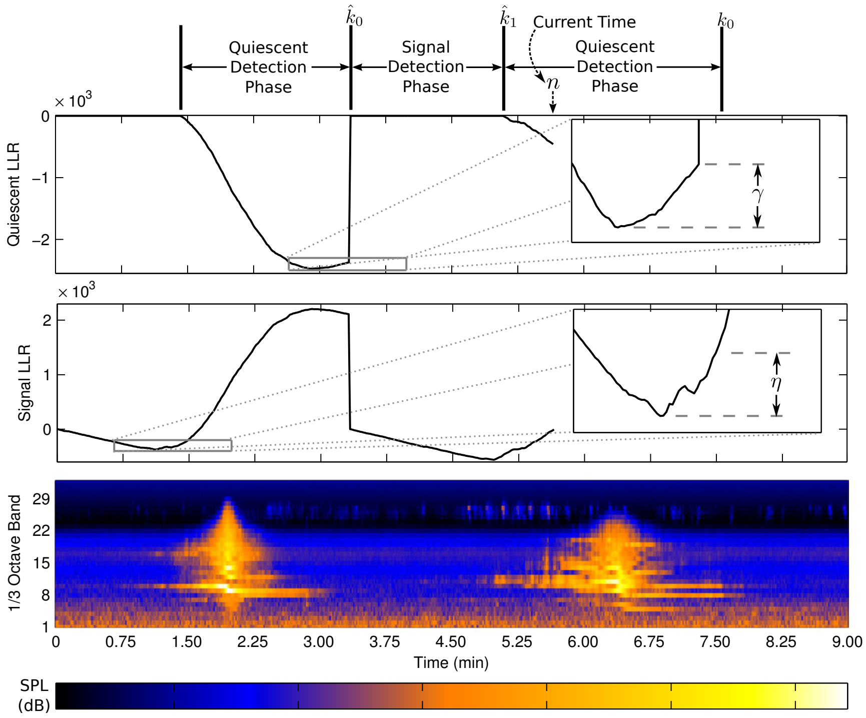


Figure 5.1. Illustration of the two phase detection approach, where the durations of several phases are shown above a 1/3 octave observation sequence (bottom), and the corresponding test statistics used to detect signal (middle) and quiescent (top) periods. The times where the onset of a quiescent period and a signal event were last detected, denoted $\hat{k}_0$ and $\hat{k}_1$, respectively, are shown relative to the current time n. A signal and quiescent period are detected when their associated LLRs increase by at least η and γ, respectively.

in Chapter 4, interference is considered purely as a nuisance in this chapter, meaning it is not specifically detected and classified. The summation over $\mathbf{h}_k^{(q)}$'s indicates that multiple types of interference may be simultaneously present, where $\mathbf{h}_k^{(q)} = \mathbf{0}$ if class q interference is absent from $\mathbf{y}_k$. As mentioned in Chapter 1, interference differs from noise in several ways, namely it is 1) typically not IID or zero mean, 2) associated with a specific set of acoustical sources that are usually not of interest, and 3) not assumed to always be present, i.e., $\beta_k = 0$ may be true.

The class label assigned to $\mathbf{y}_n$ is denoted by $c_n \in [0, P]$ at time n, where $c_n = 0$ means $\mathbf{y}_n \in \mathcal{H}_0$ and $c_n = p$ means $\mathbf{y}_n \in \mathcal{H}_1^{(p)}$. As discussed below, event-wide classification is performed, meaning the same label is assigned to all consecutive samples for which a signal detection occurred only after the next quiescent period has been detected. The reason for this is that, due to the assumed random and time-varying nature of signals, some events associated with different signal types may appear similar for subsets of their observations. Therefore, more accurate labels are assigned when taking into account the likelihood of each signal model over the course of all observations associated with an event (signatures of one signal). To facilitates this classification framework it is assumed that, under $\mathcal{H}_1^{(p)}$, only one signal is present in $\mathbf{y}_n$, meaning the signatures of two signals will never be superimposed. Furthermore, it is assumed that two different signals cannot be present in adjacent observations, i.e., $c_{n-1} = p \Rightarrow c_n \in \{0, p\}$.

Since transient signals have finite extent, the next quiescent period must be detected before the process of detecting the next signal can begin. Therefore, when a signal has been present since time $\hat{k}_1$, the following hypothesis test is used in place of (5.1) to perform quiescent detection

$$\mathcal{H}_1^{(p)} : \mathbf{y}_k = \mathbf{s}_k^{(p)} + \beta_k \sum_{q=1}^{Q} \mathbf{h}_k^{(q)} + \mathbf{w}_k, \ \hat{k}_1 \leq k \leq n \tag{5.2}$$

$$\mathcal{H}_0 : \mathbf{y}_k = \begin{cases} \mathbf{s}_k^{(p)} + \beta_k \sum_{q=1}^{Q} \mathbf{h}_k^{(q)} + \mathbf{w}_k, & \hat{k}_1 \leq k < k_0 \\ \beta_k \sum_{q=1}^{Q} \mathbf{h}_k^{(q)} + \mathbf{w}_k, & k_0 \leq k \leq n \end{cases}$$

i.e., $\mathbf{s}_k^{(p)}$'s cease to be extant at the unknown time k_0 under $\mathcal{H}_0$. In summary, signal and quiescent detection are performed when $\mathbf{y}_{n-1} \in \mathcal{H}_0$ and $\mathbf{y}_{n-1} \in \mathcal{H}_1^{(p)}$, respectively.

5.2.2. GLRTs for Hypothesis Testing

5.2.2.1. *Signal Detection.* Throughout the remainder of this section, it is assumed that $\beta_k = 0$, $\forall k$ in (5.1) and (5.2), i.e., interference is not present. This stems from the fact that the SCST method addresses interference through the use of an alternate data representation, which is presented in Section 5.3. To implement the hypothesis test in (5.1) consider the log-likelihood ratio

(LLR) for the general null and alternative hypothesis parameter sets, denoted by λ_0 and λ_p, respectively, given the data $\mathbf{Y}_{\hat{k}_0}^n$

$$\begin{aligned} L_{\hat{k}_0}^n\left(\lambda_p, k_1\right) &= \ln\left(\frac{\ell\left(\lambda_p, k_1; \mathbf{Y}_{\hat{k}_0}^n\right)}{\ell\left(\lambda_0, k_1; \mathbf{Y}_{\hat{k}_0}^n\right)}\right) \qquad (5.3) \\ &= \ln\left(\frac{f_{\lambda_0}\left(\mathbf{Y}_{\hat{k}_0}^{k_1-1}\right) f_{\lambda_p}\left(\mathbf{Y}_{k_1}^n\right)}{f_{\lambda_0}\left(\mathbf{Y}_{\hat{k}_0}^{k_1-1}\right) f_{\lambda_0}\left(\mathbf{Y}_{k_1}^n\right)}\right) = \ln\left(\frac{f_{\lambda_p}\left(\mathbf{Y}_{k_1}^n\right)}{f_{\lambda_0}\left(\mathbf{Y}_{k_1}^n\right)}\right) \end{aligned}$$

where $f_\lambda\left(\mathbf{Y}_{\hat{k}_0}^n\right)$ is a general probability distribution modeled by the parameter set $\lambda \in \{\lambda_0, \lambda_p\}$. Note that (5.3) is similar to the LLR implemented by the CUSUM method in Section 3.2.2, except that dependent observations are assumed here. This LLR is a function of two unknowns, namely the change time k_1 and the signal parameter set λ_p. The second equality in (5.3) comes from assuming that, under $\mathcal{H}_1^{(p)}$, $\mathbf{y}_n$'s before and after the unknown change time k_1 are independent, while all $\mathbf{y}_n$'s are independent under $\mathcal{H}_0$. These assumptions are the reason $\beta_k = 0$ in this section, namely since interference is typically not IID, thus invalidating the second equality when $\beta_k = 1$.

To implement (5.1), consider the GLRT for change detection with an unknown signal parameter set after the hypothesis change [33]

$$\max_{\hat{k}_0 \le k \le n} \max_p L_{\hat{k}_0}^n\left(\lambda_p, k\right) \underset{\mathbf{y}_n \in \mathcal{H}_0}{\overset{\mathbf{y}_n \notin \mathcal{H}_0}{\gtrless}} \eta \qquad (5.4)$$

where $\eta > 0$ is a predetermined signal detection threshold. Double maximization makes this test generalized and states that a signal is detected when any $L_{\hat{k}_0}^n\left(\lambda_p, \hat{k}_0\right)$ (i.e., for any $p \in [1, P]$) increases by at least η [33], and the earliest time this level of increase is witnessed marks the estimated signal onset time $\hat{k}_1$. This concept is illustrated by the plot of the signal detection statistic in the middle of Fig. 5.1, which shows the LLR for one signal type (plane) increasing as new observations containing its signatures arrive, but decreasing otherwise.

5.2.2.2. *Quiescent Detection.* Recall that a complete solution must account for the inevitably that a detected signal will cease to be extant. This process is simplified by assuming that immediate switching from one signal class to another will not be encountered in real-life cases. This involves the test in (5.2), which uses the LLR

$$\begin{aligned} F_{\hat{k}_1}^{n}\left(\lambda_{p^*}, k_0\right) &= \ln\left(\frac{\ell\left(\lambda_0, k_0; \mathbf{Y}_{\hat{k}_1}^{n}\right)}{\ell\left(\lambda_{p^*}, k_0; \mathbf{Y}_{\hat{k}_1}^{n}\right)}\right) \\ &= \ln\left(\frac{f_{\lambda_0}\left(\mathbf{Y}_{k_0}^{n}\right)}{f_{\lambda_{p^*}}\left(\mathbf{Y}_{k_0}^{n}|\mathbf{Y}_{\hat{k}_1}^{k_0-1}\right)}\right) \end{aligned} \tag{5.5}$$

where

$$p^* = \arg\max_{p} \max_{\hat{k}_0 \le k \le n} L_{\hat{k}_0}^{n}\left(\lambda_p, k\right) \tag{5.6}$$

is the maximum likelihood (ML) signal model at time n, i.e., λ_{p^*} satisfies (5.4). This means that quiescent detection is performed relative to the most likely signal class, though classification is only performed when the ML signal is no longer extant.

The test used to implement (5.2) is

$$\max_{\hat{k}_1 \le k \le n} F_{\hat{k}_1}^{n}\left(\lambda_{p^*}, k\right) \quad \begin{array}{c} \mathbf{y}_n \in \mathcal{H}_0 \\ \geq \\ < \\ \mathbf{y}_n \notin \mathcal{H}_0 \end{array} \quad \gamma \tag{5.7}$$

where $\gamma > 0$ is a predetermined quiescent detection threshold and maximization is performed with respect to the unknown onset time of the next quiescent period k_0. Equation (5.7) states that $\mathcal{H}_0$ is again accepted for samples starting at time n (i.e. $\hat{k}_0 = n$) if $F_{\hat{k}_1}^{n}\left(\lambda_{p^*}, \hat{k}_1\right)$ has increased by at least γ at this time. This concept is illustrated by the top plot in Fig. 5.1, which shows the quiescent LLR decreasing when a signal is present, but increasing during times leading up to detection of a quiescent period when signal components are absent. Note that this LLR is zero during signal detection phases since it is not used during these times.

5.2.2.3. *Signal Classification.* To exploit all available evidence for making classification decisions, labels are only assigned to a set of observations associated with the most recently detected event after using (5.7) to again accept $\mathcal{H}_0$, i.e., end of extant. The assigned label p^* corresponds to the ML model parameter set λ_{p^*} (as in (5.6)) at time $\hat{k}_0 - 1$, i.e., the time step immediately preceding the start of the newly detected quiescent period. More formally $\{c_k\}_{k=\hat{k}_1}^{\hat{k}_0-1} = p^*$.

5.2.3. Practical Considerations

As long as a reasonable method for calculating likelihoods $\ell\left(\lambda_p, k; \mathbf{Y}_{\hat{k}_0}^n\right)$, $\forall p$ can be devised, the approach outlined in this section can be used to continually detect and classify multiple signals in the presence of noise. The practicality of forming the required LLRs is largely dependent on the chosen parameterization λ_p. Defining $f_{\lambda_p}\left(\mathbf{Y}_{\hat{k}_0}^n\right)$ as a prior probability distribution parameterized by λ_p is generally infeasible without assuming independence of $\mathbf{y}_n$'s under each $\mathcal{H}_1^{(p)}$, since n is always increasing. Since the presence of interference has not yet been considered, it is possible to fit, e.g., an HMM to $\mathbf{y}_n$'s under $\mathcal{H}_1^{(p)}$, and use an approach similar to that in [35] for detecting and classifying signals. However, as mentioned in Section 1.3, the intermittent presence of multiple types of interference is a major consideration for the soundscape characterization problem considered in this thesis, and this leads to difficulties when using HMMs. In particular, if a single HMM is designed to model one signal type superimposed with different combinations of interference, the variations in the data to be modeled would be vast and difficult to capture with a set of multivariate probability distributions. On the other hand, using separate HMMs to model each unique combination of interference could lead to an abundance of models, depending on how many interference sources are considered, and frequent switching between these models as new observations are evaluated. As described in the next section, the idea behind the SCST method is to use sparse coding to simplify the temporal dependencies in the data as well as remove a large portion of the interference components, i.e., separability of signal and interference components in the sparse domain is assumed. This allows for efficient likelihood calculation without making extensive assumptions about the structure of signals to be detected.

5.3. Detection and Classification of Transient Events - Sparse Coefficient State Space

The SCST method is implemented according to the block diagram in Fig. 5.2. As can be seen, each incoming data vector $\mathbf{y}_n$ is first transformed to $\mathbf{z}_n$ using sparse coding and coefficient quantization, in order to simplify the relationships between observations as well as the values they can assume, respectively, as discussed in Sections 5.3.1 and 5.3.2. Essentially, these steps provide a realistic and flexible means for calculating the likelihoods of λ's given representative data. These likelihood values may then be updated as detailed in Section 5.3.3. As mentioned before, in this section it is assumed that $\beta_k = 1$ in (5.1) and (5.2), meaning multiple types of interference may be present at any time. Robustness to this interference is inherently handled during the sparse coding stage, as the signal and interference components of the observation can be mostly separated and associated with different atoms in the dictionary. The process then proceeds as in Section 5.2, where LLRs are used to perform signal and quiescent detection, though here $\mathbf{z}_n$'s are used in place of $\mathbf{y}_n$'s.

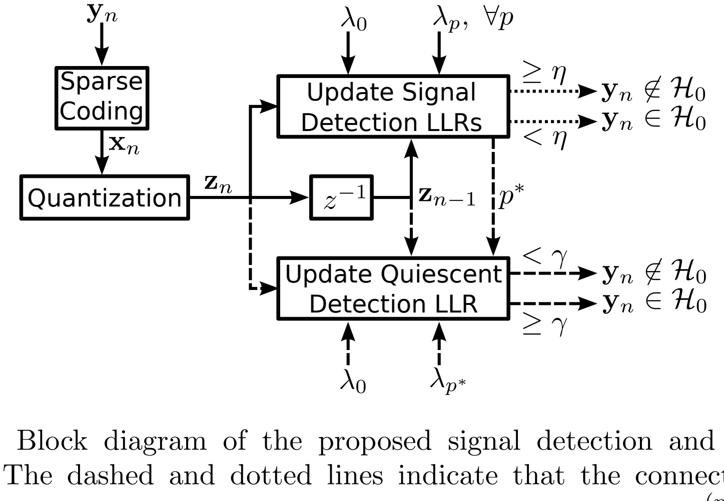


Figure 5.2. Block diagram of the proposed signal detection and classification framework. The dashed and dotted lines indicate that the connected processes are only executed during the quiescent detection (when $\mathbf{y}_{n-1} \in \mathcal{H}_1^{(p)}$) and signal detection (when $\mathbf{y}_{n-1} \in \mathcal{H}_0$) phases, respectively.

5.3.1. Sparse Coding

To simplify the dependencies between consecutive observation vectors and make the structure of nonstationary acoustical events more tractable, the SCST method first finds a sparse approximation of $\mathbf{y}_n$ when it arrives, denoted by $\mathbf{x}_n = [x_{1,n} \; \cdots \; x_{i,n} \; \cdots \; x_{M,n}]^T \in \mathbb{R}^M$. An underlying assumption is that any $\mathbf{s}_n^{(p)}$ or $\mathbf{h}_n^{(q)}$ that may be observed admits a sparse representation over some rank N dictionary matrix $\mathbf{A} = [\mathbf{A}_s, \mathbf{A}_h] \in \mathbb{R}^{N \times M}$, with $N \ll M$ and normalized columns (atoms). Furthermore, it is assumed that the atoms typically used to provide sparse representations of $\mathbf{s}_n^{(p)}$'s are mostly disjoint from those used to represent $\mathbf{h}_n^{(q)}$'s, in order to provide separability of these different components. This implies that signal and interference components can be represented in terms of two dictionaries, i.e., $\mathbf{A}_s$ and $\mathbf{A}_h$, respectively, that are relatively incoherent [79]. Note that some overlap between the atoms used to represent these two components is inevitable in many cases, but reasonably small overlap will typically not lead to a large decrease in performance. Further details concerning the recommended structure for $\mathbf{A}$ are discussed below. Apart from signal and interference separability, the merit of using $\mathbf{x}_n$'s is that they will contain many coefficients close or equal to zero. This means $x_{i,n}$ will be dependent on a relatively small set of other $x_{i',n-j}$ with time lag $j \geq 0$, and hence, the temporal evolution of the sequence $\mathbf{X}_1^n = \{\mathbf{x}_k\}_{k=1}^n$ will be easier to model and track than that of $\mathbf{Y}_1^n$.

To generate $\mathbf{x}_n$, consider the underdetermined linear system $\mathbf{A}\mathbf{v} = \mathbf{y}_n$, which has infinitely many solutions $\mathbf{v}$, meaning constraints are required to find a unique solution. Since sparsity is desired and observations are noisy, an intuitive approach is to find $\mathbf{x}_n$ using the following optimization problem [54, 80]

$$\mathbf{x}_n = \min_{\mathbf{v}} \; \|\mathbf{v}\|_0 \;\text{ subject to }\; \|\mathbf{y}_n - \mathbf{A}\mathbf{v}\|_2 \leq \delta \tag{5.8}$$

where $\delta \geq 0$ is an error tolerance proportional to $\|\mathbf{w}_n\|_2$ [54], and $\|\cdot\|_0$ is the ℓ_0-norm. The motivation for permitting a discrepancy of δ between $\mathbf{y}_n$ and $\mathbf{A}\mathbf{v}$ is to extract a $\mathbf{x}_n$ that contains fewer components representing $\mathbf{w}_n$ when compared to the case of using an equality constraint.

Roughly speaking, assuming high SNR, each $\mathbf{x}_n$ may be viewed as a sparse representation of $\mathbf{y}_n - \mathbf{w}_n$.

It is well-known [54] that (5.8) is NP-hard due to non-convexity of the ℓ_0-norm. Therefore, approximate solutions to (5.8) are required for efficient processing. The SCST method is flexible in that any pursuit method can be used to obtain $\mathbf{x}_n$. Common choices are matching pursuit algorithms [61], which are greedy approaches that select dictionary atoms sequentially, and basis pursuit algorithms [54, 80], which transform (5.8) into a convex problem by replacing the ℓ_0-norm with the ℓ_1-norm. Since the criteria for selecting a proper value of δ are dependent on the approach used to solve (5.8), literature that discusses a given sparse coding algorithm in detail should be referenced for this purpose.

To obtain consistently sparse $\mathbf{x}_n$'s and maximize signal discrimination, $\mathbf{A}$ must be intelligently designed relative to any signal and interference vectors that may be observed. In this chapter, it is assumed that

$$\mathbf{A} = [\mathbf{S}_1 \ \cdots \ \mathbf{S}_P \ \mathbf{H}_1 \ \cdots \ \mathbf{H}_Q] \tag{5.9}$$

where $\mathbf{S}_p$ and $\mathbf{H}_q$ are subdictionaries capable of providing sparse representations of class p signal vectors $\mathbf{s}_n^{(p)}$'s and class q interference vectors $\mathbf{h}_n^{(q)}$'s, respectively. Such subdictionaries may be extracted by applying, e.g., K-SVD [55] or any other sparse dictionary learning algorithm (e.g., [81, 82]) to $\mathbf{y}_n$'s representing the associated signal/interference types. Parametric dictionaries (e.g., wavelets) may also be used by associating certain atoms with the broad categories of signal and interference. Without loss of generality, it is assumed that the first M_s and last M_h columns of $\mathbf{A}$ are associated with the composite signal and interference dictionaries, i.e., $\mathbf{A}_s = [\mathbf{S}_1 \ \cdots \ \mathbf{S}_P] \in \mathbb{R}^{N \times M_s}$ and $\mathbf{A}_h = [\mathbf{H}_1 \ \cdots \ \mathbf{H}_Q] \in \mathbb{R}^{N \times M_h}$, respectively, with $M = M_s + M_h$. Consequently, the sparse coding process incorporates robustness to interference by encoding the majority of signal and interference energy using the first M_s and last M_h coefficients in $\mathbf{x}_n$, respectively. Likelihoods used for signal detection may then be based only on the M_s signal coefficients, while the M_h

interference coefficients are ignored as seen below, i.e., valid separation between these two source types is assumed. As explained in Section 5.3.3, the separation between these components need not be perfect [79], as the learned signal models can account for the fact that some signal energy will be present in $\{x_{M_s+i,n}\}_{i=1}^{M_h}$, while all learned models can account for the fact that some interference energy will be present in $\{x_{i,n}\}_{i=1}^{M_s}$. On the other hand, encoding most of the interference energy in $\{x_{M_s+i,n}\}_{i=1}^{M_h}$ allows for improved discrimination between different signal types and noise by discarding information that is a nuisance to detection and classification.

5.3.2. Quantization of Sparse Coefficients

Just as $\mathbf{x}_n$ is extracted from $\mathbf{y}_n$ to simplify the dependencies between consecutive observation vectors, the *sparse coefficient state vector* $\mathbf{z}_n = [z_{1,n} \ \cdots \ z_{i,n} \ \cdots \ z_{M_s,n}]^T \in \mathbb{R}^{M_s}$ is in turn generated by quantizing the coefficients in $\mathbf{x}_n$ corresponding to signal atoms. Instead of assuming $x_{i,n}$'s obey a convenient but unlikely distribution (e.g., Gaussian [56]), quantization ensures they may be parameterized in a simple but accurate manner using a collection of categorical (i.e., L-level discrete) distributions, while still retaining sufficient information for signal detection and classification. More explicitly, sparse coefficient states are obtained as

$$z_{i,n} = \begin{cases} 0, & |x_{i,n}| \leq \epsilon \\ H\left(x_{i,n}\right), & \text{otherwise} \end{cases}, \ i \in [1, M_s] \tag{5.10}$$

where

$$H(x) = l \text{ if } x \in (t_{l-1}, t_l]\,, \ l \in [1, L-1] \tag{5.11}$$

is a $(L-1)$-level quantization function dependent on the distribution of x under different hypotheses (defined below), and ϵ is a predetermined threshold used to determine those coefficients that are inactive ($x_{i,n} \approx 0$). The purpose of ϵ is to give coefficients close or equal to zero their own state in order to exploit the sparsity of $\mathbf{X}_1^n$ and simplify parameterization, as an overwhelming percentage of

$x_{i,n}$'s will be near zero assuming the matrix $\mathbf{A}$ is appropriately designed. Setting ϵ too low can lead to $\mathbf{z}_n$'s that lack sparsity if $\mathbf{y}_n$'s contain noise and an error tolerant version of (5.8) is used, while setting ϵ too high can lead to discarding important discriminatory features. Practically speaking, a suitable value of ϵ can be that which produces $z_{i,n} = 0$ for some large (SNR dependent) percentage of $x_{i,n}$'s extracted from observations in the training set containing noise alone.

The quantization function $H(x)$ is characterized by transition levels $\mathbf{t} = [t_0, t_1, \dots, t_{L-1}]$, with $-\infty = t_0 < t_1 < \dots < t_{L-1} = \infty$ and $L \geq 2$, and uses reconstruction levels $\mathbf{r} = [1, \dots, L-1]$, though the latter is chosen for simplicity as the actual values used for reconstruction are irrelevant to detection and classification performance [83]. Clearly, smaller L leads to simpler parameterization of the data but a greater loss of discriminatory information. In general, L should be set as large as possible while avoiding an abundance of sample-poor cases when forming categorical distributions (used in Section 5.3.3) representing $z_{i,n}$'s from training data. In other words, since the true probability distributions for $z_{i,n}$'s are rarely if ever available, quantization may be viewed as a necessary step for dealing the realities of limited training data in real-world applications, while refraining from making assumptions about these distributions. Note that, when $L = 2$, no quantizer is used and $z_{i,n} \in \{0, 1\}$.

On the other hand, it is important to ensure that $z_{i,n}$'s contain as much information useful for signal detection and classification as possible for a given L. To this end, the maximum J-divergence quantizer [84] is used that, in the case of multiple hypotheses, specifies $\mathbf{t}$ to maximize the sum of the pairwise divergence between sets of distributions corresponding to $z_{i,n}$'s representing different classes. The importance of J-divergence is largely attributed to results [85, 86] linking a maximum of this measure to minimum error probabilities when discriminating between two hypotheses, i.e., bounds on the latter can be expressed in terms of the former [83]. In general, SCST discriminates between different hypotheses by finding the likelihood of a given *pattern* of sparse coefficient states.

However, the goal of quantization is to use a single function to generate states with marginal distributions that are optimal (in the sense of J-divergence) for this discrimination.

Define $X_i^{(p)}$ and $X_i^{(0)}$ as random variables representing atom coefficients under $\mathcal{H}_1^{(p)}$ and $\mathcal{H}_0$, respectively, with realizations that are sparse coefficients $x_{i,n}$'s. The quantizer function $H(\cdot)$ in (5.11) is characterized by the transition vector $\mathbf{t}$ that maximizes [84]

$$D(\mathbf{t}) = \sum_{i=1}^{M} \sum_{\substack{p,p'=0 \\ p\neq p'}}^{P} d_i^{(p,p')}(\mathbf{t}) \tag{5.12}$$

where

$$d_i^{(p,p')}(\mathbf{t}) = \sum_{l=1}^{L} \left(r_{i,l}^{(p')}(\mathbf{t}) - r_{i,l}^{(p)}(\mathbf{t}) \right) \ln \left(\frac{r_{i,l}^{(p')}(\mathbf{t})}{r_{i,l}^{(p)}(\mathbf{t})} \right)$$

is the J-divergence between two distributions of quantized coefficients belonging to different classes, p and p', and

$$r_{i,l}^{(p)}(\mathbf{t}) = \Pr\left[t_{l-1} < X_i^{(p)} \leq t_l \right] = \int_{t_{l-1}}^{t_l} f_{X_i^{(p)}}(x)dx \tag{5.13}$$

is the probability that $X_i^{(p)}$, with probability density function $f_{X_i^{(p)}}(x)$, lies in the interval $(t_{l-1}, t_l]$. As can be seen, $D(\mathbf{t})$ is the sum of the distances between distributions for each quantized atom coefficient under each pair of hypotheses. The use of separate distributions for each coefficient in (5.12) is unique to this work and is done to exploit the fact that different signals often have sparse representations in terms of different atoms, especially for dictionaries constructed as in (5.9). In other words, if $f_{X^{(p)}}(x)$ represents the distribution of all sparse coefficients for signal type p, then the distance between $f_{X^{(p)}}(x)$ and $f_{X^{(p')}}(x),\ p \neq p'$ may be small, but it can often be assumed that the distance between some $f_{X_i^{(p)}}(x)$ and $f_{X_i^{(p')}}(x),\ p \neq p'$ for a given i is much larger. For instance, say the ith coefficient $x_{i,k}$ is associated with an atom from the subdictionary $\mathbf{S}_p$. Owing to the structure of $\mathbf{A}$ and assumed sparsity of $\mathbf{x}_k$'s, it is more common in this scenario to have $x_{i,k} \neq 0$ when a class p signal is present, but $x_{i,k} = 0$ when a class $p' \neq p$ signal is present, meaning

$f_{X_i^{(p')}}(0) \gg f_{X_i^{(p)}}(0)$. Thus, finding $\mathbf{t}$ to maximize (5.12) allows for generating $z_{i,k}$'s that are optimal for class discrimination for a given i.

5.3.3. Probability of Coefficient State Sequences

This subsection explicitly defines the probability $f_\lambda\left(\mathbf{Z}_{k_1}^n\right)$ of the sparse coefficient state sequence $\mathbf{Z}_{k_1}^n = \{\mathbf{z}_k\}_{k=k_1}^n$ (or $\mathbf{Z}_{k_0}^n$), given the parameter set $\lambda \in \{\lambda_p, \lambda_0\}$, used to form LLRs equivalent to those in (5.3) and (5.5) for hypothesis testing under the SCST framework. In essence, it is shown how the proposed coefficient state representation facilitates realistic formation of these tests even for long and high dimensional data sequences. Each model parameter set defines a BN [78], denoted by $\lambda_p = \{G_p, \Theta_p\}$. Here, G_p is a *directed acyclic graph* [78, 87] with nodes $Z_{i,k-j}^{(p)}$, $i \in [1, M_s]$, that are categorical random variables with time delay $j \in \{0, 1\}$ and corresponding realizations that are sparse coefficient states $z_{i,k-j} \in [0, L-1]$ from (5.10). Edges in G_p describe the dependencies between $Z_{i,k-j}^{(p)}$'s, i.e., the "parents" of each coefficient state. The parameters of the conditional distributions associated with the random variables $Z_{i,k-j}^{(p)}$'s are elements of the set Θ_p, and are described in more detail below. A BN allows for efficiently calculating a complicated joint probability $f_{\lambda_p}\left(\mathbf{Z}_{k_1}^n\right)$ by decomposing it into a product of conditional probabilities of $z_{i,k}$'s given other dependent states, which is much simpler to evaluate in practice. BNs are appropriate for transient detection from multivariate observations as the graph G_p is well-suited for describing causal temporal relationships owing to constraints that certain nodes (random variables) must be processed earlier than others [78].

We first show how to decompose the probability distribution used to form the numerator of the SCST test statistic equivalent to that in (5.3), which can be written as

$$f_{\lambda_p}\left(\mathbf{Z}_{k_1}^n\right) = f_{\lambda_p}\left(\mathbf{z}_{k_1}\right) \prod_{k=k_1+1}^{n} f_{\lambda_p}\left(\mathbf{z}_k | \mathbf{Z}_{k_1}^{k-1}\right) \tag{5.14}$$

where $f_{\lambda_p}\left(\mathbf{z}_{k_1}\right)$ is the prior probability of $\mathbf{z}_{k_1}$ under $\mathcal{H}_1^{(p)}$. Assuming G_p imposes a first order dependency structure, $\mathbf{z}_k$ is only dependent on $\mathbf{z}_{k-1}$, meaning

$$\begin{aligned} f_{\lambda_p}\left(\mathbf{z}_k|\mathbf{Z}_{k_1}^{k-1}\right) &= f_{\lambda_p}\left(\mathbf{z}_k|\mathbf{z}_{k-1}\right) \\ &= f_{\lambda_p}\left(z_{1,k}|\mathbf{z}_{k-1}\right)\prod_{i=2}^{M_s} f_{\lambda_p}\left(z_{i,k}|\{z_{i',k}\}_{i'=1}^{i-1},\mathbf{z}_{k-1}\right) \end{aligned} \tag{5.15}$$

where the second equality is a result of using the chain rule to decompose $f_{\lambda_p}\left(\mathbf{z}_k|\mathbf{z}_{k-1}\right)$ into a product of conditional probabilities of $z_{i,k}$ given $\mathbf{z}_{k-1}$ and previous elements in $\mathbf{z}_k$. The first order assumption in (5.15) is used for simplicity of derivations, and may be dropped if it is invalid for a particular application.

We now exploit the fact that the dependency structure of $z_{i,k}$'s can be simplified according to the BN λ_p being evaluated. More specifically, owing to sparsity of $\mathbf{z}_k$'s, many $z_{i',k-j}$, $i' \neq i$ associated with class $p' \neq p$ signal atoms will be zero when a type p signal is present, meaning the corresponding random variables $Z_{i',k-j}^{(p)}$'s for λ_p carry little to no information about $Z_{i,k}^{(p)}$'s associated with atoms in the same-class subdictionary $\mathbf{S}_p$. While sparsity is the predominant factor that enables a simplified dependency structure, there may be other application-specific attributes that allow for independence of $Z_{i,k}^{(p)}$'s. For instance, for the data in Chapter 2, certain broadband components of plane signatures are typically only present after the onset of specific narrowband mid-frequency components, meaning a node representing the former may be considered conditionally independent of other nodes besides that representing the latter. Regardless, the idea is to measure the dependence between pairs of coefficients during training to determine the edges that connect nodes in a given G_p.

The above justification means (5.15) may be reduced to

$$f_{\lambda_p}\left(\mathbf{z}_k|\mathbf{z}_{k-1}\right) = \prod_{i=1}^{M_s} f_{\boldsymbol{\theta}_{i|\mathcal{S}}^{(p)}}\left(z_{i,k}|\mathcal{S}_i^{(p)}\right) \tag{5.16}$$

where $\mathcal{S}_i^{(p)} = \left\{ z_{i',k-j} : J\left(Z_{i,k}^{(p)}, Z_{i',k-j}^{(p)}\right) > \mu \right\}$, $j \in \{0, 1\}$ with $i' < i$ when $j = 0$ (owing to (5.15)), and $J(Z, Z')$ is a measure of dependence between random variables Z and Z', e.g., mutual information is used for the results in Chapter 6. Therefore, $\boldsymbol{\theta}_{i|\mathcal{S}}^{(p)} \in \Theta_p$ is a length L categorical parameter vector for the conditional probability distribution associated with the ith coefficient state under $\mathcal{H}_1^{(p)}$, given $\mathcal{S}_i^{(p)}$. In other words, $\boldsymbol{\theta}_{i|\mathcal{S}}^{(p)}$ encodes the probability that $Z_{i,k}^{(p)} = l$ for $l \in [0, L-1]$, given that surrounding coefficient states $Z_{i',k-j}^{(p)}$, that $Z_{i,k}^{(p)}$ is found to be dependent on, are equal to specific values $z_{i',k-j} \in \mathcal{S}_i^{(p)}$. Clearly, there is a separate $\boldsymbol{\theta}_{i|\mathcal{S}}^{(p)}$ for each i, p, and possible set $\mathcal{S}_i^{(p)}$. Note that dependence between $Z_{i,k-j}^{(p)}$'s associated with different subdictionaries may exist.

In general, $J(Z, Z')$ can be any measure that best captures the dependence of Z on Z' for specific distributions of each, so long as it is easily calculated from training data to form G_p. In terms of the graph G_p, if $J\left(Z_{i,k}^{(p)}, Z_{i',k-j}^{(p)}\right)$ exceeds a predetermined threshold μ then $Z_{i',k-j}^{(p)}$ is a parent of $Z_{i,k}^{(p)}$ (an edge in G_p connects them) and $z_{i',k-j} \in \mathcal{S}_i^{(p)}$. This concept is illustrated by the example in Fig. 5.3, which shows the dependency structure used to calculate a single term in (5.15), and the simplified dependency structure imposed by the BN for calculating a single term in (5.16). This figure also shows that $\mathcal{S}_i^{(p)} \subseteq \left\{ \{z_{i',k}\}_{i'=1}^{i-1}, \mathbf{z}_{k-1} \right\}$. Selection of the threshold μ can be based on examining the empirical probability density function of $J(Z, Z')$ for training data to look for statistically significant values. Typically, there is a high probability associated with $J(Z, Z') \approx 0$ due to sparsity. Setting μ too high results in ignoring potentially useful discriminatory information. Setting μ too low can lead to large sets $\mathcal{S}_i^{(p)}$, an abundance of conditional distributions, and generally poor sampling of these distributions, thus creating a poor fit of the model λ_p to the training data.

To complete the decomposition of (5.14), the prior probability of observing $\mathbf{z}_{k_1}$ for λ_p is required, and may be written as

$$\begin{aligned} f_{\lambda_p}\left(\mathbf{z}_{k_1}\right) &= f_{\lambda_p}\left(z_{1,k_1}\right) \prod_{i=2}^{M_s} f_{\lambda_p}\left(z_{i,k_1} | \{z_{i',k_1}\}_{i'=1}^{i-1}\right) \\ &= \prod_{i=1}^{M_s} f_{\boldsymbol{\phi}_{i|\mathcal{T}}^{(p)}}\left(z_{i,k_1} | \mathcal{T}_i^{(p)}\right) \end{aligned} \tag{5.17}$$

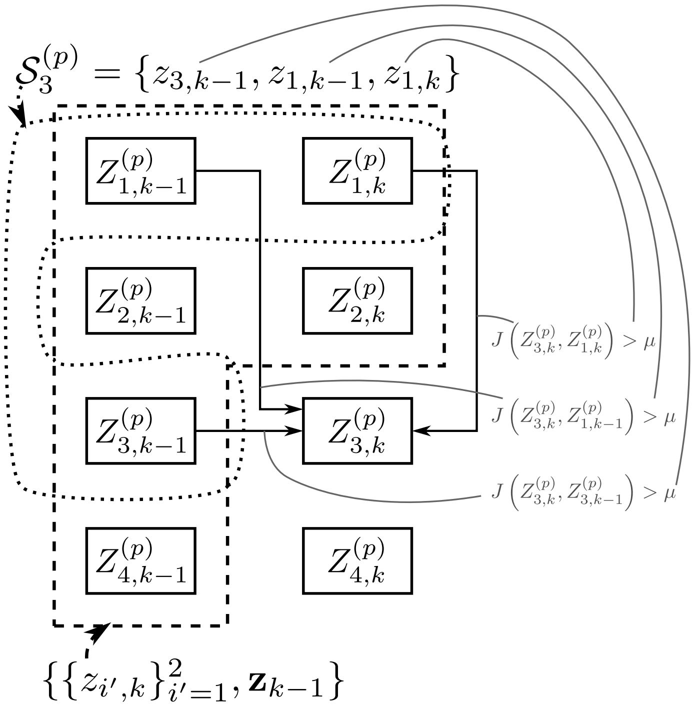

$\mathcal{S}_3^{(p)} = \{z_{3,k-1}, z_{1,k-1}, z_{1,k}\}$
$Z_{1,k-1}^{(p)}$
$Z_{1,k}^{(p)}$
$Z_{2,k-1}^{(p)}$
$Z_{2,k}^{(p)}$
$Z_{3,k-1}^{(p)}$
$Z_{3,k}^{(p)}$
$Z_{4,k-1}^{(p)}$
$Z_{4,k}^{(p)}$
$J\left(Z_{3,k}^{(p)}, Z_{1,k}^{(p)}\right) > \mu$
$J\left(Z_{3,k}^{(p)}, Z_{1,k-1}^{(p)}\right) > \mu$
$J\left(Z_{3,k}^{(p)}, Z_{3,k-1}^{(p)}\right) > \mu$
$\{\{z_{i',k}\}_{i'=1}^{2}, \mathbf{z}_{k-1}\}$

FIGURE 5.3. Example dependency structure imposed on $Z_{3,k}^{(p)}$ by λ_p for $\mathbf{z}_k \in \mathbb{R}^4$. Dashed lines enclose variables $Z_{3,k}^{(p)}$ is dependent on using the decomposition in (5.15). Dotted lines enclose the reduced set of variables $Z_{3,k}^{(p)}$ is dependent on according to the measure $J\left(Z_{3,k}^{(p)}, Z_{i',k-j}^{(p)}\right)$ (e.g., mutual information), thus defining edges in G_p (represented by arrows) and the set $\mathcal{S}_3^{(p)}$.

where $\mathcal{T}_i^{(p)} = \left\{z_{i',k_1} : J\left(Z_{i,k_1}^{(p)}, Z_{i',k_1}^{(p)}\right) > \mu\right\}$, $i' < i$ and $Z_{i,k_1}^{(p)}$ and $Z_{i',k_1}^{(p)}$ are random variables with corresponding realizations z_{i,k_1} and z_{i',k_1}, respectively, that are states of different coefficients in the first observation associated with a class p signal event. As before, $\boldsymbol{\phi}_{i|\mathcal{T}}^{(p)} \in \Theta_p$ is a length L categorical parameter vector for the prior probability distribution associated with the ith coefficient state under $\mathcal{H}_1^{(p)}$, given $\mathcal{T}_i^{(p)}$. The prior probabilities are defined similar to the conditional probabilities in (5.16) except they are conditioned on $\mathcal{T}_i^{(p)}$ rather than $\mathcal{S}_i^{(p)}$, where the former does not contain coefficient states z_{i',k_1-1}'s from the previous vector. This is due to the fact that the first vector in a signal event is independent of previous vectors and, consequently, the interelement dependency structure of $\mathbf{z}_{k_1}$ may be different from that of subsequent vectors in the event. The elements of $\mathcal{T}_i^{(p)}$ are still dictated by G_p, and hence, a given λ_p contains all the necessary components for calculating the probability of observing $\mathbf{Z}_{k_1}^n$ under $\mathcal{H}_1^{(p)}$. The full distribution parameter set associated with λ_p can now be written as

$$\Theta_p = \left\{ \left\{ \boldsymbol{\theta}_{i|\mathcal{S}}^{(p)} \right\}_{i,\mathcal{S}}, \; \left\{ \boldsymbol{\phi}_{i|\mathcal{T}}^{(p)} \right\}_{i,\mathcal{T}} \right\}.$$

We now show how to decompose the denominator of the SCST test statistic equivalent to that in (5.3), i.e., the probability of $\mathbf{Z}_{k_1}^n$ given λ_0. Since interference terms are mostly nullified by the sparse coding process, and since noise is IID, we can write

$$f_{\lambda_0}\left(\mathbf{Z}_{k_1}^n\right) = \prod_{k=k_1}^{n} f_{\lambda_0}\left(\mathbf{z}_k\right). \tag{5.18}$$

Using a similar concept to that in (5.17), each term on the right side of (5.18) can be expressed as

$$\begin{aligned} f_{\lambda_0}\left(\mathbf{z}_k\right) &= f_{\lambda_0}\left(z_{1,k}\right) \prod_{i=2}^{M_s} f_{\lambda_0}\left(z_{i,k} | \{z_{i',k}\}_{i'=1}^{i-1}\right) \\ &= \prod_{i=1}^{M_s} f_{\boldsymbol{\phi}_{i|\mathcal{T}}^{(0)}}\left(z_{i,k} | \mathcal{T}_i^{(0)}\right) \end{aligned} \tag{5.19}$$

where $\mathcal{T}_i^{(0)} = \left\{ z_{i',k} : J\left(Z_{i,k}^{(0)}, Z_{i',k}^{(0)}\right) > \mu \right\}, \; i' < i$ is a set defined similar to $\mathcal{T}_i^{(p)}$, but for $\mathcal{H}_0$ coefficient state sequences. Naturally, $\boldsymbol{\phi}_{i|\mathcal{T}}^{(0)} \in \Theta_0$ is a length L categorical parameter vector defined similar to $\boldsymbol{\phi}_{i|\mathcal{T}}^{(p)}$ and the elements of $\mathcal{T}_i^{(0)}$ are dictated by the edges in G_0.

The required BNs, λ_0 and λ_p's, can be learned [78] using a set of training sequences for each hypothesis. The dependence measure $J(Z, Z')$ between random variables representing the coefficient states is fully observable, meaning G_p has a closed form given a specific set of training data. Each parameter vector in Θ_p can then be found using ML estimation by tabulating the number of times each coefficient is equal to a specific value given the associated set of dependent coefficient states. This training procedure allows for imperfect separation between signal and interference when finding $\mathbf{x}_n$'s since λ_p will model the dependency structure of $\mathbf{Z}_{k_1}^n$ when a class p signal is present, possibly superimposed with multiple types of interference. In other words, even if $\mathbf{Z}_{k_1}^n$

does not fully represent all of the signal components originally present in $\mathbf{Y}_{k_1}^n$, and additionally contains some interference components, training λ_p using such superimposed events accounts for this.

5.3.4. Sequential GLRT Implementation using Sparse Coefficient States

One of the major advantages of the GLRTs in (5.4) and (5.7) is that they can be implemented sequentially for applications where data is continually arriving and intermittently present transient signals must be constantly detected and classified. For binary hypothesis testing problems, this sequential updating is similar to the CUSUM procedure (see Section 3.2.2), also known as Page's test [34], which was originally formulated for the case of independent observations. This "CUSUM-like" procedure [35] (so called because dependent observations are assumed) used for sequential implementation of the GLRTs produces the same results as using the standard GLRTs in (5.4) and (5.7), when given the same data and parameters. This subsection presents the proposed SCST implementation of the GLRTs when using BNs and a coefficient state sequence $\mathbf{Z}_{k_1}^n$, rather than general model parameter sets and the original data sequence $\mathbf{Y}_{k_1}^n$, as in Section 5.2.2.

To implement a sequential version of the signal detection phase of the SCST method, calculating the GLRTs in (5.3) is replaced by calculating test statistics given by

$$B_p(n) = \max\{0,\ B_p(n-1) + b_p(n)\},\ n = \hat{k}_0, \hat{k}_0 + 1, \ldots \tag{5.20}$$

and initialized as $B_p(\hat{k}_0 - 1) = 0,\ \forall p$. This statistic is updated at time n using the nonlinearity

$$b_p(n) = \begin{cases} \ln\left(\dfrac{f_{\lambda_p}(\mathbf{z}_n|\mathbf{z}_{n-1})}{f_{\lambda_0}(\mathbf{z}_n)}\right), & B_p(n-1) > 0 \\[2ex] \ln\left(\dfrac{f_{\lambda_p}(\mathbf{z}_n)}{f_{\lambda_0}(\mathbf{z}_n)}\right), & B_p(n-1) = 0 \end{cases}$$

where the same assumptions used in the previous subsection apply here, e.g., a first order dependency structure for signals. Alternatively, $b_p(n)$ can be rewritten as

$$b_p(n) = \begin{cases} \displaystyle\sum_{i=1}^{M_s} \ln\left(\frac{f_{\boldsymbol{\theta}_{i|\mathcal{S}}^{(p)}}\left(z_{i,n}|\mathcal{S}_i^{(p)}\right)}{f_{\boldsymbol{\phi}_{i|\mathcal{T}}^{(0)}}\left(z_{i,n}|\mathcal{T}_i^{(0)}\right)} \right), & B_p(n-1) > 0 \\ \displaystyle\sum_{i=1}^{M_s} \ln\left(\frac{f_{\boldsymbol{\phi}_{i|\mathcal{T}}^{(p)}}\left(z_{i,n}|\mathcal{T}_i^{(p)}\right)}{f_{\boldsymbol{\phi}_{i|\mathcal{T}}^{(0)}}\left(z_{i,n}|\mathcal{T}_i^{(0)}\right)} \right), & B_p(n-1) = 0 \end{cases}.$$

Owing to the decomposition presented in (5.14)–(5.19), $B_p(n)$ accumulates at the same rate as the LLR in (5.3) (when given the same data), though it resets whenever $B_p(n) \leq 0$. Therefore, while in Section 5.2.2 a signal is detected whenever any time segment of (5.3) increases by η, here a signal is detected at time n whenever

$$\max_p B_p(n) \geq \eta \tag{5.21}$$

i.e., the cumulative value of $B_p(n)$ simply must exceed the threshold η. More details on the equivalence of using the test statistics in (5.3) and (5.20) can be found in Section 3.2.2.

Naturally, to implement a sequential version of the quiescent detection phase of the SCST method, calculating the GLRTs in (5.7) is replaced by calculating test statistics

$$T_p(n) = \max\{0,\ T_p(n-1) + t_p(n)\},\ n = \hat{k}_1, \hat{k}_1 + 1, \ldots \tag{5.22}$$

that are initialized as $T_p(\hat{k}_1 - 1) = 0,\ \forall p$, and updated using the nonlinearity

$$t_p(n) = \ln\left(\frac{f_{\lambda_0}(\mathbf{z}_n)}{f_{\lambda_p}(\mathbf{z}_n|\mathbf{z}_{n-1})}\right) = \sum_{i=1}^{M_s} \ln\left(\frac{f_{\boldsymbol{\phi}_{i|\mathcal{T}}^{(0)}}\left(z_{i,n}|\mathcal{T}_i^{(0)}\right)}{f_{\boldsymbol{\theta}_{i|\mathcal{S}}^{(p)}}\left(z_{i,n}|\mathcal{S}_i^{(p)}\right)} \right).$$

Unlike $b_p(n)$, the value of $t_p(n)$ does not depend on the value of the corresponding test statistic at time $n-1$ since conditional distributions are always used under $\mathcal{H}_1^{(p)}$ in the quiescent detection phase, as shown in (5.5). As before, $T_p(n)$ accumulates at the same rate as the LLR in (5.5), meaning the absence of any signal is declared at time n whenever

$$T_{p^*}(n) \geq \gamma \tag{5.23}$$

where

$$p^* = \arg\max_p \, B_p(n).$$

As in Section 5.2.2, when $\mathcal{H}_0$ is again accepted a class label is assigned based on the ML signal type at time $\hat{k}_0 - 1$ and the process reverts back to looking for a new signal of unknown type according to (5.1). This phase switching process can continue indefinitely or until the end of the observation sequence has been reached, if applicable.

5.4. Conclusions

This chapter introduced a new method for detection and classification of transient acoustical events from multivariate observations using the patterns of corresponding coefficient state sequences to determine the likelihood of each known signal model. The motivation behind this approach stems from the fact that coefficient state sequences provide a simple way to represent nonstationary components and facilitate realistic calculation of likelihoods, even for lengthy vector sequences. This is especially important for applications where acoustical events associated with a given signal type are very erratic and have complex temporal evolutions. Furthermore, few assumptions need to be made concerning the statistics of observation components compared to, e.g., the SRCT method in Chapter 4. The proposed method inherently provides robustness to multiple competing interference sources, owing to the separation capabilities of sparse coding when using an appropriately designed dictionary. Use of this sparse coding also means that the SCST method can provide estimates of the signal and interference components in each observation without any additional overhead.

Like the SRCT method, the SCST method satisfies all of the requirements for solutions to the soundscape characterization problem mentioned in Section 1.3. The downsides of the SCST method relative to the SRCT method are its increased computational complexity (see Section 6.7 for details) and inability to provide class labels for interference sources. The SCST method is the last of two main comprehensive solutions for the problem considered in this thesis. Therefore, the next chapter presents an extensive benchmarking of the proposed methods with a Gaussian mixture model approach. This will allow for a detailed discussion on the advantages and disadvantages of each method for performing detection, classification, and estimation of transient acoustical sources.

CHAPTER 6

Performance Evaluation

6.1. Introduction

As mentioned in Chapter 1, the development of comprehensive approaches for detection and classification of multiple time-varying transient sources using multivariate data was strongly motivated by a national park soundscape characterization application. This necessitated the inclusion of many intrinsic processing capabilities (see Section 1.3), such as handling the simultaneous presence of competing interference and exploiting the unique structure of sequential multivariate data. It is therefore essential to determine the effectiveness of the developed methods for performing source characterization tasks using the soundscape data in Chapter 2, namely detection, classification, and estimation of signal events. More specifically, this chapter presents the results of applying the sequential random coefficient tracking (SRCT) and sparse coefficient state tracking (SCST) methods, introduced in Chapters 4 and 5, respectively, to the Kenai Fjords site four (KEFJ004) and Great Sand Dunes site one (GRSA001) data sets. Exceptional performance on these data sets would not only indicate that the proposed methods are suitable for deployment on monitoring stations for real-time soundscape characterization, but also that they may be appropriate for use in other applications with similar challenges, e.g., medical diagnosis using magnetic resonance imaging [17].

This chapter is organized as follows. A brief description of the Gaussian Mixture Model (GMM)-based method [13, 15, 45], that is used to benchmark the performance of the proposed SRCT and SCST methods, is first presented in Section 6.2. The test setup that is used to generate experimental results, i.e., the parameters and procedures used, is described in Section 6.3. Section 6.4 then presents an evaluation of the test statistics used by each method in order to determine their power for detection of transient acoustical events. Section 6.5 provides the results of applying each method to segments of 1/3 octave data from each data set to demonstrate their effectiveness in a practical sense, i.e., their ability to generate an accurate summary of a soundscape in terms of the frequency of occurrence of each signal type. Section 6.6 evaluates the ability of the SRCT and SCST methods to provide separate estimates of the signal and interference components of a given data segment. In Section 6.7, a brief analysis of the computational complexity of each method is provided. Finally, concluding remarks are given in Section 6.8.

6.2. Baseline Method: Gaussian Mixture Models

A GMM-based source characterization scheme is adopted as the benchmark method in this thesis since these models are commonly used in applications involving environmental sound [15], wildlife call [13], and speech [45] recognition. Similar to the SRCT method, the ability to calculate the likelihood of a GMM given an observation vector allows them to be used in a hierarchy of tests to determine the composition of an observation in terms of signal, interference, and noise. This allows for a benchmarking of the SRCT test statistics that would otherwise be difficult, since very few methods use such tests. For this reason, this section uses many of the same concepts, notation, and assumptions introduced in Chapter 4. In particular, under the GMM-based framework, it is assumed that a 1/3 octave vector at time k can generally be represented as

$$\mathbf{y}_k = \alpha_k \mathbf{s}_k + \beta_k \mathbf{h}_k + \mathbf{w}_k$$

where $\mathbf{s}_k \in \mathbb{R}^N$ and $\mathbf{h}_k \in \mathbb{R}^N$ are random vectors defined similar to $\mathbf{y}_k \in \mathbb{R}^N$, but represent the signatures produced by the unknown signal and interference sources to be characterized, respectively; α_k and β_k are binary variables that indicate the presence $(\alpha_k, \beta_k = 1)$ or absence $(\alpha_k, \beta_k = 0)$ of a signal and interference source, respectively; and $\mathbf{w}_k \overset{\text{IID}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{R}_\mathbf{w})$ is an independent and identically distributed (IID) measurement noise vector, where $\mathbf{R}_\mathbf{w} \in \mathbb{R}^{N \times N}$ is a known full-rank noise covariance matrix.

It is the ability of GMMs to model arbitrary distributions with multiple modes that makes them suitable for recognizing features extracted from inconsistent signatures that are a mainstay of acoustical source characterization applications. A GMM is a weighted sum of K Gaussian densities, and is specified by $\lambda_{\mathbf{s}}^{(p)} = \left\{ w_{\mathbf{s},i}^{(p)}, \boldsymbol{\mu}_{\mathbf{s},i}^{(p)}, \boldsymbol{\Sigma}_{\mathbf{s},i}^{(p)} \right\}_{i=1}^{K}$ for signal type p, where $w_{\mathbf{s},i}^{(p)}$, $\boldsymbol{\mu}_{\mathbf{s},i}^{(p)}$, and $\boldsymbol{\Sigma}_{\mathbf{s},i}^{(p)}$ are the ith component weight, mean vector, and covariance matrix, respectively. Along the same lines as the SRCT method, interference and dual source GMMs many be similarly defined as $\lambda_{\mathbf{h}}^{(q)}$ and $\lambda_{\mathbf{s},\mathbf{h}}^{(p,q)}$, respectively, where q is the interference type. The likelihood of a given set of signal GMM parameters, given a vector observation $\mathbf{y}_k$, may be found as [15]

$$\ell\left(\lambda_{\mathbf{s}}^{(p)}; \mathbf{y}_k\right) = \sum_{i=1}^{K} w_{\mathbf{s},i}^{(p)} f\left(\mathbf{y}_k; \boldsymbol{\mu}_{\mathbf{s},i}^{(p)}, \boldsymbol{\Sigma}_{\mathbf{s},i}^{(p)}\right) \tag{6.1}$$

where $f(\cdot)$ represents a multivariate Gaussian density function that is parameterized by $\boldsymbol{\mu}_{\mathbf{s},i}^{(p)}$ and $\boldsymbol{\Sigma}_{\mathbf{s},i}^{(p)}$.

In order to detect the simultaneous presence of signal and interference sources, a separate GMM is needed for modeling $\mathbf{y}_k$'s that correspond to all of the different source composition hypotheses (see (4.3) in Section 4.3), besides the noise alone hypothesis $\mathcal{H}_0$. The likelihood under $\mathcal{H}_0$ may be found using a standard unimodal multivariate Gaussian likelihood function for parameters that are the fixed and known noise statistics mentioned above. In each case, a mixture of Gaussian densities can represent observations that contain both a random noise component, as well as the component(s) associated with one or two random sources. A hierarchical scheme, similar to that implemented by the SRCT method, may then be applied to each $\mathbf{y}_k$ using likelihoods calculated as in (6.1) for a set of log-likelihood ratio tests (LLRT). While the dominant source and source quantity LLRTs for the GMM-based method have the same forms as those used for the SRCT method, i.e., (4.25) and (4.26), respectively, detection and classification are performed using single source GMMs. More specifically, GMM-based detection is performed using the test statistic

$$\max_{\lambda} \frac{\ell\left(\lambda; \mathbf{y}_k\right)}{\ell\left(\lambda_0; \mathbf{y}_k\right)}, \ \lambda \in \{\lambda_{\mathbf{s}}^{(p)}\}_{p=1}^{P} \cup \{\lambda_{\mathbf{h}}^{(q)}\}_{q=1}^{Q} \tag{6.2}$$

where λ_0 is the parameter set under $\mathcal{H}_0$, and P and Q are the total number of signal and interference sources, respectively. Classification is then performed by determining the most likely source of each type as

$$p^* = \arg\max_{p} \frac{\ell\left(\lambda_{\mathbf{s}}^{(p)}; \mathbf{y}_k\right)}{\ell\left(\lambda_0; \mathbf{y}_k\right)} \qquad q^* = \arg\max_{q} \frac{\ell\left(\lambda_{\mathbf{h}}^{(q)}; \mathbf{y}_k\right)}{\ell\left(\lambda_0; \mathbf{y}_k\right)}.$$

The reason for this discrepancy between the SRCT and GMM-based methods is that, in the former, the dual source models and likelihood values can accommodate the presence of a single source, as coefficient estimates for absent sources will be negligible (see Section 4.3.3) whereas a dual source GMM represents observations that necessarily contain both signal and interference components.

6.3. Test Setup

This section describes the experimental setup used to evaluate the performance of the SRCT, SCST, and GMM methods for characterizing transient acoustical sources. Although the procedures used to apply a given method are largely the same for both the KEFJ004 and GRSA001 data sets, they were analyzed independently, meaning the models (e.g., GMMs) and the parameters used (e.g., detection thresholds) were unique to each data set. Unless otherwise noted, the models and parameters discussed in this section were used to generate all the results presented in this thesis.

6.3.1. Kenai Fjords Site 4

To apply the SRCT, SCST, and GMM methods to the KEFJ004 data set (see Section 2.3.1), disjoint training and testing sets were formed using a collection of two-hour-long data segments found throughout the 19 days of data. In order to provide robust training and a challenging testing environment, segments were selected for both sets that contained a relatively large number of signal and interferences sources and contained events with highly variable signatures. In particular, among all the observations used in this study (training and testing), 36.7% contained noise alone, 4.62% contained signal plus noise, 53.0% contained interference plus noise, and 5.73% contained signal, interference, and noise. Observations containing only interference sources were more common due to rather relentless rain and thunder at the KEFJ004 site, which is also the reason interference was present for the majority of observations where a signal was present. The training set consisted of 10 data segments (about 4.39% of total data) and was used to extract models and determine proper parameter values for a given method, as described in more detail below. The testing set consisted of 40 (about 17.5% of total data) segments and was used to evaluate the performance of each method. No testing segments contained an event used to extract source models, and at least one segment from each of the 19 available days of data was used.

In order to apply the SRCT method to the KEFJ004 data, the subspace and autoregressive (AR) parameter matrix dictionaries for each source type, as well as the observation noise statistics, need to be extracted from a set of "clean" training events that contain only the signatures of interest for a particular model. These training events were chosen to ensure a given source model adequately captures the diversity of the associated signatures. For example, in the KEFJ004 site, a significant amount of within-class diversity exists for propeller plane events, not only because this is the most commonly occurring type of signal, but also owing to the presence of different types of planes. Consequently, more plane training events are required when compared to other signal types. To extract a source's subspace in (4.1), eigenvalue decomposition [88] was applied to a covariance matrix representing the data from several clean events corresponding to that source type. The three eigenvectors corresponding to the largest eigenvalues were then used as basis vectors for that source model. Using the parameter identification procedure in [89], second order linear vector AR model parameters (used in (4.2)) were also extracted from the aforementioned set of training events. Noise statistics were estimated using observation segments that did not contain any signal or interference components. The detection, dominant source, and source quantity thresholds were found to be $\gamma = -1.07$, $\tau = 3.63$, and $\eta = 54.6$, respectively, for the SRCT method. For all methods, thresholds were selected such that no signals in the training segments were missed.

For the GMM-based method, single source models were formed for each signal and interference type by applying the expectation maximization (EM) algorithm [90] to the same training events used to form the SRCT models. Dual source GMMs were extracted from observations formed by superimposing different combinations of the clean signal and interference observations used to form the single source GMMs. The number of components in a given GMM was selected by applying the silhouette method [91] to the corresponding set of observations the GMM was extracted from, and was generally between two and five. The detection, dominant source, and source quantity thresholds were found to be $\gamma = -4.55$, $\tau = 19.0$, and $\eta = -22.7$, respectively, for the GMM method.

For the SCST method, the KEFJ004 training data was used to form the sparse coding dictionary $\mathbf{A}$, extract the Bayesian networks (BN) λ_p's and λ_0 [78] (see Section 5.3.3), and choose the detection thresholds $\eta = 52.3$ and $\gamma = 20.0$ in (5.21) and (5.23), respectively. In particular, events within the training segments containing the signatures of one signal source, often superimposed with one or more types of interference, were used to learn an associated BN λ_p. Similarly, training events containing only interference and noise were used to learn the BN λ_0. To form $\mathbf{A}$, K-SVD [55] was applied separately to different sets of observations, each representing a single type of signal or interference, to extract source-specific dictionaries $\mathbf{S}_p \in \mathbb{R}^{N \times 25}$, $\forall p$ and $\mathbf{H}_q \in \mathbb{R}^{N \times 15}$, $\forall q$. The concatenation of these atoms yields $\mathbf{A}$ as in (5.9). Selection of the number of atoms in each dictionary was performed according to the guidelines outlined in [55]. Fewer atoms were extracted for each interference source since the associated signatures are generally less diverse than those of signals. Basis pursuit denoising [80] was used to perform sparse coding in the dictionary learning stage (K-SVD has the same flexibility as SCST) as well as to extract each $\mathbf{x}_n$ in (5.8). Based on the criterion that $z_{i,n} = 0$ for 99% of sparse coefficients representing observations in the training set containing noise alone, $\epsilon = 3.24$ was selected. To determine parent-child relationships in the BNs λ_p's, mutual information was used as the dependence measure with a threshold of $\mu = 0.1$, which corresponds to the largest 2% of values witnessed for this measure.

As mentioned before, to maximize class discrimination, we desire L in (5.11) to be as large as possible while avoiding sample poor distributions used to form (5.16), (5.17), and (5.19). Realistically, even with an abundance of training data, there may be no available samples to form some of the conditional distributions, e.g., in cases where a specific combination of dependent coefficient values $\mathcal{S}_i^{(p)}$ never occurs for class p signals due to the structure of their signatures. Therefore, the criterion used in these experiments is that at least one of the conditional distributions for each $Z_{i,k}^{(p)}$ (every coefficient and class combination) must be formed using $\geq 4L$ samples. When this criterion is met, the remaining distributions that are considered sample poor are set to uniform. This procedure led to $L = 4$.

Since the data vectors used for this study represent frequency subband acoustical energy at different times they are not zero-mean, and hence, the noise mean was subtracted from each observation as it arrived, before being processed by a given method. Also, the hidden Markov model (HMM)-based sequential decision fusion scheme in Section 3.3 was applied to the signal detection/classification results produced by the SRCT and GMM methods, since they both assign labels separately to individual observations, as opposed to the SCST method, which inherently makes event-wide decisions. As a reminder, this fusion process finds the likelihood that a certain signal type is present, given a sequence of preliminary decisions (class labels), using a procedure that is similar to that proposed in [35]. A separate three-state HMM was learned for each signal type using the Baum-Welch algorithm. Different HMMs were used for the SRCT and GMM methods, and each HMM was formed using preliminary decision sequences generated by a given method for segments in the training set. This approach is appropriate for on-line processing and is done both to smooth out the detection/classification results, and to associate a cluster of detections that have the same label with a single event for more concise and meaningful classification results. In the case of the GMM method, this HMM-based smoothing provides a means to incorporate temporal dependencies between decisions that would otherwise be independent.

6.3.2. Great Sand Dunes Site 1

As mentioned, the same general experimental setup described above was used to apply each method to the GRSA001 data (see Section 2.3.2), and hence, only the differences between the two experimental setups (e.g., parameter values used) will be described here. It is important to note that, unlike the KEFJ004 data, no annotations exist for the GRSA001 data that denote the presence/absence of interference in each observation. This not only means that the distribution of the GRSA001 data in terms of observation composition hypotheses is unknown, but also that this data set cannot be used to directly evaluate the detection, dominant source, and source quantity test statistics used by the SRCT and GMM methods. However, as will be seen, both data sets can indeed be used to evaluate the performance of each method for detecting signal components.

The GRSA001 training and testing sets consisted of 10 data segments (about 4.90% of total data) and 38 segments (about 18.6% of total data), respectively, and served the same purposes as in the KEFJ004 data. At least two segments from each of the 17 available days of data were used to form the testing set. When applying the SRCT method to the GRSA001 data, each source was characterized by a two-dimensional subspace and a second order vector AR model, that were extracted as described above. The SRCT detection, dominant source, and source quantity thresholds were found to be $\gamma = 98.0$, $\tau = 78.8$, and $\eta = 276$, respectively. For the GMM method, these thresholds were found to be $\gamma = 101$, $\tau = 10.2$, and $\eta = -20.5$, respectively. For the SCST method, the signal and quiescent detection thresholds were set to $\eta = 43.7$ and $\gamma = 20.0$, respectively. As before, the sparse coding dictionary $\mathbf{A}$ was formed by concatenating the source specific dictionaries $\mathbf{S}_p \in \mathbb{R}^{N\times 25}$, $\forall p$ and $\mathbf{H}_q \in \mathbb{R}^{N\times 15}$, $\forall q$, that were all extracted using K-SVD.

The threshold for determining the zero state was set to $\epsilon = 8.61$, $L = 4$ quantization levels were used to form coefficient states, and $\mu = 0.1$ was used to determine the edges in the BNs. HMM-based decision fusion was also used to aggregate and smooth the signal detection/classification results produced by the SRCT and GMM methods for the GRSA001 data.

6.4. Test Statistic Evaluation

This section presents a performance evaluation of the test statistics implemented by each method for detecting sources in the testing segments associated with a given data set. Results are presented in terms of receiver operator characteristic (ROC) curves, which show how the probability of detection (P_D) and probability of false alarm (P_{FA}) change as the associated decision threshold is modified. Unlike the results presented in the next section, here performance is measured using statistics generated for individual observations rather than entire events, i.e., the temporal position of an observation is irrelevant and only its associated detection statistics are considered. For this reason, the HMM-based decision fusion does not influence the results presented in this section, as this processing step does not modify any of the test statistics.

Since the SRCT method uses three tests to determine the composition of an observation in terms of signal, interference, and noise, the performance of each of these tests is compared with that of the equivalent tests implemented by the GMM method. In this case, P_D indicates the probability of correctly accepting the hypothesis associated with the numerator of a given LLRT, i.e., $\mathcal{H}_3^{(p,q)}$, $\mathcal{H}_1^{(p^*)}$, and $\mathcal{H}_3^{(p^*,q^*)}$ for (4.23), (4.25), and (4.26), respectively. On the other hand, P_{FA} indicates the probability of falsely accepting the hypothesis associated with the denominator of a given LLRT, i.e., $\mathcal{H}_0$, $\mathcal{H}_2^{(q^*)}$, and $\mathcal{H}_1^{(p^*)}$ *or* $\mathcal{H}_2^{(q^*)}$ for (4.23), (4.25), and (4.26), respectively.

The SCST method does not use test statistics that are equivalent to those used by the SRCT and GMM methods, and hence, a separate evaluation metric is needed to compare the performance of all three methods. In particular, signal detection performance is used, where P_D refers to the probability of correctly detecting the presence of a signal irrespective of any interference that may be present, while P_{FA} refers to the probability of falsely detecting a signal when none is present. Though SCST signal detection is dependent on both the signal and quiescent detection statistics, only a single threshold may be modified to generate the ROC, and hence, the threshold for the latter remained fixed at the value indicated in the previous section for a given data set. Similarly, the SRCT and GMM signal detection ROCs were generated by only modifying the statistic for the initial detection test, i.e., (4.23) and (6.2), respectively, to determine the associated impact on signal detection performance, while the thresholds for the other tests in the hierarchy remained fixed.

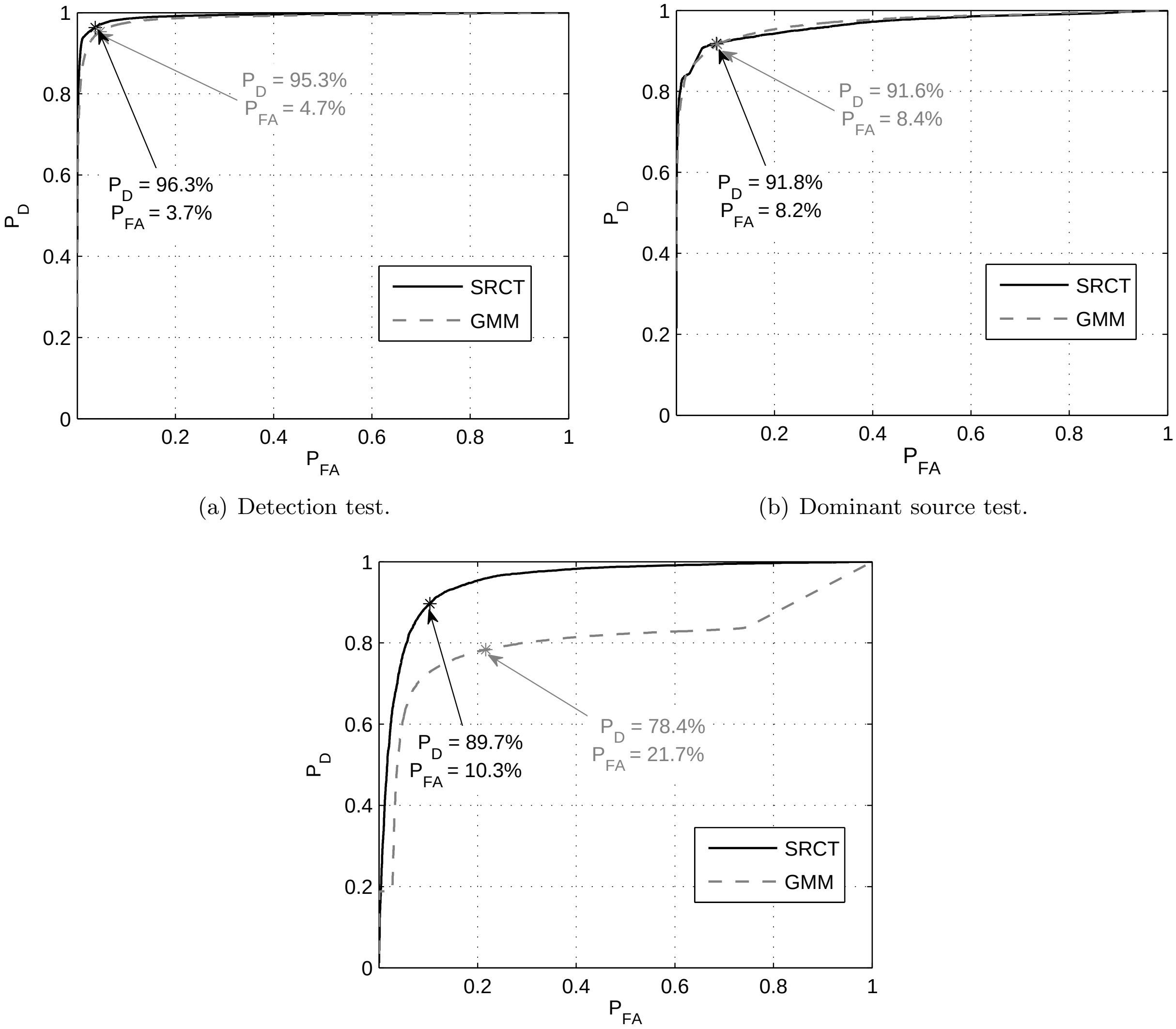
In all cases, the evaluation metrics considered in this section are the area under the ROC curve (AUC) and the P_D and P_{FA} at its "knee-point". The AUC is important since it represents the discrimination ability of a test, while the knee-point corresponds to a decision threshold where $P_D + P_{FA} = 1$. It is important to note a slight discrepancy in terminology between the types of evaluations performed in this section, that is rooted in the way the SRCT and SCST methods define their associated hypotheses. When testing the SRCT statistics directly, $\mathcal{H}_1^{(p^*)}$ strictly means that only a type p signal and noise are present (no interference), while for the signal detection evaluation where all three methods are compared, $\mathcal{H}_1^{(p^*)}$ means that a type p signal and noise are present, but interference may or may not be present. See (4.3) and (5.1) for precise definitions of the hypotheses used by the SRCT and SCST methods, respectively.

6.4.1. Kenai Fjords Site 4

The ROC curves comparing the performance of the three SRCT and GMM test statistics are shown in Figs. 6.1(a)-6.1(c). As can be seen from Fig. 6.1(a), the knee-points of the ROC curves for the SRCT and GMM detectors (signal and/or interference) are ($P_D = 96.3\%$, $P_{FA} = 3.7\%$) and ($P_D = 95.3\%$, $P_{FA} = 4.7\%$), respectively, while the AUCs are 0.992 and 0.987, respectively. Although both detectors operate similarly, the SRCT detector performs slightly better since the source estimates used to form the likelihoods can accommodate arbitrary combinations of signal and interference. Nonetheless, the single source GMMs still produce sufficiently high likelihoods for observations containing one or more sources. In both cases missed detections are primarily caused by samples associated with mild interference, such as very light rain, as well as the presence of novel source signatures. The latter is due to the inability of basis vectors for the SRCT method, or distributions for the GMM method, to capture source signatures that rarely occur during training. False alarms are mostly caused by ambient noise statistics changing slightly over time, owing to the properties of natural soundscape data. However, this deviation is small enough such that the detectors remain fairly robust to this temporal change.

The ROC curves in Fig. 6.1(b) for the SRCT and GMM dominant source tests exhibit knee-points at ($P_D = 91.8\%$, $P_{FA} = 8.2\%$) and ($P_D = 91.6\%$, $P_{FA} = 8.4\%$), respectively, while the AUCs are 0.965 and 0.968, respectively. The GMM and SRCT methods perform similarly in the dominant source stage, as the models used by each are appropriate for classification in these single source scenarios. This result reinforces the suitability of GMM-based methods for the interference-free acoustical source recognition tasks that they are typically applied to, owing to the ability of GMMs to represent complicated signatures. For both methods, dominant source errors can mostly be attributed to closeness between some realizations of rain and thunder interference to the models associated with helicopter and jet signal types, which occasionally leads to $\mathcal{H}_1^{(p^*)}$ being accepted when $\mathcal{H}_2^{(q^*)}$ is true. This confusion is typically only the case when unusually strong thunder is encountered.

Fig. 6.1(c) shows the ROC curves for the SRCT and GMM source quantity test statistics, where the knee-points are ($P_D = 89.7\%$, $P_{FA} = 10.3\%$) and ($P_D = 78.3\%$, $P_{FA} = 21.7\%$), respectively, and the AUCs are 0.953 and 0.806, respectively. The GMM source quantity ROC curve is smoothest from $P_D = 19.0\%$ to $P_D = 84.0\%$ since this is the region where the dual source GMMs produce numerically significant likelihoods, and hence, the LLRT is not affected by round off error. These results show that the main difference between the SRCT and GMM-based methods is their



(a) Detection test.

(b) Dominant source test.

(c) Source quantity test.

Figure 6.1. ROC curves for each type of test statistic produced by the SRCT and GMM methods when applied to the KEFJ004 data.

respective performance in determining source quantity, where the SRCT vastly outperforms the GMM-based method. This is due in part to the SRCT exploiting the dependency structure of each source's signatures when forming the LLRT. More importantly, the dual source GMMs have difficulty representing the extreme variations associated with superimposed signatures from two different sources. For instance, not only is there a significant amount of diversity between events associated with a given source type, but both signal and interference sources present in an observation can be at different relative stages in their evolution (i.e., temporal position within an event), and can have arbitrary amplitudes. A given dual source GMM must capture all combinations of these variations, leading to a lower likelihood for the GMM given any specific observation, whereas the SRCT method directly estimates the source signatures to account for this diversity. For the SRCT method, in some cases a weak signal superimposed with strong interference triggers false acceptance of $\mathcal{H}_2^{(q^*)}$ over $\mathcal{H}_3^{(p^*,q^*)}$, thus missing the signal. Other errors for this test are again due to novel source signatures and their closeness to subspaces of other models. Specifically, strong plane and helicopter signals occasionally falsely trigger $\mathcal{H}_3^{(p^*,q^*)}$ over $\mathcal{H}_1^{(p^*)}$, and very strong thunder also causes false acceptance of $\mathcal{H}_3^{(p^*,q^*)}$. Clearly, performance issues arising from close subspaces are more prevalent when the system must contend with more source types, and these issues are exacerbated when the subspace and observation dimensionalities are closer.

To compare the performance of all three methods under consideration, Fig. 6.2 shows their associated signal detection ROC curves. As can be seen, the knee-points of the ROC curves for the SCST, SRCT, and GMM methods are ($P_D = 94.0\%$, $P_{FA} = 5.99\%$), ($P_D = 86.7\%\%$, $P_{FA} = 13.3\%\%$), and ($P_D = 84.2\%$, $P_{FA} = 15.8\%$), respectively, while their AUCs are 0.967, 0.863, and 0.889, respectively. The relative difference in performance between the SRCT and GMM methods for signal detection can be mostly attributed to the same factors described above. On the other hand, the SCST achieves a much higher P_D at a given P_{FA} primarily because it exploits the dependencies between the signal components in temporally adjacent observations to yield a cumulative test statistic. Consequently, even when an event contains some observations with weak signal components the SCST method maintains a sufficiently high detection statistic throughout such an event. In contrast, the SRCT and GMM methods perform detection on each observation independently, leading to more missed detections within some events, though the former performs better overall at its knee-point since it forms the detection statistic for a given observation using previous observations. Missed detections for the SCST method are primarily due to a delay in signal detection that is inherent with transient detection schemes using a likelihood ratio [34], which causes

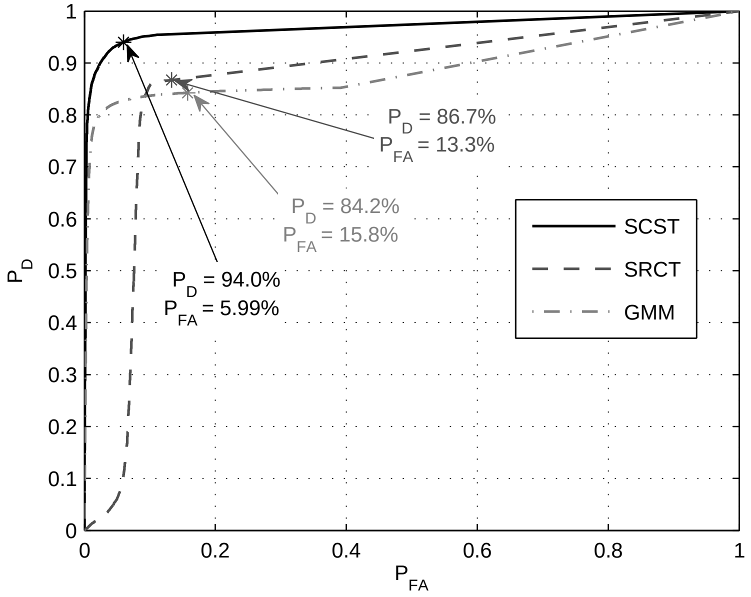


FIGURE 6.2. Signal detection ROC curves for the KEFJ004 data.

a small number of samples to be missed at the beginning of each signal event. Similarly, false alarms generated by the SCST method are mainly caused by delays to quiescent detection, leading to a few false detections at the end of each signal event.

6.4.2. Great Sand Dunes Site 1

Recall that the GRSA001 data cannot be used to directly evaluate the SRCT and GMM test statistics since the presence of interference was not tabulated for this data set. Therefore, only the signal detection performance of each method is considered here. As can be seen from Fig. 6.3, the knee-points of the GRSA001 signal detection ROC curves for the SCST, SRCT, and GMM methods are ($P_D = 92.8\%$, $P_{FA} = 7.21\%$), ($P_D = 84.4\%$, $P_{FA} = 15.6\%$), and ($P_D = 79.6\%$, $P_{FA} = 20.4\%$), respectively, while their AUCs are 0.962, 0.863, and 0.845, respectively. Many of the same reasons the SCST method performed signal detection most accurately on the KEFJ004 data set also apply here. However, for the GRSA001 data set, perhaps the biggest reason the SCST method produced so few false alarms relative to the other methods is its ability to remain robust to the simultaneous presence of multiple types of interference. In contrast, both the SRCT and GMM methods assume

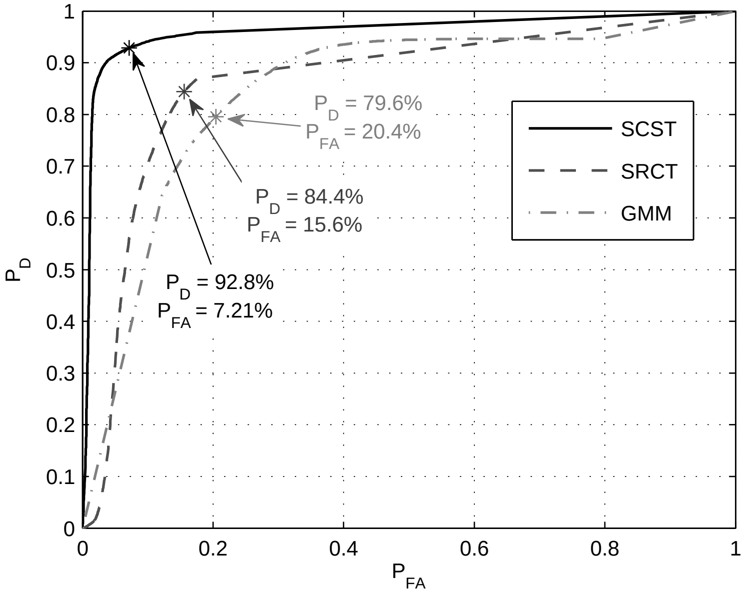


Figure 6.3. Signal detection ROC curves for the GRSA001 data.

a maximum of one type of signal and one type of interference can be simultaneously present. Due to the wide variety of interference source types associated with the GRSA001 soundscape, the presence of multiple types of interference in a given observation would sometimes lead to false alarms for methods other than SCST, since the superimposed signal and interference model would produce a higher likelihood than the interference alone model. A larger performance gap between the SRCT and GMM methods is also witnessed for the GRSA001 data when compared to the KEFJ004 data, which can be primarily attributed to the extremely prominent interference present in the former data set, and better handling of this interference by the SRCT method. More specifically, variable and often violent wind led to a higher P_D at a given P_{FA} for the SRCT method, since the dual source GMMs were often insufficient for identifying the presence of weak jet signatures occluded by such heavy wind.

6.5. Classification Using Entire Data Segments

While the above ROC analyses demonstrated the effectiveness of various test statistics implemented by each method when applied to individual observations in a given data set, here the overall performance of each method for correctly detecting and classifying entire signal events in testing sequences is evaluated. This provides an indication of how each method performs on a real soundscape analysis problem, where the goal is to tabulate the number of times and when each signal type is present. These results are based on the true signal locations, as determined by the annotation process detailed in Section 2.3. For instance, each method estimates the time of arrival, duration, and class label of a given signal event. If at least half of the set of observations associated with a manually annotated event (truth) are also in the set of observations associated with a detected signal event, and they additionally have the same class label, then the annotated event is considered correctly detected and classified. Missed detections result when too few or no observations in the annotated event are assigned a label associated with a signal, and misclassifications occur when the wrong label is assigned a majority of the time. False alarms occur when a signal event is thought to be present where there is none. Results are presented mainly in terms of confusion matrices, which are discussed in more detail below, though a few visual examples are also provided that show the results of applying each method to two-hour long segments from each data set to yield class labels for observations therein.

6.5.1. Kenai Fjords Site 4

The overall detection and classification results for the KEFJ004 data set are presented in terms of the confusion matrices in Table 6.1. Each entry in this table indicates the number of times a certain type of signal event was assigned a specific label by a given method (SCST / SRCT / GMM). Since “none” means no signal of interest (either present or assigned), the first column in each confusion matrix indicates instances where signal events are missed, whereas the first row indicates false alarms (no annotated event present). The shaded diagonal entries indicate the number of events of each signal type that are assigned the correct label, which show overall correct signal classification rates of 90.7%, 89.6%, and 79.8% for the SCST, SRCT, and GMM methods, respectively. False alarm rates are reported in terms of the percentage of all event detections (i.e.,

Table 6.1. KEFJ004 confusion matrix showing the total number of instances each signal type was assigned a given label by each method (SCST / SRCT / GMM).

		Assigned			
		None	**Plane**	**Heli**	**Jet**
Truth	**None**	-	3 / 1 / 18	1 / 0 / 7	8 / 7 / 1
	Plane	4 / 4 / 13	**117 / 120 / 112**	3 / 0 / 1	2 / 2 / 0
	Heli	3 / 9 / 10	1 / 2 / 4	**32 / 27 / 24**	2 / 0 / 0
	Jet	0 / 2 / 2	1 / 0 / 6	1 / 0 / 1	**17 / 17 / 10**

entries in the last three columns in each confusion matrix) that are false, which are 6.38%, 4.55%, and 14.1% for the SCST, SRCT, and GMM methods, respectively. This is also the reason "-" appears for the "none" diagonal entry in Table 6.1.

As can be seen from Table 6.1, the overall classification results produced by the SCST and SRCT methods are similar, but both are significantly better than those produced by the GMM method. The SRCT performs better than the GMM mainly because of the relative differences in performance of the associated source quantity tests. For example, the GMM missed more signals when interference was present than when absent, while the SRCT missed fewer signal events overall, since it was able to adapt to dual source scenarios due to the increased flexibility offered by the estimation process used to find the likelihood values. Confusion between signal types was also more common for the GMM method since these models were not quite as effective at estimating the signal types in the presence of interference. The GMM-based method produced significantly more false alarms owing to a less powerful source quantity test and the fact that the threshold for this test was set to detect all signals in the training segments. Most false alarms produced by the GMM-based method were associated with planes since strong thunder that is frequently present (interference alone) tends to resemble superimposed rain and plane signatures (dual source). Most false alarms produced by the SRCT method were for jets since the subspace for this source is close to weak thunder signatures.

Despite the superior signal detection performance of the SCST method reported in the previous section, the overall classification results produced by this method were not much better than those

TABLE 6.2. Correspondence between colors used in detection/classification strips and each signal type.

Signal Type	None	Plane	Helicopter	Jet
Color Code		🟥	🟩	🟦

of the SRCT method for the KEFJ004 data set. The main reason for this is that the test statistic evaluations were not influenced by the HMM-based sequential decision fusion applied to the SRCT and GMM class labels. In other words this decision fusion corrected some classification errors for individual observations to noticeably improve the overall performance of the SRCT and GMM methods. Since the differences in performance between each method is larger on the GRSA001 data, a more detailed discussion regarding their relative strengths and weaknesses is reserved until the results on this data set are presented.

To demonstrate a practical source characterization application, and to offer visual insight into the cause of some errors made by each method, Figs. 6.4 and 6.5 demonstrate the detection and classification results obtained for two different two-hour long testing segments from the KEFJ004 data set. Each segment presents a fairly difficult detection/classification problem with signatures of multiple signal sources that are mostly superimposed with those of competing interference. Each figure contains both the original data sequence (bottom) and corresponding detection/classification "strips" (top) that indicate true signal locations, as well as those estimated by each method, using different colored segments. When a strip is white it means that no source was detected for those samples, whereas colored strips indicate the presence of a specific type of signal as designated by Table 6.2. The method that produced the class label sequence represented by a given detection/classification strip is displayed to its left.

Comparing the "truth" detection/classification strips in Figs. 6.4 and 6.5 to those associated with each source characterization method shows that, in general, each method detects the signals for the appropriate time intervals and assigns accurate classification labels. A few errors can be noted in these examples, namely the SRCT method missing two faint and short helicopter events around the 41 min mark of the first hour and 15 min mark of the second hour in Fig. 6.4. These errors are due to confusion of the helicopter signatures with intense thunder alone, meaning the last stage of the process incorrectly rejected the appropriate dual source hypothesis in this case. The SRCT method also underestimated the durations of the first three helicopter events in Fig. 6.4. The SCST method performs very well on the data shown in Fig. 6.4, and produces only one false alarm that was a result of some novel interference signatures that resembled those of a distant plane. The GMM method also performs fairly well, though the first and third helicopter events in Fig. 6.4 are each broken into two separate detected events.

The detection/classification strips in Fig. 6.5 show that each method performs fairly well even in the presence of especially heavy rain and thunder. The main errors of note for this data segment are the SRCT method missing a jet and a helicopter, mostly due to their signatures being superimposed with those of heavy rain and thunder. The SCST method misclassified a jet as plane and produced a jet false alarm at the 20 min mark and 26 min mark of the second hour, respectively. The former error was likely caused by associating some of the jet's energy with the interference atoms during the sparse coding stage, leaving few atom coefficients left for discriminating between signal types. Finally, it can be seen that the GMM misclassified the helicopter and both jet events in Fig. 6.5 as plane, most probably due to the presence of heavy interference and the inadequacy of dual source GMMs for modeling the time-varying signatures of two superimposed source types.

6.5.2. Great Sand Dunes Site 1

The overall detection and classification results for the GRSA001 data set are shown in the confusion matrices in Table 6.3. The shaded diagonal entires indicate overall correct signal classification rates of 93.0%, 89.0%, and 80.4% for the SCST, SRCT, and GMM methods, respectively, while the false alarm rates are 3.75%, 11.0%, and 13.9% for the SCST, SRCT, and GMM methods, respectively. As can be seen, the overall classification results produced by the SCST method are noticeably better than those produced by the SRCT method, and significantly better than those

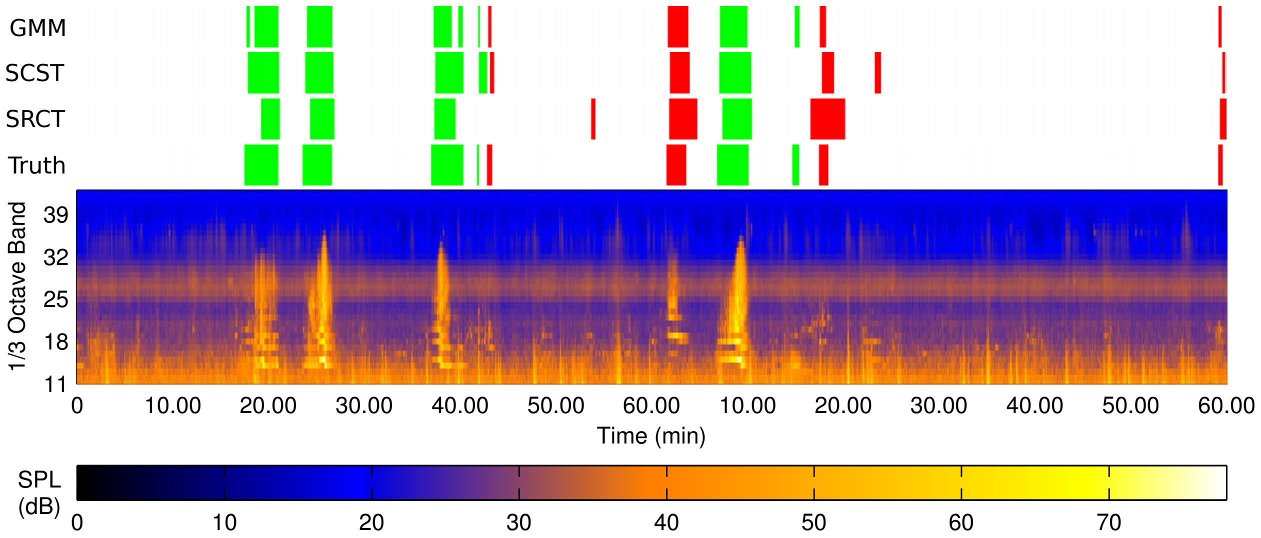


Figure 6.4. Results of applying each method to KEFJ004 data collected on 7/27/08 during hours 15-17.

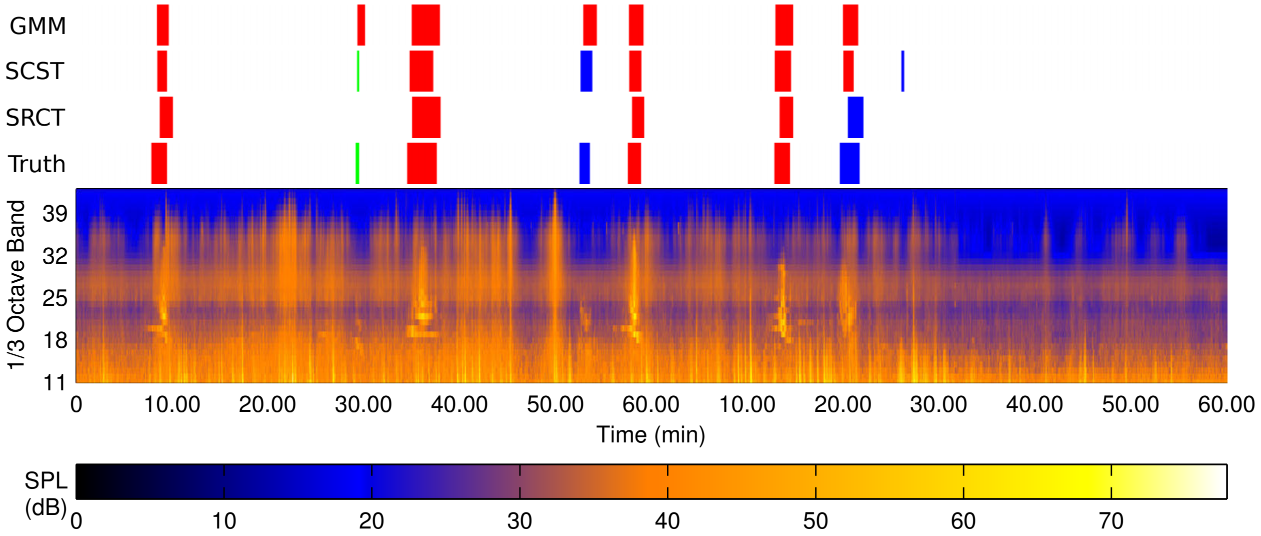


Figure 6.5. Results of applying each method to KEFJ004 data collected on 7/29/08 during hours 14-16.

produced by the GMM method, which is mainly due to the increased flexibility offered by the SCST model. The missed detections and false alarms in each case can be attributed to factors that influenced the ROC curves discussed above. As far as classifying detected signals, the gap in performance is caused by the drastically different approaches taken by each method. Although GMMs can approximate arbitrary distributions, there are still severe limitations on the accuracy of a finite mixture for modeling complicated acoustical signatures. In contrast, the SCST method makes no assumptions concerning the distributions of the signals, interference, and noise, but instead simplifies the data representation just enough so that likelihoods can be realistically computed. In other words, simplifying the data representation has provided superior class discrimination when compared to restricting the plausible structure of observation components. Moreover, the SCST method is generally better suited for adapting to sudden changes in the structure of source signatures considered in this study owing to, e.g., Doppler effects. For instance, the 1/3 octave bands that contain significant energy can rapidly change if a signal source has a high velocity and becomes relatively close to the receiver. For SCST, such a quick change conveniently manifests itself as a change in the atoms used in the sparse representation, which can easily be modeled by a BN. On the other hand, the SRCT method (for example) assumes the coordinates of the source signatures relative to a specific subspace evolve according to a linear AR vector model, which typically leads to greater errors when estimating rapidly changing source signatures.

The presence of interference was also less detrimental to the classification performance of the SCST method. The SRCT method misclassified a fair number of jets as planes, and vice versa, when strong wind was present, as the superposition of plane and wind signatures can resemble those of a jet. The sparse coding process used by the SCST method was able to associate the majority of the wind signatures to the interference atoms in these cases, meaning it did not impact signal detection and classification. It should also be noted that the SCST method is typically more robust to the presence of novel signatures within an event since it considers the joint likelihood of all observations therein. In contrast, the SRCT and GMM-based methods make decisions on individual observations (though the former also considers previous observations when assigning labels), and aggregate results using postprocessing.

To conclude this section, Figs. 6.6 and 6.7 present detection/classification strips produced by each method for two data segments in the GRSA001 data set. The data in Fig. 6.6 represents a very complex source characterization scenario, where heavy wind is present most of the time,

Table 6.3. GRSA001 confusion matrix showing the total number of instances each signal type was assigned a given label by each method (SCST / SRCT / GMM).

		Assigned		
		None	**Plane**	**Jet**
Truth	**None**	-	0 / 10 / 28	11 / 24 / 12
	Plane	3 / 9 / 11	**38 / 30 / 27**	2 / 4 / 5
	Jet	16 / 16 / 43	0 / 4 / 0	**242 / 238 / 215**

with extremely heavy wind during the first 40 minutes. Furthermore, many jets are present, and the signatures of two jets overlap in a couple of cases, namely near the 16 and 28 min marks of the second hour. The SCST method performs near perfectly on this data segment, with the only small errors being splitting each of two jet events into two separate detections; errors that could be corrected by using a larger quiescent detection threshold. The SRCT method also does fairly well on the data in Fig. 6.6, but also splits two jet events into separate detections, misclassifies the only plane as a jet, and misses the last jet event in the sequence. The former error was caused by the extremely heavy wind, meaning the superimposed signatures were closest to the jet subspace used by the SRCT method. A jet is missed by the SRCT method since its signatures were very weak, and reminiscent of wind. The GMM method performs the worst on the data in Fig. 6.6, as it misclassifies the plane as jet, generates three false alarms, and divides six separate events into two separate detections each. As before, the errors generated by the GMM method can mostly be attributed to inadequacies of the GMM for discriminating between single and dual source hypotheses when signatures of both sources overlap significantly.

The last example in Fig. 6.7 again demonstrates the superior performance of the SCST method on the GRSA001 data, as it only misclassified one jet event as plane near the 42 min mark of the second hour. The SRCT method also does very well, but some events are split into multiple detections again, and it produces three false alarms near the end of the segment. The latter error is due to the fact that the SRCT method assumes that a maximum of one type of interference will be present at a given time, whereas the observations at the end of this data segment contain both elk

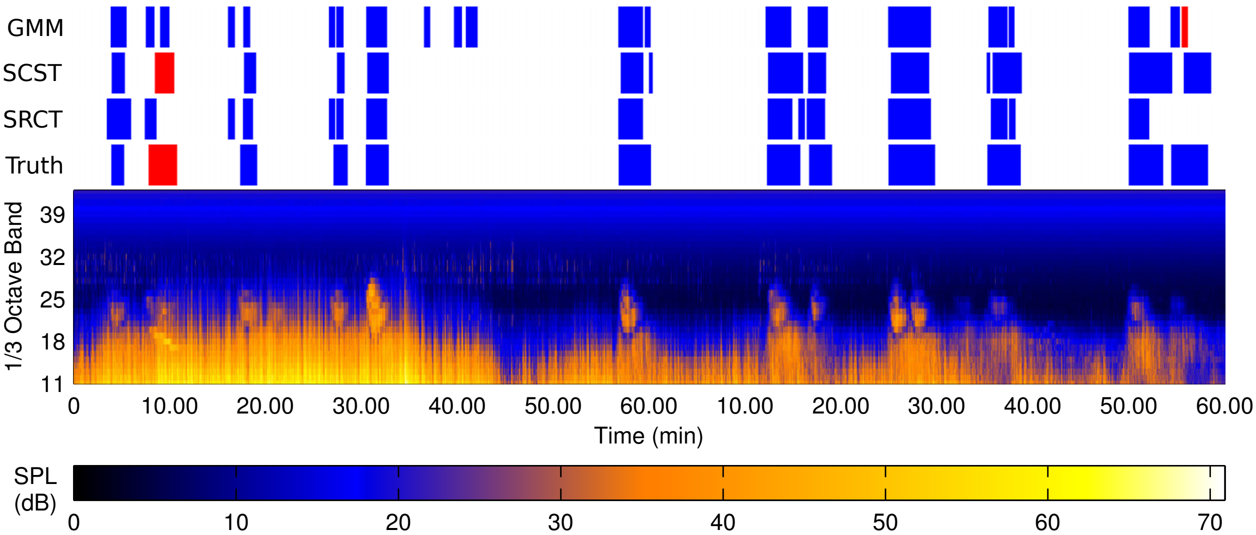


FIGURE 6.6. Results of applying each method to GRSA001 data collected on 9/28/08 during hours 20-22.

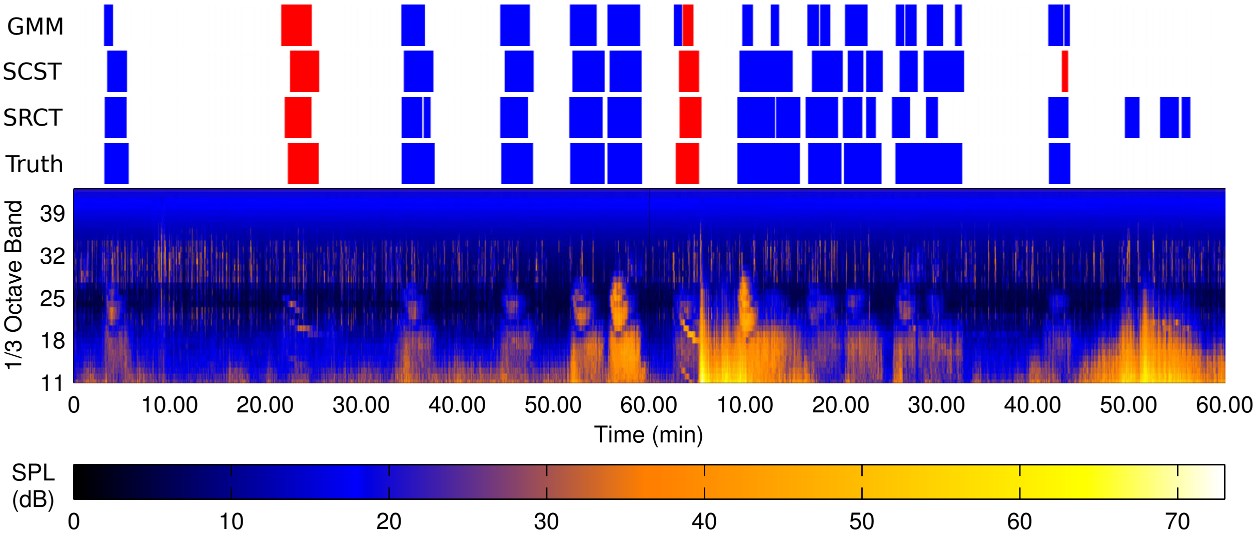


FIGURE 6.7. Results of applying each method to GRSA001 data collected on 9/29/08 during hours 21-23.

call and wind interference. This led to false acceptance of a dual source hypothesis, as the associated model was a better match to the data than a single interference model. The GMM method again splits many of the signal events in Fig. 6.7 into multiple detections, and underestimates the duration of many other events. However, the GMM does assign mostly accurate class labels.

6.6. Source Separation Results

While the SCST and SRCT methods were developed primarily for detection and classification of transient source, they each posses inherent capabilities for separating superimposed signal and interference signatures in a given observation, thereby producing isolated estimates of each component. Therefore, this section presents an evaluation of the ability of these methods for performing such separation. The SRCT method can estimate the signatures of one signal and one interference component from a given observation using the recovery equations in (4.28), which exploit the estimated basis coefficients corresponding to the accepted hypothesis. In general, the separation capabilities of the SCST method are dictated by the structure of the dictionary used for sparse coding. For example, if this dictionary consists of a set of source-specific subdictionaries that are mutually incoherent [54], then energy in an observation that is associated with different source types will be assigned to the appropriate subdictionaries during the sparse coding stage. An estimate of a given source's signatures can then be formed by using only the atoms associated with this source, i.e., by setting the coefficients of other atoms to zero before reconstruction. Since the dictionaries used to generate the results in this chapter were not specifically designed to be mutually incoherent, here we only emphasizes the ability of the sparse coding stage of the SCST method to separate the signal and interference components of an observation sequence, where each may contain any number of sources of the designated type. Results are presented in terms of explicit improvements to the signal-to-noise ratio (SNR) when each method is applied to sets of synthetically generated events containing superimposed signal and interference signatures, as well as a visual analysis of separated sequences.

6.6.1. SNR Improvement Using Synthetic Data

The separation performance of each method was tested in the following manner. First, for a given data set and for each source type (signal and interference), a set of data segments was identified and extracted from the testing segments that contain only the signatures of that source (plus ambient noise). Ten and twenty events per source type were used for the KEFJ004 and GRSA001 data sets, respectively, as signals occurred far less often in the former, and hence, isolated signal events were rare. Next, a unique set of synthetically superimposed events was generated for each possible signal and interference type pair (PQ sets total), each of which contained the same number events as the associated signal type, i.e., ten and twenty for the KEFJ004 and GRSA001 data sets, respectively. To generate the set associated with signal type p and interference type q in a given data set, the signatures of a unique type p signal event were added to the signatures of a randomly chosen type q interference event. Since KEFJ004 contains three signal types and two interference types, this procedure resulted in a set of $10 \times 3 \times 2 = 60$ superimposed events, whereas $20 \times 2 \times 4 = 160$ superimposed events were generated from the GRSA001 data.

Denote $\mathbf{Y}_1^n$, $\mathbf{S}_1^n$, and $\hat{\mathbf{S}}_1^n$ as matrices representing a length n synthetically superimposed data sequence (i.e., columns of $\mathbf{Y}_1^n$ represent individual 1/3 octave vectors), the sequence containing only signal components present in $\mathbf{Y}_1^n$, and a corresponding estimate of the signal sequence, respectively. Since $\mathbf{S}_1^n$ is known for each $\mathbf{Y}_1^n$, explicit separation results may be generated by comparing the SNR of the original superimposed event (input SNR) with the SNR of the signal estimate produced by a given method (output SNR). More explicitly, we compare

$$\text{Input SNR} = 10 \log \left(\frac{\|\mathbf{S}_1^n\|_F^2}{\|\mathbf{Y}_1^n - \mathbf{S}_1^n\|_F^2} \right) \text{dB}$$

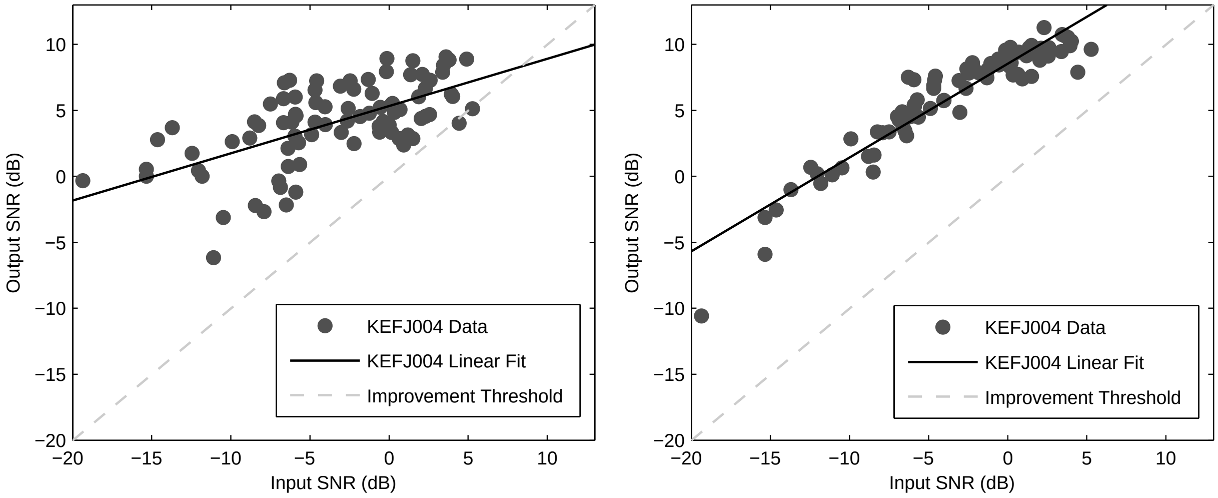
with

$$\text{Output SNR} = 10 \log \left(\frac{\|\mathbf{S}_1^n\|_F^2}{\left\|\hat{\mathbf{S}}_1^n - \mathbf{S}_1^n\right\|_F^2} \right) \text{dB}$$

where $\|\cdot\|_F^2$ means Frobenius norm, and interference present in $\mathbf{Y}_1^n$ is considered "noise" in this case.

Figs. 6.8(a) and 6.8(b) contain scatter plots that show the input SNRs versus the output SNRs achieved by the SRCT and SCST methods, respectively, when applied to the synthetically superimposed events associated with the KEFJ004 data set in order to generate an estimate of the signal component in each case. The solid line represents the linear least squares fit to the input/output SNR data points, which provides an overall indication of how the performance of each method changes with input SNR. The dashed line in each figure shows the improvement threshold, i.e., a point is only above this line if a given method improved the SNR of the sample. The average input SNR over all events was -3.32 dB, while the average output SNRs were 4.62 dB and 6.62 dB for the SRCT and SCST methods, respectively. As can be seen, the SCST method performs best overall on KEFJ004 events. However, the output SNR achieved by both methods was higher than the input SNR for each sample, with the exception of two events processed by the SRCT method that had relatively high input SNR. The improvement in the SNR for $\hat{\mathbf{S}}_1^n$'s generated by the SRCT method is clearly much greater for events with low input SNR, while events with high input SNR are improved only slightly on average. This is because the SRCT method assumes each source follows a subspace model with coefficients that obey a vector linear AR model, and hence, the estimated signal component in each case is a rough approximation of the true signal component. This means that, for high input SNR cases, the signal component is degraded nearly as much as the (relatively weak) interference component whereas, for low input SNR cases, removal of the prominent interference component is sufficient for drastically improving the overall SNR, despite the signal estimate being imprecise.

The SCST method maintains more consistent SNR improvement regardless of the input SNR for a given $\mathbf{Y}_1^n$, which is demonstrated by the slope of the associated linear fit being closer to one. This is due to the fact that the sparse coding stage is typically able to associate both high and low levels of interference with the appropriate atoms, which are then discarded when constructing the signal estimate. The downside to this sparse coding process is that a small amount of the



(a) SRCT method.

(b) SCST method.

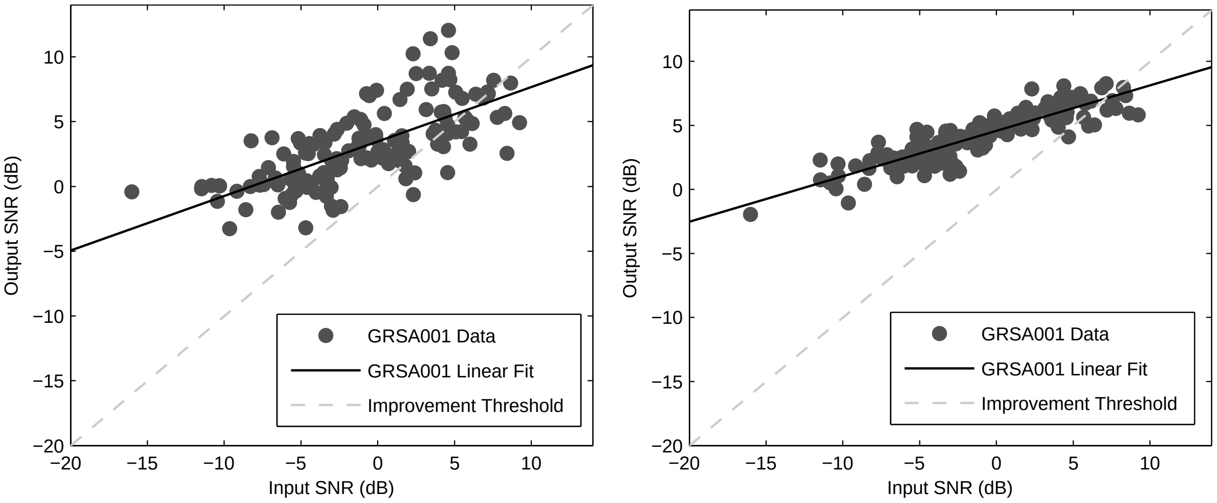
FIGURE 6.8. Improvement in SNR achieved by each method when applied to synthetically superimposed signal and interference data segments from the KEFJ004 data set.

signal energy is typically assigned to the interference atoms during sparse coding, and vice versa, which negatively impacts the output SNR regardless of the level of input SNR. This is caused by larger than desired coherence between atoms associated with signal and interference subdictionaries [54], which is common in cases when these different components have similar morphologies [79]. Although a few methods for sparse dictionary learning have been proposed [81, 82] that attempt to simultaneously constrain the coherence between different dictionaries while retaining their reconstruction capabilities, they involve solving multi-objective optimization problems that do not guarantee convergence, making it difficult to extract suitable dictionaries. Also of note is the relatively large amount of variance in output SNR produced by the SRCT method for a given input SNR level, when compared to the SCST method. This can be attributed to the inadequacy of the models used by the former method for properly estimating the subtle variations of signal events, i.e., the SRCT models do not quite capture the severe with-class diversity present in dynamic acoustical sources. The estimates formed by the SCST method, on the other hand, do not make restrictive assumptions about the behavior of these signatures, but rather simply require that the dictionary be capable of providing a sparse representation of any signal event that may be encountered.

Figs. 6.9(a) and 6.9(b) also contain input/output SNR scatter plots for the SRCT and SCST methods, respectively, but for synthetically superimposed events associated with the GRSA001 data set. The average input SNR over all events was -1.86 dB, while the average output SNRs were 3.08 dB and 4.47 dB for the SRCT and SCST methods, respectively, meaning the SCST method again performs best overall on GRSA001 events. These scatter plots show that the KEFJ004 and GRSA001 separation results have many of the same features in common, e.g., higher output SNR variance for the SRCT method. However, the GRSA001 results reveal some unique behavior, namely that the improvement in output SNR produced by the SCST is less drastic for high input SNR cases, which is also the case with the SRCT method for both data sets. This result can be attributed to the same reasons mentioned above, where some signal energy is associated with interference atoms during the sparse coding stage of SCST. This effect is more dramatic for the GRSA001 data set since the signatures of heavy wind interference are very close to those of jet signals, more so than any other signal and interference combination considered in either data set. Greater improvements are witnessed for low input SNR cases since the sparse coding still removes most of the interference, which is sufficient for bringing the signal estimate much closer to the true signal.

6.6.2. Examples using Real Data

To supplement the above discussion on separation performance, Figs. 6.10–6.13 provide visual examples of the signal and interference signatures estimated by the SRCT and SCST methods when applied to two data segments from each of the data sets. For comparison, examples showing the signatures of each source type in isolation can be found in Tables 2.1 and 2.2 for the KEFJ004 and GRSA001 data sets, respectively. Note that the SRCT method inherently estimates interference signatures using the same basis coefficient estimation approach as in the signal estimation case, while



(a) SRCT method.

(b) SCST method.

FIGURE 6.9. Improvement in SNR achieved by each method when applied to synthetically superimposed signal and interference data segments from the GRSA001 data set.

the SCST interference estimate is a result of using only the interference atoms in the dictionary, along with their corresponding sparse coefficients, to reconstruct the data sequence. Additionally, the original data sequence in each figure has the noise mean subtracted to better highlight the signal and interference components.

Fig. 6.10 shows a relatively simple sequence from the KEFJ004 data set, where some light rain and birdsong signatures are briefly superimposed with signals that are planes, helicopters, and a jet. The SCST signal and interference estimates in Figs. 6.10 (b) and 6.10(c), respectively, appear to be very accurate representations, as the former contains almost no interference, while the latter contains only some faint helicopter signatures around the 35 and 48 min marks in the first hour. The SRCT signal estimate in Fig. 6.10(d) contains no noticeable interference, but the signal signatures are somewhat blurred, which is due to basing the estimates on a somewhat restrictive model, as described above. The SRCT interference estimate in Fig. 6.10(e) appears to be a mostly faithful representation, apart from dark patches in the low frequency regions where rain is present, namely

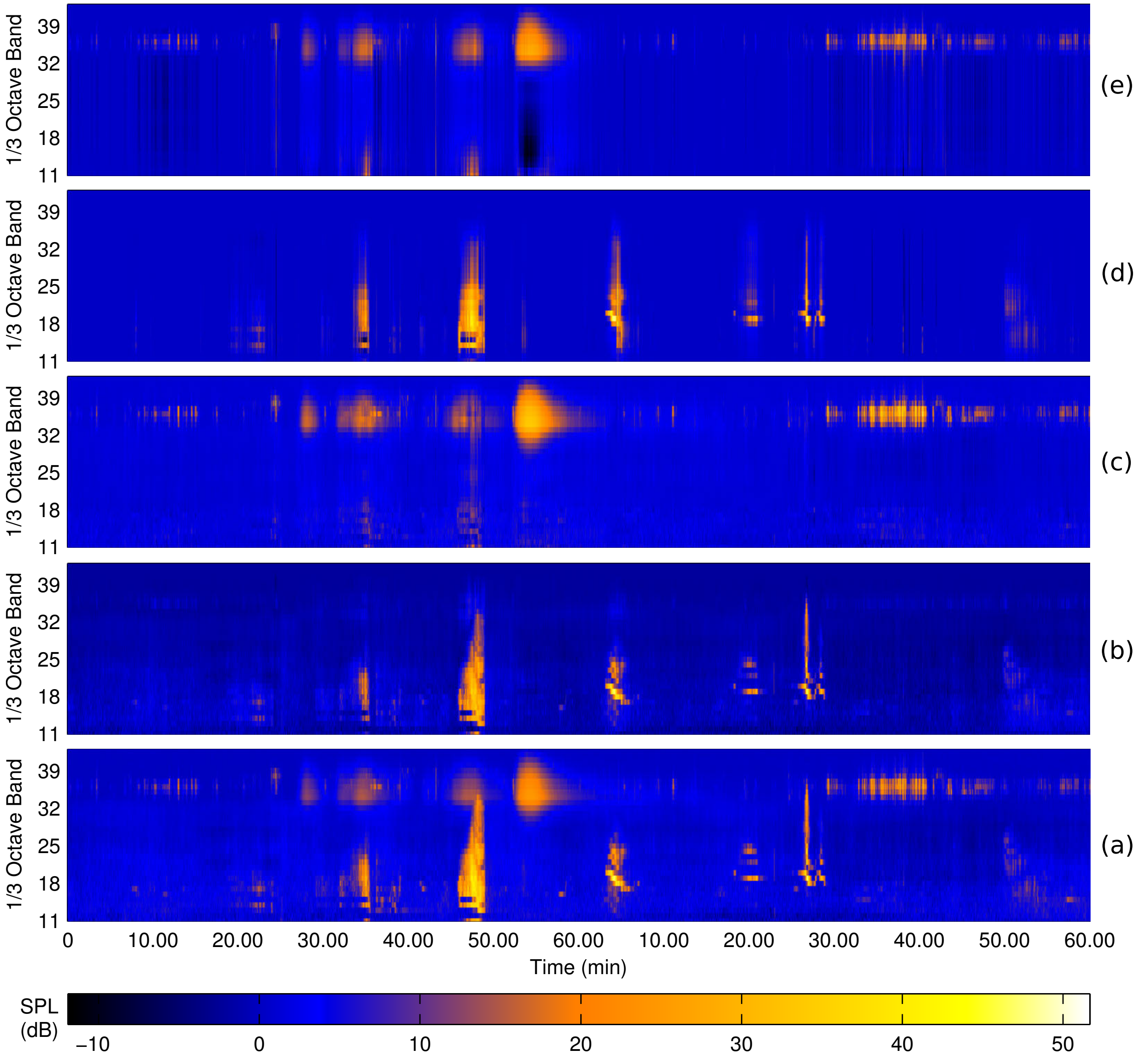


FIGURE 6.10. Separation of signal and interference components present in the data segment in (a), which corresponds to KEFJ004 recordings during hours 10-12 on 7/25/08. (b) SCST signal image, (c) SCST interference image, (d) SRCT signal image, (e) SRCT interference image.

around the 10 and 55 min marks of the first hour. Such errors are caused by the limitations of a three-dimensional subspace model for constructing observations with high variance.

Fig. 6.11 shows the separation results for another data segment in the KEFJ004 data set, where moderate rain and thunder are present the entire time, and superimposed with the signatures of a number of planes. The SCST method again does a fantastic job of separating the two components,

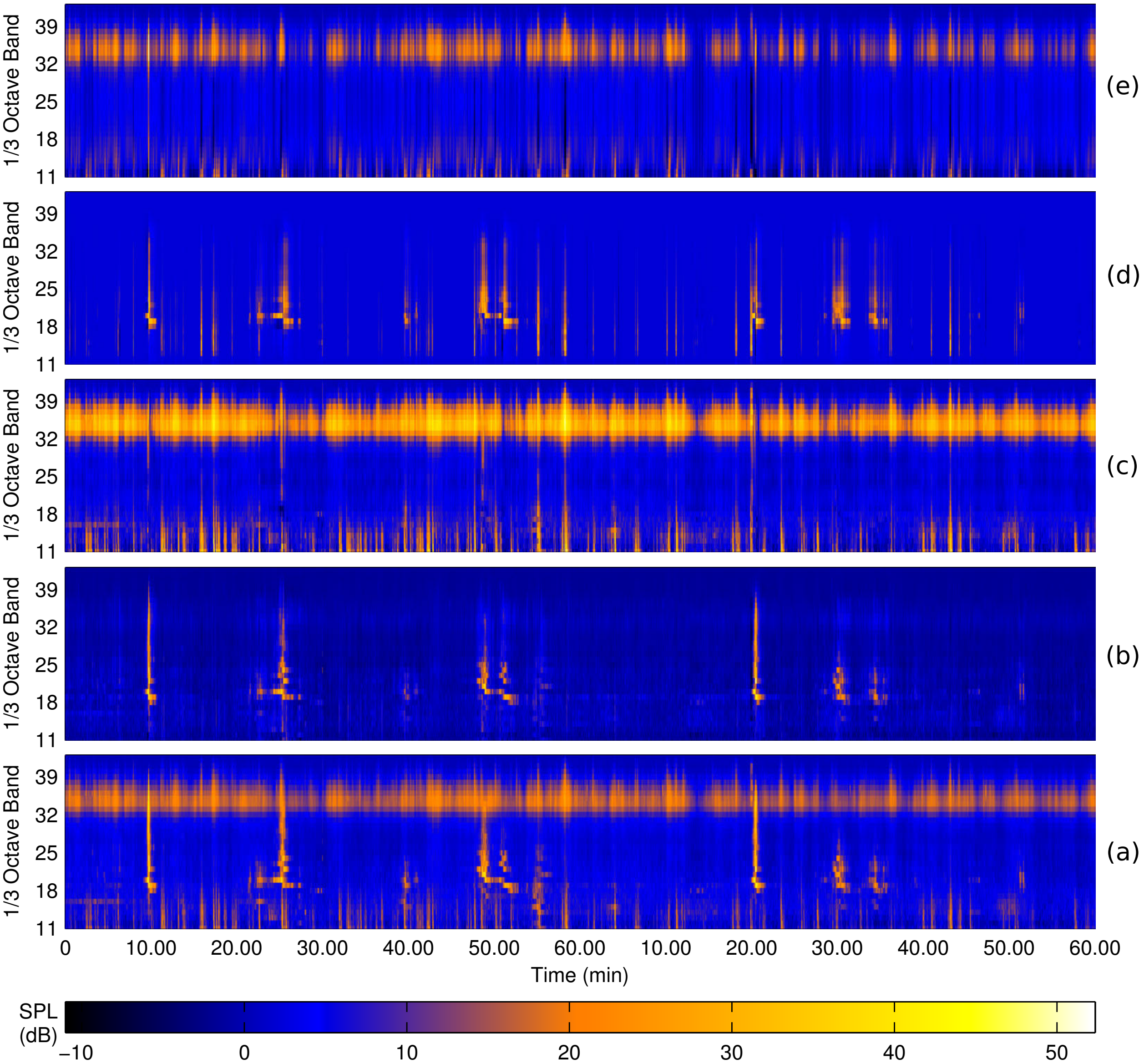


FIGURE 6.11. Separation of signal and interference components present in the data segment in (a), which corresponds to KEFJ004 recordings during hours 14-16 on 7/30/08. (b) SCST signal image, (c) SCST interference image, (d) SRCT signal image, (e) SRCT interference image.

as no major errors are present. The SRCT method mostly produces seemingly accurate signal and interference estimates, though as before the former is somewhat blurry. Furthermore, some of the energy caused by thunder interference is present in the signal estimate owing to it resembling the broadband portion of plane signatures, and hence, it was associated with this signal's model.

Fig. 6.12 shows the separation results for a segment in the GRSA001 data set, where loud elk calls and light wind represent the interference, and a number of plane and jet events represent the signals. The SCST method is able to associate most of the signatures of the elk calls and heavier wind with the interference sequence, but light wind signatures are present in the estimated signal sequence. This is most noticeable from the 29–42 min mark during the first hour in Fig. 6.12(b), and is caused by high coherence between some atoms in the wind and jet subdictionaries. The estimates produced by the SRCT method share many of the same properties mentioned above, i.e., rather blurry signal estimates and some association of wind interference with the signal sequence.

Finally, Fig. 6.13 shows one last separation example from the GRSA001 data set, where moderate to extremely heavy wind is superimposed with the signatures of many jets and one plane. The SCST estimate in Fig. 6.13(b) shows a seemingly accurate representation of the signal component, but also includes of some heavy wind interference energy. Although this type of error means less accurate estimation performance, it typically does not negatively effect classification performance, as the $\mathcal{H}_0$ model for SCST is trained on such sequences containing heavy interference. The SRCT signal estimate in Fig. 6.13(d) is perhaps the most blurry of all, owing to the severe overlap between signal and interference signatures, and consequent difficulty of separating these two components. The interference estimates in Figs. 6.13(c) and Fig. 6.13(e), produced by the SCST and SRCT methods, respectively, both seem fairly accurate, though the latter has some gaps in the mid-frequency range when heavy wind was present, since this energy was associated with the signal estimate. Overall, these results demonstrate that both methods are able to generate reasonable estimates of the signal and interference components of an observation sequence, though the SCST method performs best overall.

6.7. Computational Complexity

As a final evaluation measure, the computational cost of the SRCT, SCST, and GMM methods for processing a single observation $\mathbf{y}_k \in \mathbb{R}^N$ is considered. This analyzes assumes all the models and

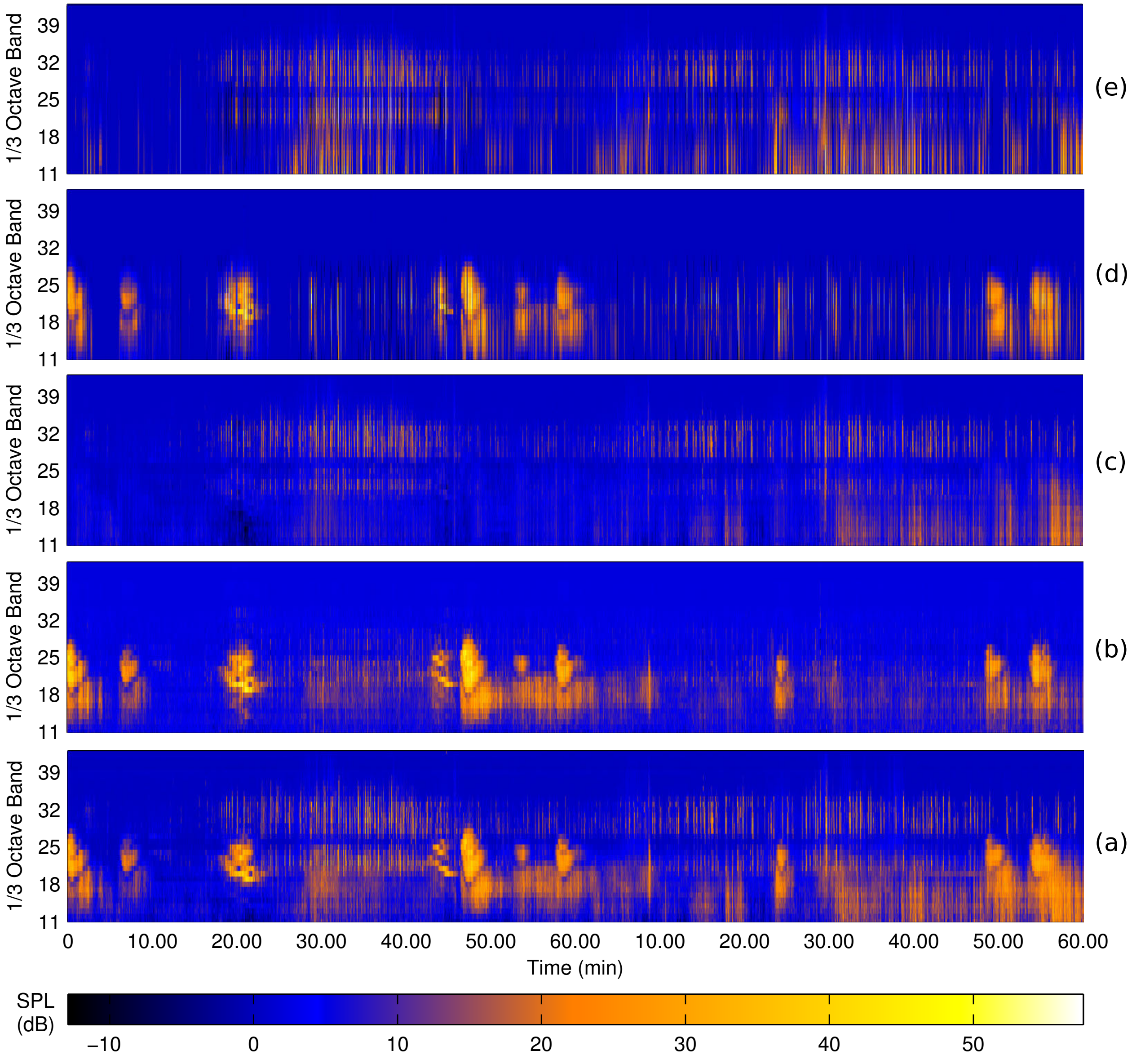


FIGURE 6.12. Separation of signal and interference components present in the data segment in (a), which corresponds to GRSA001 recordings during hours 19-21 on 9/24/08. (b) SCST signal image, (c) SCST interference image, (d) SRCT signal image, (e) SRCT interference image.

parameters (e.g., the sparse coding dictionary used by the SCST method) are all computed off-line, as was the case for the results reported above. For the SRCT method, this cost is driven by the processes with the largest associated growth rates, that are required to find the variables $\boldsymbol{\zeta}_k(\boldsymbol{\theta}_k)$'s in (4.8) for the PQ dual source hypotheses (out of all $P+Q+PQ$ hypotheses). In particular, the most costly processes are implementation of the Kalman filters and calculating $\boldsymbol{\zeta}_k(\boldsymbol{\theta}_k)$'s directly, which

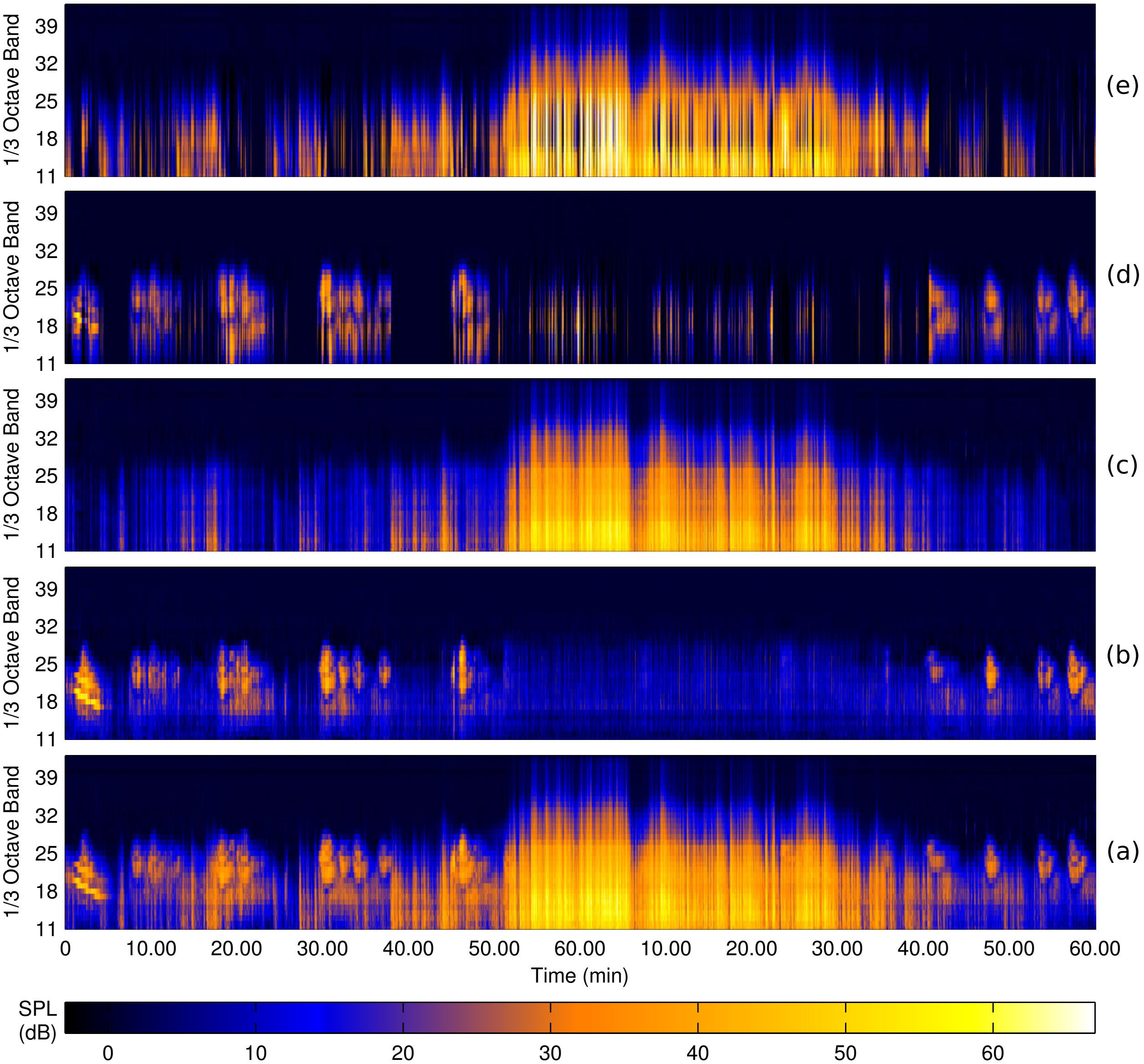


FIGURE 6.13. Separation of signal and interference components present in the data segment in (a), which corresponds to GRSA001 recordings during hours 16-18 on 10/02/08. (b) SCST signal image, (c) SCST interference image, (d) SRCT signal image, (e) SRCT interference image.

require $O(PQD^{2.373})$ and $O(PQN^{2.373})$ operations, respectively, if efficient implementations are used, where D is the model subspace dimension. Therefore, since $N > D$ necessarily, $O(PQN^{2.373})$ represents the SRCT algorithm cost. For the GMM method, the matrix inversions and determinants required to find the likelihoods may be calculated off-line, but a likelihood must be found for each component in a GMM. Assuming an average of K components in each GMM, the cost of this method is then $O(KPQN^2)$ operations, which is very comparable to that of the SRCT method for the data considered in this thesis.

When considering the SCST method, the cost of the sparse coding process dominates its overall computational complexity. If basis pursuit denoising [80] is used, then this process involves finding the solution to a quadratic programming problem, which can be accomplished with a wide variety of algorithms, each with different complexities that also depend on tolerated error δ between the observation and sparse representation. Orthogonal matching pursuit generally requires fewer computations and is recommended for applications where N and the number of atoms M in the sparse coding dictionary are large. Though there are still various ways to implement orthogonal matching pursuit, the cost of computationally efficient algorithms is about $O(LNM)$ [92], where L is the number of iterations (sparsity level), that depends on δ. In the absolute worst case scenario, $L = N$, meaning $O(MN^2)$ operations are required for sparse coding. Otherwise, SCST simply requires updating the signal and quiescent detection statistics, which is very simple since the distribution parameters are computed off-line, requiring only $O(PM_s)$ operations, where $M_s \leq M$ is the number of atoms associated with signal components.

Overall, the relative computational complexity of each method depends on M (for SCST only) and the number of signal sources P and interference sources Q, though it is reasonable to assume that the cost of SCST is similar to the SRCT and GMM methods for many applications, so long as an efficient sparse coding algorithm is used. For the results in this chapter, the SCST (using basis pursuit denoising), SRCT, and GMM methods took an average of 5.87×10^{-2}, 5.70×10^{-3}, and 2.60×10^{-2} seconds, respectively, to process a single observation using MATLAB on a computer with a 3.2 GHz quad-core processor and 8 GB of RAM. Clearly, the SCST method was the slowest in this case, but it still processed the data about 17 times faster than the data sampling rate of one observation per second.

6.8. Conclusions

This chapter presented a comprehensive performance evaluation to determine the relative strengths and weakness of the SRCT, SCST, and GMM methods for performing transient source detection, classification, and estimation using 1/3 octave data sequences representing real acoustical recordings of national park soundscapes. The SRCT and SCST methods were introduced in Chapters 4 and 5 of this thesis, respectively, while the GMM method was briefly introduced in Section 6.2, and was used as the benchmark method since this approach is common in acoustical source characterization applications. The performance of the detection, dominant source, and source quantity test statistics implemented by the SRCT method were first benchmarked against the equivalent tests implemented by the GMM method using ROC curves and the KEFJ004 data set. The SCST method was left out of this comparison since it does not rely similar test statistics. It was found that the SRCT and GMM approaches to calculating likelihoods for the detection and dominant source tests performed equally well, with both achieving AUCs > 0.965. However, the SRCT method performed much better on the source quantity test, as the associated AUC was 0.953, while the GMM AUC was only 0.806. This means that the SRCT method is generally more proficient at detecting the presence of a signal buried in interference.

A separate ROC analysis was used to directly compare the performance of all three methods for detecting signals, regardless of the presence of interference. For the KEFJ004 data set, the SCST, SRCT, and GMM methods achieved AUCs of 0.967, 0.863, and 0.889, respectively, while for the GRSA001 data set these methods achieved AUCs of 0.962, 0.863, and 0.845, respectively. The SRCT and GMM methods performed similarly for signal detection since they use a similar hierarchy of tests. However, in this case, the GMM compensated for its inferior performance on the source quantity test with higher P_D for very low P_{FA} values, when compared to the SRCT method. Still, the SCST method clearly displayed the best overall signal detection performance, owing to the cumulative nature of its detection statistic, which remains sufficiently high even when an event contains some observations with weak signal components.

The overall detection/classification performance of each method was then evaluated on each data set using confusion matrices. For the KEFJ004 data set overall correct classification rates of 90.7%, 89.6%, and 79.8% were achieved by the SCST, SRCT, and GMM methods, respectively, while the false alarm rates were 6.38%, 4.55%, and 14.1% for these methods, respectively. The performance of the SRCT method approached that of the SCST method on this data set since the former benefited from the HMM-based decision fusion introduced in Section 3.3. The GMM method performed the worst overall mainly due to a weak source quantity test, as mentioned above, leading to poor performance in heavy interference. A greater gap in performance was witnessed for the GRSA001 data set, where overall correct signal classification rates of 93.0%, 89.0%, and 80.4% were achieved by the the SCST, SRCT, and GMM methods, respectively, while the false alarm rates were 3.75%, 11.0%, and 13.9%, respectively. The main reason the SCST method performed best on the GRSA001 data is its ability to remain robust to the simultaneous presence of multiple types of interference, while the other two methods assume a maximum of one type of interference is present in a given observation.

The ability of the SRCT and SCST methods to generate accurate estimates of the signal and interference components of an observation sequence was then evaluated. Testing event sequences were synthetically generated for each data set that each contained the signatures of one type of signal and one type of interference, and each method was tasked with estimating the signatures of the former. On average, the SRCT method was able to improve the SNR of signal sequences associated with the KEFJ004 data set from -3.32 dB to 4.62 dB, while the SCST method improved the SNR to 6.62 dB. For sequences associated with the GRSA001 data set, the average input SNR was -1.86 dB, and the SRCT and SCST methods were able to improve this average to 3.08 dB and 4.47 dB, respectively. Clearly, the SCST method provided the best source separation results overall, which is mostly due to the SRCT method's reliance on a linear vector AR model to estimate the temporal evolution of a signal event, whereas the SCST model is far more flexible.

Overall, determining which method to use depends on the requirements of a specific application. If interference class labels are required, or low computational cost is a priority, then use of the SRCT method is recommended. On the other hand, if the data contains observations corrupted with multiple types of interference simultaneously, or contains weak signal events that could benefit from the use of a cumulative detection statistic, then the SCST method is recommended. The SCST method is also preferred when sources of interest do not obey the relatively more strict model assumed by the SRCT method, e.g., for speech recognition applications.

CHAPTER 7

Conclusions and Suggestions for Future Work

7.1. Summary and Observations

This thesis considered methods and concepts for characterization of multiple time-varying transient sources using sequential multivariate data. This involves detecting on the onset of new transient events, estimating their durations, assigning corresponding class labels, and possibly forming estimates of their signatures. The primary motivation for conducting the present study was tabulating the presence and properties of extrinsic acoustical sources present in national parks. Since the associated soundscapes were recorded for months at a time, and are represented using 1/3 octave vector sequences, developing approaches to handle this compressed data format was essential. This problem carries many intrinsic challenges such as the frequent presence of prominent sources of interference whose signatures are superimposed with sources of interest (signals), and erratic source signatures that lead to extreme within-class diversity and between-class similarities. Prior to the work in this thesis, these complications prevented development of a comprehensive solution capable of properly and automatically analyzing the national park soundscape data, despite the fact that such capabilities would be useful in a variety of applications, e.g., medical diagnosis using magnetic resonance images and target detection using sonar data.

The developments in this thesis address various aspects of the transient source characterization problem, and include complete solutions that can accommodate all of the intricacies of national park soundscape data, yet remain flexible enough to allow straightforward adaptation to other application areas. The major contributions of this work are summarized below.

(1) **A new sequential random coefficient tracking (SRCT) method.**

The SRCT method [74] was introduced in Chapter 4 as a comprehensive approach to transient source characterization using sequential multivariate data, that satisfies all of the established requirements for successful solutions outlined in Section 1.3. This method is capable of detecting, classifying, and estimating the signatures of a maximum of one signal and one interference source in each observation, by establishing source composition hypotheses and associated models for each combination. The models are based on assuming the signatures of each source lie in low dimensional subspaces, and that the associated random basis coefficients obey a linear autoregressive (AR) vector model. A Kalman filter estimates these basis coefficients for each source type and under each hypothesis, which allows for generating the necessary model parameters. The likelihoods of various parameter sets given the observation are used to form statistics for conducting a hierarchy of tests to determine the composition of the observation in terms of a signal, interference, and noise component. The estimated coefficients for the accepted hypothesis may also be used in conjunction with the associated basis vectors to form estimates of the actual source signatures present in a given observation, thus performing separation in dual source cases.

The SRCT method is most useful when analyzing data that contains at most one type of signal and one type of interference simultaneously, and additionally when a class label for interference is desired. The SRCT method is also the least computationally intensive of all the comprehensive approaches introduced in this thesis, and is therefore preferred for large dimensional data, or when a large number of source types must be considered. It is also simple to extract SRCT models from training data and apply them to new testing sequences, since only two parameters are required in the former case (subspace dimension and AR model order), while only three parameters are required in the latter case (a threshold for each test in the hierarchy).

(2) **A new sparse coefficient state tracking (SCST) method.**

The SCST method was developed and introduced in Chapter 5 to address a few practical concerns associated with using the SRCT method to characterize complex soundscape data. In particular, the SRCT method is not robust to the simultaneous presence of multiple types of interference, which can be an issue when analyzing, e.g., some natural soundscape data containing a large number of different types of acoustical interference related to weather effects and wildlife vocalizations. Furthermore, the SRCT source and noise models may not be appropriate in certain scenarios. For example, although Appendix B demonstrates that it is generally reasonable to assume noise is multivariate Gaussian in the 1/3 octave domain, the presence of certain ambient acoustical sources (e.g., waterfalls) may invalidate this assumption. Lastly, the ability to perform source characterization using other multivariate data formats that do not have Gaussian noise, e.g., Mel-frequency cepstral coefficients [51], increases the relevance of a given approach.

Like the SRCT method, the SCST method also meets the requirements outlined in Section 1.3, but does so by placing very few restrictions on the source signatures and noise. Instead, the data is simplified by first finding a sparse approximation of a given observation, which makes the temporal evolution of nonstationary acoustical signatures more tractable. This sparse coding stage provides inherent robustness to the simultaneous presence of multiple types of interference since these components may be separated from those of any signal that is present, assuming the coding dictionary is properly designed. The resulting sparse coefficients are then quantized to levels that are designed to maximize the discriminatory information they contain as measured by the J-divergence [85, 86] between different hypotheses, i.e., coefficient "states" are extracted. A Bayesian network may be trained for each source type that models the conditional distributions of sparse coefficient states extracted from corresponding 1/3 octave signatures, when given previous coefficient states. This allows for calculating the likelihood of a particular source model given a data sequence whose length constantly increases. Likelihood ratio tests (LRT) are formed to detect entire signal events during a quiescent phase (absence of a signal), and subsequently detect the next quiescent phase to effectively estimate the duration of the current signal event.

The SCST method has proven to be highly versatile and offer superior signal detection, classification, and estimation performance in most soundscape characterization scenarios. However, it is still only capable of detecting a maximum of one type of signal at a time, and is not capable of classifying interference. Additionally, it is more computationally intensive than the SRCT method in many circumstances, unless a very efficient sparse coding strategy is used [92].

(3) **Sequential decision fusion using hidden Markov models (HMM).**

Chapter 3 explained that it is sometimes beneficial to make classification decisions on individual observations separately, namely when the compositions of these observations in terms of signal, interference, and noise components change frequently. However, this approach can lead to decision sequences that do not provide an accurate description of the data contents in terms of the quantity and properties of acoustical sources that were present. Therefore, Section 3.3 introduced a new sequential decision fusion scheme that is capable of aggregating a stream of decisions to yield detection and classification results in terms of entire acoustical events. This approach finds the likelihood of each of several different HMMs [36], that are associated with different signal types, given a sequence of preliminary decisions (signal class labels) assigned to individual observations. These likelihoods are then used in to form a set of LRTs [35] that can be tracked to identify segments of data that contain unique acoustical events. This represents a new approach to multi-class sequential decision fusion, as most of the existing work in this area [43] considers binary hypothesis tests and/or fixed length data.

The proposed fusion was used to enhance the decision sequences made by the SRCT and Gaussian mixture model (GMM)-based methods that resulted from their application to the soundscape data in Chapter 2. As shown in Section 6.4, when the signal detection performance of the SCST and SRCT methods were analyzed using receiver operator characteristics (ROC), it was found that the former provided superior performance on both data sets, as decision fusion did not impact these tests. On the other hand, as indicated in Section 6.5, the overall detection and classification results on the KEFJ004 data set were similar for both of these methods, since the SRCT results were improved by the proposed decision fusion.

(4) **Comprehensive performance evaluation.**

In Chapter 6, the SRCT and SCST methods were benchmarked against a GMM-based method to determine the effectiveness of each for detecting, classifying, and estimating the signatures of transient sources. These results represent the first comprehensive performance evaluation that was conducted using 1/3 octave soundscape data collected in national parks, and furthermore, the first time any automated source characterization method was successfully applied to this data. In summary, by using receiver operator characteristic (ROC) curves it was found that the SRCT and GMM methods performed similarly when it came to detection (signal and interference) and determining the dominant source that is present in a given observation, but the former performed much better when it came to determining the quantity of sources, making it more proficient at detecting signals in the presence of interference. When it came to signal detection performance (regardless of interference), the SCST performed best overall by achieving an area under the ROC curve (AUC) of 0.967 for the KEFJ004 data, whereas AUCs of 0.863 and 0.889 were achieved by the SRCT and GMM methods, respectively.

The overall detection and classification performance of each method were also reported in terms of confusion matrices, which showed that the SCST and SRCT methods achieved similar performance on the KEFJ004 data, though both of these methods performed much better than the GMM-based approach. The largest performance differences were witnessed on the GRSA001 data, where overall correct classification rates of 93.0%, 89.0%, and 80.4% were achieved by the SCST, SRCT, and GMM methods, respectively, with corresponding false alarm rates of 3.75%, 11.0%, and 13.9%, respectively. The main reasons for this performance gap are the ability of the SCST method to remain robust to multiple types of interference that are simultaneously present, as well as the flexibility of the Bayesian network model it is based on.

Chapter 6 also demonstrated the ability of the SRCT and SCST methods to generate estimates of the signatures of signal sources present in an observation sequence that is corrupted by heavy interference. The SCST method again performed best in this regard by improving the average signal-to-noise ratio (SNR) of interference-laden images associated with the KEFJ004 data from -3.32 dB to 6.62 dB, whereas the average SNR of SRCT signal estimates was 4.62 dB. Similar performance was also witnessed for the GRSA001 data set. Finally, the computational complexity of each method was compared, where the SCST method was found to have to highest cost, both theoretically and in practice. It was found that the SCST (using basis pursuit denoising), SRCT, and GMM methods took an average of 5.87×10^{-2}, 5.70×10^{-3}, and 2.60×10^{-2} seconds, respectively, to process a single observation using MATLAB on a computer with a 3.2 GHz quad-core processor and 8 GB of RAM.

7.2. Future Work

There are several important theoretical and experimental research areas related to this problem that can be pursued as part of future research efforts. These include, but are not limited to:

(1) **Extend existing source characterization methods to handle multiple signal events that are simultaneously present.**

Perhaps the biggest deficiency of the proposed source characterization methods for properly analyzing the data in Chapter 2 is that they are unable to detect when multiple signals are simultaneously present. In such cases, both the SRCT and SCST methods will report the presence of the most dominant source within a given time interval, while weaker sources will go unnoticed; although this happens very infrequently for the present application. Fig. 7.1 shows an example of a 1/3 octave data sequence that contains several instances of overlapping signatures from different signals, which is common for data collected in parks with high air traffic, e.g., Grand Canyon and Yosemite National Parks. This is a challenging problem that

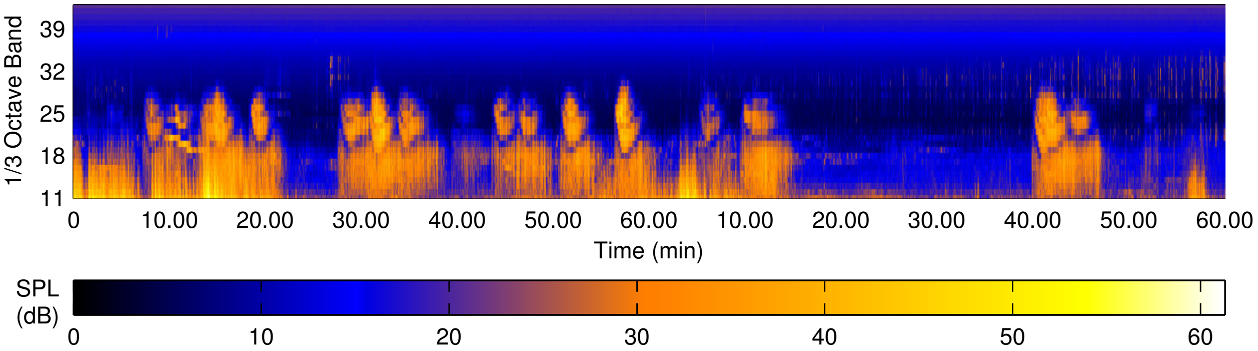


FIGURE 7.1. Example 1/3 octave data sequence containing multiple instances of overlapping signatures associated with different signal sources.

has not been addressed in cases where the data arrives sequentially and is multivariate. Some research in the area of blind source separation [93] may be a proper candidate for extending, but most existing work only applies to fixed-size time series data and/or when multiple independent realizations (e.g., multiple microphones) of the data are available.

(2) **Evaluate the performance of the SCST method using a more efficient sparse coding method.**

The experimental results produced by the SCST method reported in Chapter 6 were obtained by using basis pursuit denoising [54, 80] to obtain sparse approximations, which involves solving a computationally expensive quadratic programming problem. This sparse coding strategy is well-suited to the SCST method in that it consistently produces sufficiently sparse sequences that can be easily modeled using a Bayesian network, but its complexity means slow processing for high-dimensional data and potentially significant challenges associated with direct implementation of the SCST algorithm on acoustical monitoring stations. It would therefore be worthwhile to evaluate the detection and classification performance of the SCST method when using a computationally efficient sparse coding strategy, such as a fast implementation of orthogonal matching pursuit [92]. Ideally, the performance loss resulting from the new sparse coding strategy would be negligible, while introducing the benefits mentioned above.

(3) **Design fast implementations of the proposed algorithms.**

Section 1.1 thoroughly discussed the primary motivations for developing the algorithms presented in this thesis, among which was circumventing the current approach for National Park soundscape characterization, which is based on manual analysis by operators. The results in Chapter 6 demonstrated that automatic post-mission soundscape characterization using the proposed algorithms is indeed feasible, but it would be even more beneficial to implement the best performing and most versatile algorithm directly on acoustical monitoring stations for real-time soundscape analysis during deployment. Some components of the algorithms, such as the sparse coding step of the SCST method, would be complicated to implement efficiently using, e.g., a hardware description language. On the other hand, this task would have numerous benefits such as real-time reporting of abnormal acoustical conditions within a park and more widespread analysis of various sites.

(4) **Develop a robust method for extracting mutually incoherent sparse coding dictionaries for use with the SCST method.**

The results in this thesis that were produced by the SCST method used K-SVD [55] to extract source specific sparse coding dictionaries, which in general are designed to minimize the error between a training observation set and its sparse reconstruction for a given level of sparsity. While such dictionaries generally provided sparse coefficient state sequences that offered sufficient discrimination between the signatures of different source types, it was often the case that signatures associated with a given source type were not exclusively assigned to the associated source-specific dictionary during the sparse coding process. This is a result of excessive coherence between atoms associated with different sub-dictionaries [54], which is common when their associated signatures have similar morphologies [79]. While methods for sparse dictionary learning exist [81, 82] that attempt to extract a set of dictionaries that simultaneously have small mutual coherence and retain their reconstruction capabilities, they involve solving multi-objective optimization problems that do not guarantee convergence, making their practical application difficult. It would therefore be interesting to develop a more robust method of extracting such mutually incoherent source-specific dictionaries to improve source separation and class discrimination when used in the sparse coding stage of the SCST method.

(5) **Use kernel methods to perform sparse coding in a higher dimensional space where signal and interference components are linearly separable.**

The results in Section 6.6 demonstrated the ability of the sparse coding process used by the SCST method to mostly separate signal and interference components of an observation sequences, thus allowing for robust detection of signals whose signatures are corrupted by nuisance sources. However, the separation was typically not perfect, as a small amount of the signal components present in the original observation would often be associated with the interference atoms, and vice versa. Apart from designing mutually incoherent dictionaries, another possible solution to this issue is to use the kernel trick [94] to implicitly map the data into a higher dimensional space where the two components are linearly separable. More specifically, a multivariate kernel sparse representation framework [95] can be used to find the sparse representation of an observation using atoms in a high dimensional space, which essentially performs nonlinear separation of the different components in this observation. Successful application of this approach would not only improve signal estimation performance, but also provide better detection and classification performance for the SCST method, since sparse coefficients should represent signals to be detected more distinctly in this case.

(6) **Investigate the benefits of imposing a temporal sparsity and consistency constraint for the SCST method.**

The first step of the SCST process involves finding a sparse representation of a given vector in the sequence as it arrives, which is done independently of other observations. The overall structure of a transient event is then modeled by a Bayesian network that describes the temporal evolution of the extracted sparse vectors. It would therefore be interesting to additionally impose a sparsity constraint over time in order to ensure successive sparse vectors represent only the salient features of an entire event. While simultaneously imposing sparsity between and within observations has been studied before [96], there in no known research that deals with this problem in the context of transient detection and classification using sequential data.

(7) **Design an optimal multivariate data representation for source characterization.** As noted in Chapter 2, the 1/3 octave data format is adequate but not necessarily optimal for performing automated source characterization, as its use was originally motivated by data storage limitations and its utility for visual analysis by humans. Since automated approaches for analyzing the data now exist, it would be advantageous to design a data representation that is optimal for this task to deploy on future acoustical monitoring stations, thus improving the overall performance of the system. A simple example of the benefits of this work can be seen in Fig. 7.2, which shows two different representations of a time series containing the simulated signatures of two sources in motion that emit different narrowband frequencies. The bottom image shows the 1/3 octave representation of the data, where the signatures of the two sources overlap in frequency to a large degree. The top image shows a representation of the simulated data obtained by applying the Karhunen-Loeve transform [88], with basis vectors that were extracted from the time-domain signatures of each source type. This alternate representation clearly separates the signatures of the two sources into different areas of the feature space, thus allowing for improved discrimination. Although the approach used for this example is not practical, in that a relevant transform needs to provide a unique representation of all source types that may be present (rather than two simple narrowband sources), it nonetheless demonstrates that alternate data representations may exist that are more useful than 1/3 octave for source characterization.

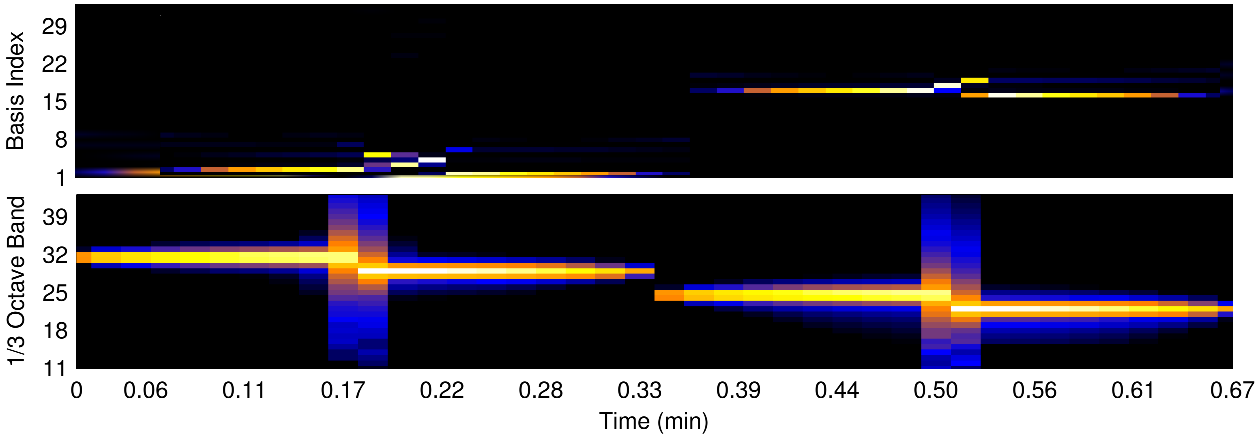


FIGURE 7.2. Comparison of Karhunen-Loeve (top) and 1/3 octave (bottom) representations of simulated signatures associated with 0.2 kHz and 1 kHz sources in motion.

(8) **Use models learned from one data set to perform source characterization on another.**

The results in Chapter 6 were based on scenarios where some training data segments from a given data set (site) were used to extract models and set parameters for a each method, which were in turn used to analyze a disjoint set of testing segments from the same site. While such an approach is still a realistic way to perform source characterization, and a much faster way to do so than manual analysis of the data by an operator, the proposed methods would have even greater utility if it were unnecessary to enact the training procedure for each site separately. Therefore, it would be interesting to evaluate the performance of each method when trained using only data from a single site, yet tested on data from multiple other sites. Clearly, the same sources of interest that each method is trained to recognize should be present in these testing sites in order to obtain valid results.

(9) **Incorporate a learning system to improve characterization of novel events and source types.**

Any source characterization approach mentioned in this thesis will only perform well if trained using data that is representative of testing sequences it is expected to analyze. This is the main reason applying a system trained on data from one site might not perform as well on data from another site, as mentioned above. It might therefore be worthwhile to develop an in-situ learning framework [94] for updating existing source models and forming new ones. The main challenge of this task is determining which subset of newly encountered data is suitable for updating a given model. This is because the performance of the system can quickly degrade if learned samples do not belong to the class associated with the updated model, for instance. On the other hand, updating the models would be a fairly simple process once relevant data has been identified.

(10) **Application of the proposed methods to other problems.**

As mentioned in Chapter 1, the methods developed in this thesis are relevant to a multitude of applications that use sequential multivariate data, e.g., speech recognition, habitat monitoring, medical diagnosis, and battlefield surveillance. This is particularly true of the SCST method, as it makes few assumptions concerning the structure of signals, interference, and noise present in the data, making it readily adaptable to other application areas. An example of such an alternate data sequence can be seen in Fig. 7.3, which shows a synthetic aperture sonar (SAS)-like image [97] generated using multi-channel sonar data. Such images represent the coherence between different pings off various spatial locations on the seafloor, where higher levels of coherence (red pixels) typically occur at locations where underwater objects are present. In [97], objects are detected by applying the matched subspace detector in Appendix D, to individual pings. It would be interesting to study the benefits of using the SCST method to perform characterization of signals/objects of interest, relative to the standard approaches used in a given field of study.

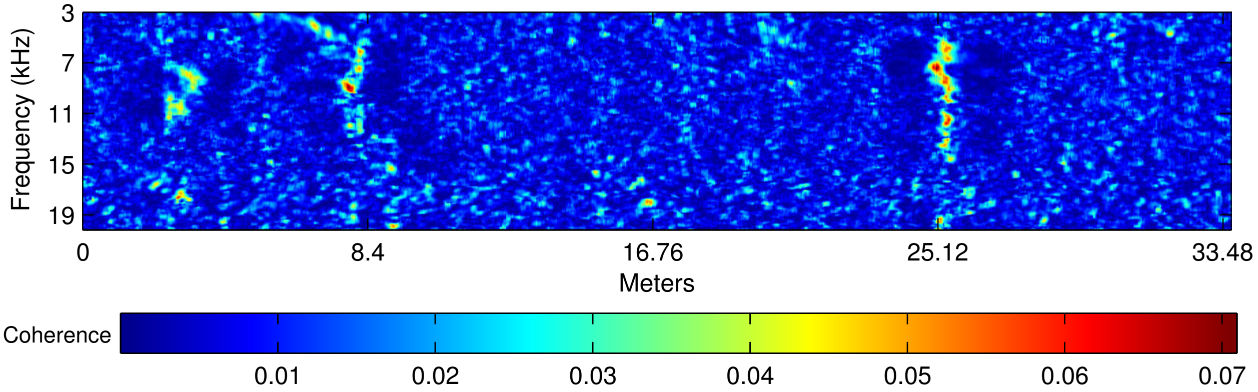


Figure 7.3. SAS-like data sequence that is suitable for analysis via the proposed source characterization methods.

Bibliography

[1] B. H. Juang and L. R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. on Acoustics, Speech, and Signal Process.*, vol. 33, no. 6, pp. 1404–1413, December 1985.

[2] B. Gajić and K. K. Paliwal, "Robust speech recognition in noisy environments based on subband spectral centroid histograms," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 2, pp. 600–608, March 2006.

[3] K. K. Paliwal, "Spectral subband centroid features for speech recognition," *Proc. of IEEE Int'l Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, vol. 2, pp. 617–620, May 1998.

[4] M. Wu, D. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 11, no. 3, pp. 229–241, May 2003.

[5] J. Bouvrie, T. Ezzat, and T. Poggio, "Localized spectro-temporal cepstral analysis of speech," *Proc. of IEEE Int'l Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, pp. 4733–4736, May 2008.

[6] E. A. Edmonds, L. Y. Pan, and S. M. O'Brien, "Automatic feature extraction from spectrograms for acoustic-phonetic analysis," *11th Int'l Conf. on Pattern Recognition (IAPR)*, pp. 701–704, September 1992.

[7] M. Ramona, G. Richard, and B. David, "Vocal detection in music with support vector machines," *Proc. of IEEE Int'l Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, pp. 1885–1888, April 2008.

[8] T. S. Brandes, "Feature vector selection and use with hidden Markov models to identify frequency-modulated bioacoustic signals amidst noise," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 16, no. 6, pp. 1173–1180, August 2008.

[9] S. E. Anderson, A. S. Dave, and D. Margoliash, "Template-based automatic recognition of birdsong syllables from continuous recordings," *J. of the Acoustical Society of America (JASA)*, vol. 100, no. 2, pp. 1209–1219, August 1996.

[10] Z. Chen and R. C. Maher, "Semi-automatic classification of bird vocalizations using spectral peak tracks," *J. of the Acoustical Society of America (JASA)*, vol. 120, no. 5, pp. 2974–2984, November 2006.

[11] E. D. Chesmore, "Application of time domain signal coding and artificial neural networks to passive acoustical identification of animals," *Applied Acoustics*, vol. 62, no. 12, pp. 1359–1374, December 2001.

[12] J. A. Kogan and D. Margoliash, "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study," *J. of the Acoustical Society of America (JASA)*, vol. 103, no. 4, pp. 2185–2196, April 1998.

[13] P. Somervuo, A. Härmä, and S. Fagerlund, "Parametric representation of bird sounds for automatic species recognition," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 6, pp. 2252–2263, November 2006.

[14] H. Wang, J. Elson, L. Girod, D. Estrin, and K. Yao, "Target classification and localization in habitat monitoring," *Proc. of IEEE Int'l Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, vol. 4, pp. 844–847, April 2003.

[15] S. Chu, S. Narayanan, and C. C. J. Kou, "Environmental sound recognition with time-frequency audio features," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 17, no. 6, pp. 1142–1158, August 2009.

[16] J. D. Krijnders, M. E. Niessen, and T. C. Andringa, "Sound event recognition through expectancy-based evaluation of signal-driven hypotheses," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1552–1559, August 2010.

[17] S. G. Lingala, Y. Hu, E. DiBella, and M. Jacob, "Accelerated dynamic MRI exploiting sparsity and low-rank structure: k-t SLR," *IEEE Trans. on Medical Imaging*, vol. 30, no. 5, pp. 1042–1054, May 2011.

[18] L. Owsley, L. Atlas, and C. Heinemann, "Use of modulation spectra for representation and classification of acoustic transients from sniper fire," *Proc. of IEEE Int'l Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, vol. 4, pp. 1129–1132, March 2004.

[19] D. E. Lake, "Harmonic phase coupling for battlefield acoustic target identification," *Proc. of IEEE Int'l Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, vol. 4, pp. 2049–2052, May 1998.

[20] H. Wu and J. Mendel, "Classification of battlefield ground vehicles using acoustic features and fuzzy logic rule-based classifiers," *IEEE Trans. on Fuzzy Systems*, vol. 15, no. 1, pp. 56–72, February 2007.

[21] M. R. Azimi-Sadjadi, A. Pezeshki, and N. Roseveare, "Wideband DOA estimation algorithms for multiple moving sources using unattended acoustic sensors," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 44, no. 4, pp. 1585–1599, October 2008.

[22] E. Lynch, D. Joyce, and K. Fristrup, "An assessment of noise audibility and sound levels in U.S. national parks," *Landscape Ecol.*, vol. 26, pp. 1297–1309, August 2011.

[23] S. Le Cam, C. Collet, and F. Salzenstein, "Detection of transient signals: local approach using a Markovian tree with frequency selectivity," *IEEE Int'l Workshop on Machine Learning for Signal Process.*, pp. 1–6, September 2009.

[24] D. E. Asraf and M. G. Gustafsson, "Detection of multiple transient signals with unknown arrival times," *IEEE Trans. on Inform. Theory*, vol. 51, no. 1856–1860, May 2005.

[25] E. Fishler and H. Messer, "Detection and parameter estimation of a transient signal using order statistics," *IEEE Trans. on Signal Process.*, vol. 48, no. 5, pp. 1455–1458, May 2000.

[26] S. Colonnese and G. Scarano, "Transient signal detection using higher order moments," *IEEE Trans. on Signal Process.*, vol. 47, no. 2, pp. 515–520, February 1999.

[27] Z. Wang and P. K. Willett, "All-purpose and plug-in power-law detectors for transient signals," *IEEE Trans. on Signal Process.*, vol. 49, no. 11, pp. 2454–2466, November 2001.

[28] C. Lei, J. Zhang, and Q. Gao, "Unknown and arbitrary sparse signal detection against background noise," *IEEE 10th Int'l Conf. on Signal Process. (ICSP)*, pp. 46–49, October 2010.

[29] M. I. Plett, "Transient detection with cross wavelet transforms and wavelet coherence," *IEEE Trans. on Signal Process.*, vol. 55, no. 5, pp. 1605–1611, May 2007.

[30] V. Bruni, S. Marconi, and D. Vitulano, "Time-scale atoms chains for transients detection in audio signals," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 3, pp. 420–433, March 2010.

[31] P. W. C. Han, B. Chen, and D. Abraham, "A detection optimal min-max test for transient signals," *IEEE Trans. on Inform. Theory*, vol. 44, no. 2, pp. 866–869, March 1998.

[32] C. Hory, N. Martin, and A. Chehikian, "Spectrogram segmentation by means of statistical features for non-stationary signal interpretation," *IEEE Trans. on Signal Process.*, vol. 50, no. 12, pp. 2915–2925, December 2002.

[33] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application.* Prentice-Hall, Inc. Englewood Cliffs, N.J., 1993.

[34] G. Lorden, "Procedures for reacting to a change in distribution," *Annals of Mathematical Statistics*, vol. 42, pp. 1897–1908, June 1971.

[35] B. Chen and P. Willet, "Detection of hidden Markov model transient signals," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 36, no. 4, pp. 1253–1268, October 2000.

[36] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[37] B. Chen and P. Willet, "Superimposed HMM transient detection via target tracking ideas," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 37, no. 3, pp. 946–956, July 2001.

[38] R. L. Streit and P. K. Willett, "Detection of random transient signals via hyperparameter estimation," *IEEE Trans. on Signal Process.*, vol. 47, no. 7, pp. 1823–1834, July 1999.

[39] D. Luengo, C. Pantaleon, I. Santamaria, L. Vielva, and J. Ibañez, "Multiple composite hypothesis testing: a competitive approach," *J. of VLSI Signal Process. Systems*, vol. 37, no. 2/3, pp. 319–331, June 2004.

[40] A. B. Salberg, A. Hanssen, and L. L. Scharf, "Robust multidimensional matched subspace classifiers based on weighted least-squares," *IEEE Trans. on Signal Process.*, vol. 55, no. 3, pp. 873–880, March 2007.

[41] J. Cartmill, N. Wachowski, and M. Azimi-Sadjadi, "Buried underwater object classification using a collaborative multiaspect classifier," *IEEE Journal of Oceanic Engr.*, vol. 34, no. 1, pp. 32–44, January 2009.

[42] N. Wachowski and M. R. Azimi-Sadjadi, "A likelihood-based decision feedback system for multi-aspect classification of underwater targets," *IEEE Int'l Joint Conf. on Neural Networks (IJCNN)*, pp. 3232–3239, June 2009.

[43] P. K. Varshney, *Distributed Detection and Data Fusion*, 1st ed. Springer-Verlag, Inc. New York, NY, 1997.

[44] L. L. Scharf and B. Friedlander, "Matched subspace detectors," *IEEE Trans. on Signal Process.*, vol. 42, no. 8, pp. 2146–2157, August 1994.

[45] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 3, no. 1, pp. 72–83, January 1995.

[46] R. A. Altes, "Detection, estimation, and classification with spectrograms," *J. of the Acoustical Society of America (JASA)*, vol. 67, no. 4, pp. 1232–1246, April 1980.

[47] M. Pollak and D. Siegmund, "Approximations to the expected sample size of certain sequential tests," *Annals Statistics*, vol. 3, no. 6, pp. 1267–1282, November 1975.

[48] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proc. of IEEE*, vol. 98, no. 6, pp. 1031–1044, June 2010.

[49] H. Zhang, N. M. Nasrabadi, T. S. Huang, and Y. Zhang, "Transient acoustic signal classification using joint sparse representation," *Proc. of IEEE Int'l Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, pp. 2220–2223, May 2011.

[50] J. M. K. Kua, E. Ambikairajah, J. Epps, and R. Togneri, "Speaker verification using sparse representation classification," *Proc. of IEEE Int'l Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, pp. 4548–4551, May 2011.

[51] S. Zubair and W. Wang, "Audio classification based on sparse coefficients," *Sensor Signal Processing for Defence*, pp. 1–5, September 2011.

[52] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Sparse representation for target detection in hyperspectral imagery," *IEEE J. of Selected Topics in Signal Process.*, vol. 5, no. 3, pp. 629–640, June 2011.

[53] M. Brown and N. P. Costen, "Exploratory basis pursuit classification," *Pattern Recognition Letters - Special issue: Artificial neural networks in pattern recognition*, vol. 26, no. 12, pp. 1907–1915, September 2005.

[54] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, no. 1, pp. 34–81, February 2009.

[55] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. on Signal Process.*, vol. 54, no. 11, pp. 4311–4322, November 2006.

[56] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. on Signal Process.*, vol. 46, no. 4, pp. 886–902, April 1998.

[57] L. Daudet, "Sparse and structured decompositions of signals with the molecular matching pursuit," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 5, pp. 1808–1816, September 2006.

[58] S. J. Godsill, A. T. Cemgil, C. Févotte, and P. J. Wolfe, "Bayesian computational methods for sparse audio and music processing," *15th European Signal Process. Conf.*, pp. 345–349, September 2007.

[59] C. Févotte and S. Godsill, "Sparse linear regression in unions of bases via Bayesian variable selection," *IEEE Signal Process. Letters*, vol. 13, no. 7, pp. 441–444, July 2006.

[60] D. Oldoni, B. De Coensel, M. Rademaker, B. De Baets, and D. Botteldooren, "Context-dependent environmental sound monitoring using SOM coupled with LEGION," *Proc. of Int'l Joint Conf. on Neural Networks (IJCNN)*, pp. 1–8, July 2010.

[61] J. M. Adler, B. D. Rao, and K. Kreutz-Delgado, "Comparison of basis selection methods," *Thirtieth Asilomar Conf. on Signals, Systems and Computers*, vol. 1, pp. 252–257, November 1996.

[62] K. Veggeberg, "Octave analysis explored," *Evaluation Engineering*, pp. 40–43, August 2008.

[63] *Larson Davis Model 831 Technical Reference Manual*, PCB Piezotronics.

[64] Acoustical Society of America, "Specification for octave-band and fractional-octave-band analog and digital filters," *ANSI S1.11-2004*, 2004.

[65] B. Friedlander and B. Porat, "Performance analysis of transient detectors based on a class of linear data transforms," *IEEE Trans. on Inform. Theory*, vol. 38, no. 2, pp. 665–673, March 1992.

[66] Acoustical Society of America, "Specifications for integrating-averaging sound level meters," *ANSI S1.43-1997*, 1997.

[67] R. A. Davis, T. C. M. Lee, and G. A. Rodriguez-Yam, "Structural break estimation for nonstationary time series models," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 223–239, March 2006.

[68] A. Wald and J. Wolfowitz, "Optimum character of the sequential probability ratio test," *Annals of Mathematical Statistics*, vol. 19, pp. 326–339, September 1948.

[69] E. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, pp. 100–115, January 1954.

[70] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, pp. 289–337, 1933.

[71] P. R. Runkle, B. K. Bharadwaj, L. Couchman, and L. Carin, "Hidden Markov models for multi-aspect target classification," *IEEE Trans. on Signal Process.*, vol. 47, no. 7, pp. 2035–2040, July 1999.

[72] M. Robinson, M. R. Azimi-Sadjadi, and J. Salazar, "Multi-aspect target discrimination using hidden Markov models and neural networks," *IEEE Trans. on Neural Networks*, vol. 16, no. 2, pp. 447–459, March 2005.

[73] A. G. Tartakovskii, "Sequential testing of many simple hypotheses with independent observations," *Probl. Inform. Transm.*, vol. 24, no. 4, pp. 299–309, 1989.

[74] N. Wachowski and M. Azimi-Sadjadi, "Characterization of multiple transient acoustical sources from time-transform representations," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 21, no. 9, pp. 1966–1978, September 2013.

[75] A. P. Sage and J. L. Melsa, *Estimation Theory with Applications to Communications and Control.* McGraw-Hill, 1971.

[76] A. Ben-Israel and T. Greville, *Generalized Inverses: Theory and Applications*, 2nd ed. Springer, 2003.

[77] R. T. Behrens and L. L. Scharf, "Signal processing applications of oblique projection operators," *IEEE Trans. on Signal Process.*, vol. 42, no. 6, pp. 1413–1424, June 1994.

[78] D. E. Holmes and L. C. Jain, *Innovations in Bayesian Networks*, 1st ed. Springer Berlin Heidelberg, 2008.

[79] D. L. Donoho and G. Kutyniok, "Analysis of ℓ_1 minimization in the geometric separation problem," *42nd Annual Conf. Inform. Sciences and Systems*, pp. 274–279, March 2008.

[80] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. on Inform. Theory*, vol. 52, no. 1, pp. 6–18, January 2006.

[81] D. Barchiesi and M. D. Plumbley, "Learning incoherent dictionaries for sparse approximation using iterative projections and rotations," *IEEE Trans. on Signal Process.*, vol. 61, no. 8, pp. 2055–2065, April 2013.

[82] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," *IEEE Int'l Conf. on Computer Vision*, pp. 543–550, November 2011.

[83] H. V. Poor and J. B. Thomas, "Applications of Ali-Silvey distance measures in the design of generalized quantizers for binary decision systems," *IEEE Trans. on Communications*, vol. 25, no. 9, pp. 893–900, September 1977.

[84] J. N. Tsitsiklis, "Extremal properties of likelihood-ratio quantizers," *IEEE Trans. on Communications*, vol. 41, no. 4, pp. 550–558, April 1993.

[85] H. Kobayashi and J. B. Thomas, "Distance measures and related criteria," *Proc. Fifth Annual Allerton Conf. Circuit and System Theory*, pp. 491–500, October 1967.

[86] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *J. Royal Stat. Soc. Series B*, vol. 28, no. 1, pp. 131–142, April 1966.

[87] A. C. Davison, *Statistical Models.* Cambridge University Press, 2003.

[88] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: Johns Hopkins, 1996.

[89] M. R. Azimi-Sadjadi, "New results in strip Kalman filtering," *IEEE Trans. on Circuits and Systems*, vol. 36, no. 6, pp. 893–897, June 1989.

[90] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[91] P. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Computational and Applied Mathematics*, vol. 20, no. 1, pp. 53–65, November 1987.

[92] J. Wang, S. Kwon, and B. Shim, "Generalized orthogonal matching pursuit," *IEEE Trans. on Signal Process.*, vol. 60, no. 12, pp. 6202–6216, December 2012.

[93] D. Pham and J. Cardoso, "Blind separation of instantaneous mixtures of nonstationary sources," *IEEE Trans. on Signal Process.*, vol. 49, no. 9, pp. 1837–1848, September 2011.

[94] M. R. Azimi-Sadjadi, J. Salazar, and S. Srinivasan, "An adaptable image retrieval system with relevance feedback using kernel machines and selective sampling," *IEEE Trans. on Image Process.*, vol. 18, no. 7, pp. 1645–1659, July 2009.

[95] N. H. Nguyen, N. M. Nasrabadi, and T. D. Tran, "Multi-sensor joint kernel sparse representation for personnel detection," *Proc. of the 20th European Signal Process. Conf. (EUSIPCO)*, pp. 739–743, August 2012.

[96] ——, "Robust multi-sensor classification via joint sparse representation," *Proc. of the 14th Int'l Conf. on Info. Fusion*, pp. 1–8, July 2011.

[97] N. Wachowski and M. R. Azimi-Sadjadi, "A new synthetic aperture sonar processing method using coherence analysis," *IEEE Journal of Oceanic Engr.*, vol. 36, no. 4, pp. 665–678, October 2011.

[98] A. Spanias, T. Painter, and A. Venkatraman, *Audio signal processing and coding.* John Wiley & Sons, 2006.

[99] A. M. Mathai and S. B. Provost, *Quadratic forms in random variables: theory and applications.* Marcel Dekker Incorporated, 1992.

[100] M. S. Bartlett and D. G. Kendall, "The statistical analysis of variance-heterogeneity and the logarithmic transformation," *Supplement to the journal of the royal statistical society*, vol. 8, no. 1, pp. 128–138, 1946.

[101] H. Anton, *Calculus: A New Horizon, 6th ed.* New York: Wiley, 1999.

[102] G. J. Székely and M. L. Rizzo, "A new test for multivariate normality," *Journal of Multivariate Analysi*, vol. 93, no. 1, pp. 58–80, March 2005.

[103] R. Fisher, *Statistical methods for research workers.* Edinburgh: Oliver & Boyd, 1925.

[104] D. Ramírez, J. Vía, I. Santamaría, and L. L. Scharf, "Locally most powerful invariant tests for correlation and sphericity of Gaussian vectors," *IEEE Trans. on Inform. Theory*, vol. 59, no. 4, pp. 2128–2141, April 2013.

[105] L. L. Scharf and C. T. Mullis, "Canonical coordinates and the geometry of inference, rate, and capacity," *IEEE Trans. on Signal Process.*, vol. 48, no. 3, pp. 824–831, March 2000.

[106] A. Brodzik, "On the Fourier transform of finite chirps," *IEEE Signal Process. Letters*, vol. 13, no. 9, pp. 541–544, September 2006.

[107] S. Kraut, L. L. Scharf, and L. T. McWhorter, "Adaptive subspace detectors," *IEEE Trans. on Signal Process.*, vol. 49, no. 1, pp. 1–15, January 2001.

APPENDIX A

Explicit Form of a One-Third Octave Vector

In this Appendix, the explicit form of a 1/3 octave vector [62] is given, which helps in understanding the characteristics of various source signatures present in the data in Chapter 2, and serves as a precursor to deriving the distribution of ambient noise in the 1/3 octave domain in Appendix B. Reasons for using the 1/3 octave representation in this thesis are discussed in Section 2.2. The derivations herein use the standards specified in [64] and [66] for fractional-octave-band filters and integrating sound level meters, respectively, as guidelines.

Though 1/3 octave vectors may be extracted from either continuous or sampled data, here it is assumed that the original data captured by the monitoring station is an audio waveform sampled at $M = 51,200$ Hz [63], and hence, may be represented by a time series vector $\mathbf{u} = [u[0] \ \cdots \ u[m-1]]^T \in \mathbb{R}^m$. In this Appendix, there is no prior assumption placed on the composition or structure of $\mathbf{u}$, meaning it can potentially contain the signatures of any number of acoustical sources plus ambient noise. Since the soundscape is being continually recorded, the length m of $\mathbf{u}$ is increasing. The elements of the 1/3 octave vector $\mathbf{y}_k = [y_k[0] \ \cdots \ y_k[N-1]]^T \in \mathbb{R}^N$ represent the average energy in $N = 33$ different 1/3 octave frequency bands over the interval of $\mathbf{u}$ corresponding to the kth second of the recorded audio waveform, i.e., $[u[kM] \ \cdots \ u[(k+1)M-1]]$.

The nth 1/3 octave frequency band has a center frequency of $f_c^{(n)} = 10^3 \cdot 2^{\left(\frac{n}{3}-10\right)}$, meaning $f_c^{(n+3)} = 2f_c^{(n)}$, i.e., the center frequency doubles (is one octave higher) every third band, hence the name 1/3 octave. The upper and lower cutoff frequencies for the nth band are given by $f_u^{(n)} = 2^{\frac{1}{6}} f_c^{(n)}$ and $f_l^{(n)} = 2^{-\frac{1}{6}} f_c^{(n)}$, respectively. For the data used in this thesis $n \in [11, 43]$, meaning each $\mathbf{y}_k$ captures energy in the frequency range of 12.4 Hz – 20.2 kHz.

Suppose $\mathbf{u} = [u[0] \ \cdots \ u[kM-1]] \in \mathbb{R}^{kM}$, meaning the kth second of data has just been recorded. Since the order of operations used to extract a 1/3 octave vector is not standardized, the first step used here is windowing in the time domain. Denote $\mathbf{u}_k = \mathbf{T}_k\mathbf{u} \in \mathbb{R}^M$ as the vector containing the samples from the kth (last) second of $\mathbf{u}$, where

$$\mathbf{T}_k = \left[\mathbf{0}_{M\times(k-1)M}\ \mathbf{I}_M\right] \in \mathbb{R}^{M\times kM}.$$

To find $y_k[n]$, $\mathbf{u}_k$ is bandpass filtered with a passband that corresponds to the nth 1/3 octave frequency band, within certain tolerances [64]. For ease of derivations in Appendix B, and to allow the use of IIR filters, this filtering operation is shown in the frequency domain. Therefore, denote $\mathbf{x}_k = \mathbf{D}_M\mathbf{u}_k$ as the Fourier representation of $\mathbf{u}_k$, where $\mathbf{D}_M = \frac{1}{\sqrt{M}}\left[w_M^{ij}\right]_{i,j=0,\ldots,M-1}$ is the $M \times M$ DFT matrix with $w_M = e^{-j2\pi/M}$. Filtering can then be performed using the matrix whose diagonal elements are the samples of the M-point frequency response of the bandpass filter associated with the nth 1/3 octave frequency band, given by

$$\mathbf{G}_n = diag\left[g_n[0]\ g_n[1]\ \cdots\ g_n[M-1]\right]. \tag{A.1}$$

The sound pressure level, in decibels (dB), in the nth 1/3 octave frequency band can then be written as [66]

$$\begin{aligned} y_k[n] &= 10\log\left\{\frac{1}{p_0^2 M}\mathbf{u}^T\mathbf{T}_k^T\mathbf{D}_M^H\mathbf{G}_n^H\mathbf{G}_n\mathbf{D}_M\mathbf{T}_k\mathbf{u}\right\} \\ &= 10\log\left\{\frac{1}{p_0^2 M}\left\|\mathbf{G}_n\mathbf{x}_k\right\|_2^2\right\} \end{aligned} \tag{A.2}$$

where $p_0 = 20\ \mu\text{Pa}$ is the reference sound pressure and $\|\mathbf{x}\|_2 = \mathbf{x}^H\mathbf{x}$ is the Hermitian inner product.

APPENDIX B

Null Distribution of One-Third Octave Data

This Appendix derives the theoretical distribution, and analyzes the experimental distribution of ambient noise[1] that is present in recorded audio waveforms, in the 1/3 octave domain. The intent is to justify certain assumptions made about the prior distribution of the noise, which are often necessary when developing robust detection and classification methods (see Chapter 4). This material builds on the derivations in Appendix A showing the explicit form of 1/3 octave vectors in terms of the original time series data, and additionally uses the standards specified in [64] and [66] for fractional-octave-band filters and integrating sound level meters, respectively. Note that, since these standards allow for variations in certain processes to accommodate flexible design goals, assumptions must occasionally be made, e.g., concerning the filter type used.

B.1. Theoretical Noise Distribution

This section derives the distribution of the nth element $y_k[n]$ of the 1/3 octave vector $\mathbf{y}_k$ when the original audio waveform follows the null distribution, i.e., $\mathbf{u} = \mathbf{v} = [v[0] \ \cdots \ v[m-1]]^T \in \mathbb{R}^m$ (see Appendix A), where $\mathbf{v} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}_m\right)$ is a white Gaussian noise sequence. In this case, it follows that $\mathbf{y}_k = \mathbf{w}_k$, where $\mathbf{w}_k = [w_k[0] \ \cdots \ w_k[N-1]]^T$ is the associated 1/3 octave noise vector. The assumptions concerning the distribution of $\mathbf{v}$ are typical for audio processing applications in the time domain [98], and are the aggregate result of many factors including thermal noise and several phenomena that cause subtle variations in the sound pressure levels of natural acoustical environments.

Since $w_k[n]$ represents sound pressure level in decibels (dB), define $z_k[n] = 10^{\frac{w_k[n]}{10}}$ as the nth element of $\mathbf{w}_k$ with the dB transformation reversed. The effects of this transformation on the

[1] Note: ambient noise excludes interference sources, as discussed in Chapter 2.

distribution of noise will be discussed later in this Appendix. Note that (A.2) can be used to write

$$z_k[n] = \frac{1}{p_0^2 M}\mathbf{v}^T\mathbf{T}_k^T\mathbf{D}_M^H\mathbf{G}_n^H\mathbf{G}_n\mathbf{D}_M\mathbf{T}_k\mathbf{v}. \tag{B.1}$$

where H is the Hermitian operator. Since $\mathbf{T}_k$ simply performs rectangular windowing, it follows that $\mathbf{v}_k = \mathbf{T}_k\mathbf{v} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}_M\right)$. Therefore, in general $z_k[n]$ is generated using a quadratic form of the Gaussian random vector $\mathbf{v}_k$, with deterministic matrix $\mathbf{D}_M^H\mathbf{G}_n^H\mathbf{G}_n\mathbf{D}_M$, making $z_k[n]$ generalized central chi-squared distributed [99].

For simplicity assume $\mathbf{G}_n$ from (A.1) represents an ideal filter for the nth 1/3 octave band, i.e.,

$$\mathbf{G}_n = diag\left[\mathbf{0}_{1\times f_l^{(n)}}\ \mathbf{1}_{1\times(f_u^{(n)}-f_l^{(n)})}\ \mathbf{0}_{1\times(M-f_u^{(n)})}\right]^T \tag{B.2}$$

is a matrix of zeros apart from a series of ones on the diagonal at locations corresponding to integer cutoff frequencies $f_l^{(n)}+1$ through $f_u^{(n)}$. Note that this type of filter satisfies the specifications in [64] for 1/3 octave bandpass filter attenuation characteristics. In this case $\mathbf{D}_M^H\mathbf{G}_n^H\mathbf{G}_n\mathbf{D}_M = \mathbf{P}_{\mathbf{D}_M^n}$ is a projection matrix for the subspace spanned by the DFT basis functions corresponding to the filter cutoff frequencies $f_l^{(n)}+1$ through $f_u^{(n)}$, i.e., the column space of $\mathbf{D}_M^n$, which contains columns $f_l^{(n)}+1$ through $f_u^{(n)}$ of $\mathbf{D}_M$. Consequently, it can be shown [99] that, since $\mathbf{P}_{\mathbf{D}_M^n}$ is symmetric and idempotent, $\mathbf{v}_k^H\mathbf{P}_{\mathbf{D}_M^n}\mathbf{v}_k \sim \chi_\nu^2$, i.e., central chi-squared distributed with $\nu = \text{tr}(\mathbf{P}_{\mathbf{D}_M^n})$ degrees of freedom. Note that $\text{tr}(\mathbf{P}_{\mathbf{D}_M^n}) = f_u^{(n)} - f_l^{(n)}$ since $\mathbf{P}_{\mathbf{D}_M^n}$ performs projection onto a space spanned by $f_u^{(n)} - f_l^{(n)}$ basis vectors.

It is well-known that, for large ν, the χ_ν^2 distribution asymptotically converges to $\mathcal{N}(\nu, 2\nu)$ [100]. Thus, for moderately large n, where the bandwidth is large, it is reasonable to assume that

$$z_k[n] \sim \mathcal{N}\left(\frac{\nu}{p_0^2 M}, \frac{2\nu}{p_0^4 M^2}\right). \tag{B.3}$$

Moreover, since $w_k[n] = 10\log(z_k[n])$, the distribution of $w_k[n]$ converges to Gaussian much faster than the distribution of $z_k[n]$ [100], as the logarithm removes much of the asymmetry. The log transformation is also beneficial here since $w_k[n] \leq 0$ for $0 < z_k[n] < 1$, meaning the resulting distribution of the transformed values is two-sided, as with a Gaussian distribution. Note that a similar conclusion concerning the distribution of $w_k[n]$ can be reached for continuous time audio using Riemann sums [101]. This means assuming ambient noise is Gaussian in the 1/3 octave domain is still reasonable even when it is known that $\mathbf{y}_k$ are extracted from continuous waveforms.

According to the order of operations used in (B.1), where windowing is performed before filtering in the frequency domain, $\mathbf{w}_k$ for different k are formed using disjoint sets of time-domain samples $\mathbf{T}_k\mathbf{v}$, meaning it can also be assumed that $w_k[n]$ are independent for different k. On the other hand, since the same $\mathbf{v}_k$ is used to generate $w_k[n]$, $\forall n$, the noise vector $\mathbf{w}_k$ is colored (element-wise) with a non-diagonal full rank covariance matrix $\mathbf{R}_\mathbf{w}$. Justification for $\mathbf{R}_\mathbf{w}$ being full rank comes from the fact that $\mathbf{G}_n$ performs windowing on disjoint 1/3 octave frequency bands, and hence, the subspaces characterized by $\mathbf{P}_{\mathbf{D}_M^n}$'s for different n are orthogonal. Unfortunately, it is difficult to draw further conclusions on the distribution of the entire vector $\mathbf{w}_k$, mainly owing to the log transformation in (A.2). For this reason, the next section presents an evaluation of the experimental distribution of $\mathbf{w}_k$ using the KEFJ004 data from Chapter 2.

B.2. Experimental Noise Distribution

To establish more concrete analytical claims concerning the distribution of $\mathbf{w}_k$, this section provides experimental validation of the assumed distribution of $\mathbf{w}_k$, i.e., $\mathbf{w}_k \overset{\text{IID}}{\sim} \mathcal{N}(\boldsymbol{\mu}_\mathbf{w}, \mathbf{R}_\mathbf{w})$, where $\boldsymbol{\mu}_\mathbf{w}$ is the noise mean vector and $\mathbf{R}_\mathbf{w}$ is the full rank noise covariance matrix. Note that $\boldsymbol{\mu}_\mathbf{w} \neq \mathbf{0}$ since $\mathbf{w}_k$ represents the average energy in different frequency bands. Here, we begin by testing the Gaussianity of $\mathbf{w}_k$, which is followed by a test for independence between $\mathbf{w}_k$'s. The latter is necessary since the above derivations assume a specific order of operations used in (B.1), as well as white Gaussian noise in the original time domain.

B.2.1. Testing for Gaussianity

The goal here is to perform the following hypothesis test

$$\begin{aligned} \mathcal{H}_0 &: \mathbf{w}_k \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{w}}, \mathbf{R}_{\mathbf{w}}\right) \\ \mathcal{H}_1 &: \mathbf{w}_k \not\sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{w}}, \mathbf{R}_{\mathbf{w}}\right) \end{aligned} \tag{B.4}$$

i.e. $\mathcal{H}_0$ means the sample distribution of $\mathbf{w}_k$ is multivariate Gaussian, as assumed, while $\mathcal{H}_1$ means the assumed distribution is not a good fit to $\mathbf{w}_k$. This is accomplished using an energy distance goodness-of-fit measure [102]. Testing for independence of $\mathbf{w}_k$'s is done separately afterward. Given arbitrary random vectors $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^N$, with known distributions, the energy distance between their associated distributions is given by

$$h(\mathbf{x}, \mathbf{y}) = 2E \left\|\mathbf{x} - \mathbf{y}\right\|_2 - E \left\|\mathbf{x} - \mathbf{x}'\right\|_2 - E \left\|\mathbf{y} - \mathbf{y}'\right\|_2 \geq 0 \tag{B.5}$$

where $\mathbf{x}'$ and $\mathbf{y}'$ are random variables that are independent of and identically distributed to $\mathbf{x}$ and $\mathbf{y}$, respectively. It follows that $\mathbf{x}$ and $\mathbf{y}$ are identically distributed if and only if $h(\mathbf{x}, \mathbf{y}) = 0$. This test is affine invariant, consistent, and simple to implement. Clearly, such a test is unnecessary if the distributions of $\mathbf{x}$ and $\mathbf{y}$ are already known, but (B.5) serves as a baseline for constructing an appropriate sample version of an energy distance goodness-of-fit measure.

Since the hypothesized distribution is $\mathbf{w}_k \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{w}}, \mathbf{R}_{\mathbf{w}}\right)$, consider a set of realizations of this vector, denoted by $\{\mathbf{w}_k^{(j)}\}_{j=1}^J$, as well as the corresponding set of normalized 1/3 octave noise vectors $\{\boldsymbol{\omega}_k^{(j)}\}_{j=1}^J$ where $\boldsymbol{\omega}_k^{(j)} = \mathbf{R}_{\mathbf{w}}^{-1/2}\left(\mathbf{w}_k^{(j)} - \boldsymbol{\mu}_{\mathbf{w}}\right)$. Lastly, consider the set of independent realizations $\{\mathbf{n}^{(l)}\}_{l=1}^L$ of the random vector $\mathbf{n} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_N\right)$. The principles of (B.5) may then be used to test the sample distribution of $\{\mathbf{w}_k^{(j)}\}_{j=1}^J$ as [102]

$$\mathcal{E}(\{\boldsymbol{\omega}_k^{(j)}\}_j, \{\mathbf{n}^{(l)}\}_l) \tag{B.6}$$
$$= \frac{2}{JL}\sum_j \sum_l \left\|\boldsymbol{\omega}_k^{(j)} - \mathbf{n}^{(l)}\right\|_2 - \frac{1}{J^2}\sum_j \sum_{j'} \left\|\boldsymbol{\omega}_k^{(j)} - \boldsymbol{\omega}_k^{(j')}\right\|_2 - \frac{1}{L^2}\sum_l \sum_{l'} \left\|\mathbf{n}^{(l)} - \mathbf{n}^{(l')}\right\|_2 .$$

As with many applications, the parameters of the hypothesized distribution $\boldsymbol{\mu}_\mathbf{w}$ and $\mathbf{R}_\mathbf{w}$ are unknown, and hence, their corresponding estimates

$$\hat{\boldsymbol{\mu}}_\mathbf{w} = \frac{1}{J}\sum_j \mathbf{w}_k^{(j)}$$
$$\hat{\mathbf{R}}_\mathbf{w} = \frac{1}{J}\sum_j (\mathbf{w}_k^{(j)} - \hat{\boldsymbol{\mu}}_\mathbf{w})(\mathbf{w}_k^{(j)} - \hat{\boldsymbol{\mu}}_\mathbf{w})^T$$

must be used instead. Defining the normalized vector $\hat{\boldsymbol{\omega}}_k^{(j)} = \hat{\mathbf{R}}_\mathbf{w}^{-1/2}\left(\mathbf{w}_k^{(j)} - \hat{\boldsymbol{\mu}}_\mathbf{w}\right)$, the set $\{\hat{\boldsymbol{\omega}}_k^{(j)}\}_j$ can be generated that is ultimately used to determine goodness of fit according to [102]

$$\mathcal{E}(\{\hat{\boldsymbol{\omega}}_k^{(j)}\}_j, \{\mathbf{n}^{(l)}\}_l) \quad \underset{\text{accept } \mathcal{H}_0}{\overset{\text{accept } \mathcal{H}_1}{\underset{<}{\geq}}} \quad \eta \tag{B.7}$$

where η is a predetermined threshold that may be based on empirical percentiles of $\mathcal{E}$ estimated by simulations using two independent sequences that are $\overset{\text{IID}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$. The test in (B.7) rejects multivariate Gaussianity of $\mathbf{w}_k$ (hypothesis $\mathcal{H}_0$) for large values of $\mathcal{E}$.

To test the Gaussianity of $\mathbf{w}_k$, the energy distance test in (B.7) was applied to a set of $J = 14,960$ (just over eight hours worth) 1/3 octave observation vectors containing noise alone from the KEFJ004 data set. As mentioned in Chapter 2, noise in this data set is caused by light wind, water flow, sensor noise, and other phenomena that is continually present and mostly random. Therefore,

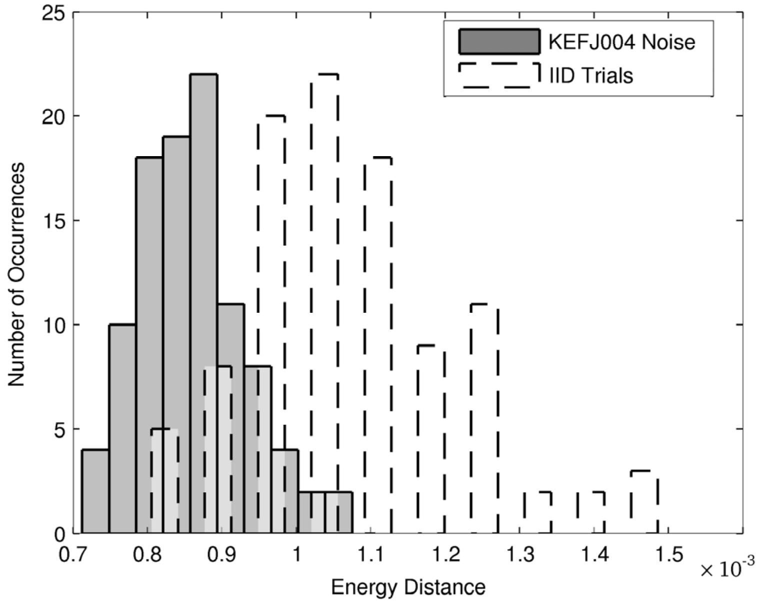


FIGURE B.1. Histograms of the energy distances between 1) two different sets of IID standard normal multivariate vectors $\{\mathbf{n}^{(l)}\}_l$ and $\{\mathbf{n}^{(l')}\}_{l'}$ (white) and 2) $\{\mathbf{n}^{(l)}\}_l$ and $\{\hat{\boldsymbol{\omega}}_k^{(j)}\}_j$ (shaded).

it may not be reasonable to assume the corresponding audio waveforms are pure white Gaussian noise, though this is part of the reason experimental validations are performed. The test was implemented using 100 trials, where each used (B.6) to measure the energy distance between the normalized KEFJ004 noise vectors $\{\hat{\boldsymbol{\omega}}_k^{(j)}\}_j$ and a different size J set (i.e. $L = J$ in (B.6) for these experiments) of vectors $\{\mathbf{n}^{(l)}\}_l$ that were generated using the `'mvnrnd'` command in MATLAB (i.e., pseudo-random), and are therefore approximately $\overset{\text{IID}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$. The shaded histogram in Fig. B.1 shows of these energy distance measures along with a white (dashed-line) histogram that shows the results of a similar set of trials measuring the distance between two independent size J sets of $\overset{\text{IID}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$ vectors. The latter results are used as a benchmark to determine if accepting $\mathcal{H}_0$ (declaring $\mathbf{w}_k \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{w}, \mathbf{R}_\mathbf{w})$) is reasonable. In particular, it is common to use a 0.05 significance level for such tests [103], meaning η may be set as the energy distance measure that divides the density of the measures $\mathcal{E}(\{\mathbf{n}^{(l)}\}_j, \{\mathbf{n}^{(l)}\}_l)$ into its lower 95% and upper 5% portions. Based on this criterion, a threshold of $\eta = 0.0013$ is used to determine whether or not to reject $\mathcal{H}_0$.

As can be seen from the shaded histogram in Fig. B.1, $\mathcal{E}(\{\hat{\boldsymbol{\omega}}_k^{(j)}\}_j, \{\mathbf{n}^{(l)}\}_l)$ is close to zero for every trial. Furthermore, the maximum value of $\mathcal{E}(\{\hat{\boldsymbol{\omega}}_k^{(j)}\}_j, \{\mathbf{n}^{(l)}\}_l)$ is far below the value of η specified by the histogram of $\mathcal{E}(\{\mathbf{n}^{(j)}\}_j), \{\mathbf{n}^{(l)}\}_l)$. Therefore, it is reasonable to accept $\mathcal{H}_0$ in (B.4) and assume that $\mathbf{w}_k \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{w}}, \mathbf{R}_{\mathbf{w}})$. However, the claim that $\mathbf{w}_k$'s are independent $\forall k$ still needs to be validated, which is investigated next.

B.2.2. Testing for Independence

The goal is now to validate the assumption that $\mathbf{w}_k, \forall k$ are IID using the locally most powerful invariant test for correlation of Gaussian vectors. It has recently been shown [104] that the corresponding test statistic is the Frobenius norm of the coherence matrix of a vector formed by concatenating all the vectors in the set to be tested. This test is often used for signal detection using multi-channel data where the null hypothesis states that observations from different sensors are independent, but can be applied to the data in Chapter 2 by treating temporally adjacent vectors as different data channels to test for independence of $\mathbf{w}_k$'s over k. Clearly, the independence and the above Gaussianity conditions must both be satisfied in order to claim that $\mathbf{w}_k \overset{\text{IID}}{\sim} \mathcal{N}(\boldsymbol{\mu}_{\mathbf{w}}, \mathbf{R}_{\mathbf{w}})$, as originally hypothesized.

Consider the set of random vectors $\{\tilde{\mathbf{w}}_k\}_{k=1}^{K}$ where it assumed that each $\tilde{\mathbf{w}}_k = \mathbf{w}_k - \boldsymbol{\mu}_{\mathbf{w}} \in \mathbb{R}^N$ is zero-mean multivariate Gaussian distributed, and define the vector $\tilde{\mathbf{w}} = \left[\tilde{\mathbf{w}}_1^T, \; \ldots, \; \tilde{\mathbf{w}}_K^T\right]^T \in \mathbb{R}^{NK}$ with covariance matrix $\mathbf{R}_{\tilde{\mathbf{w}}} = E\left[\tilde{\mathbf{w}}\tilde{\mathbf{w}}^T\right]$. Since $\tilde{\mathbf{w}}_k$ is zero-mean, and the mean and covariance matrix completely parameterize a multivariate Gaussian distribution, the sufficient statistic for testing independence of $\tilde{\mathbf{w}}_k$'s is the composite sample covariance matrix of $\tilde{\mathbf{w}}$

$$\hat{\mathbf{R}}_{\tilde{\mathbf{w}}} = \frac{1}{M} \sum_{m=1}^{M} \tilde{\mathbf{w}}^{(m)} \tilde{\mathbf{w}}^{(m)T} \tag{B.8}$$

where $\tilde{\mathbf{w}}^{(m)}$ is the concatenation of the mth measurement of each vector in the set $\{\tilde{\mathbf{w}}_k\}_{k=1}^{K}$. Defining $\hat{\mathbf{D}} = \text{diag}_N(\hat{\mathbf{R}}_{\tilde{\mathbf{w}}})$ as the symmetric block-diagonal matrix formed using the $N \times N$ matrix blocks on the diagonal of $\hat{\mathbf{R}}_{\tilde{\mathbf{w}}}$, the coherence matrix can be written as [104]

$$\hat{\mathbf{C}} = \hat{\mathbf{D}}^{-1/2}\hat{\mathbf{R}}_{\tilde{\mathbf{w}}}\hat{\mathbf{D}}^{-1/2}.$$

Typically, a coherence matrix denotes the covariance between two random vectors that have been whitened [105], which transforms each vector to remove the inter-element correlations. Here, the $N \times N$ blocks on the diagonal of $\hat{\mathbf{C}}$ are $\mathbf{I}_N$, while the off-diagonal blocks capture the coherence between different $\tilde{\mathbf{w}}_k$'s. The relevant hypothesis test in this case is then [104]

$$\begin{aligned} &\mathcal{H}_0 : \tilde{\mathbf{w}} \sim \mathcal{N}\left(\mathbf{0}_{NK}, \mathbf{D}\right) \\ &\mathcal{H}_1 : \tilde{\mathbf{w}} \sim \mathcal{N}\left(\mathbf{0}_{NK}, \mathbf{R}_1\right) \end{aligned} \tag{B.9}$$

where $\mathcal{H}_0$ and $\mathcal{H}_1$ correspond to accepting and rejecting the notion of $\tilde{\mathbf{w}}_k$'s being IID, respectively, and where $\mathbf{R}_1$ is an unknown covariance matrix under $\mathcal{H}_1$. This test is conducted according to [104]

$$\left\|\hat{\mathbf{C}}\right\|_F \begin{array}{c} \text{accept } \mathcal{H}_1 \\ \geq \\ < \\ \text{accept } \mathcal{H}_0 \end{array} \gamma \tag{B.10}$$

where γ is a predetermined threshold that may be based on empirical percentiles of the Frobenius norm $\left\|\hat{\mathbf{C}}\right\|_F$, estimated by simulations using independent sets of vectors that are $\overset{\text{IID}}{\sim} \mathcal{N}\left(\mathbf{0}, \mathbf{I}_N\right)$. The idea behind this test is that $\left\|\hat{\mathbf{C}}\right\|_F$ in (B.10) becomes larger as the off-diagonal elements of $\hat{\mathbf{C}}$ deviate from zero, which is the value of these elements for a set of truly IID vectors.

To test the independence of $\mathbf{w}_k$'s for the KEFJ004 data in Chapter 2, the coherence test in (B.10) was applied to the same set of 1/3 octave noise vectors that was used to test Gaussianity

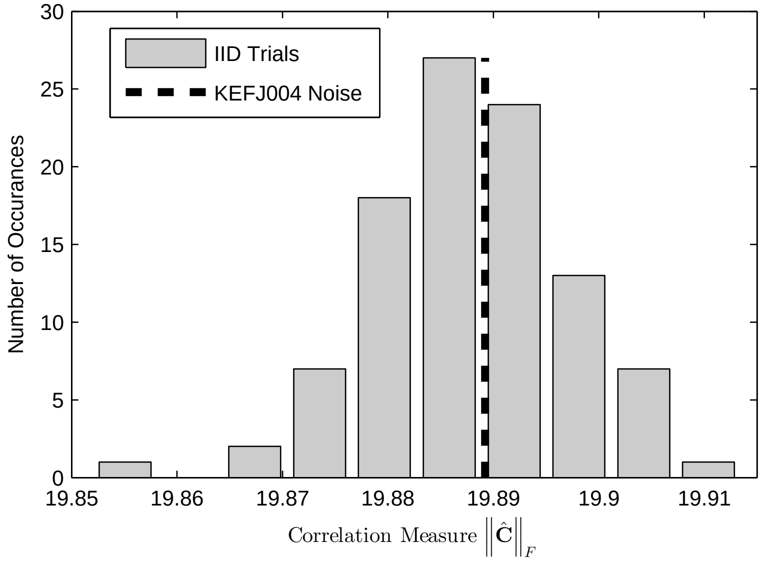


FIGURE B.2. Histogram of the coherence measures obtained for a set of IID multivariate Gaussian vectors, together with the coherence measure obtained for KEFJ004 noise vectors.

in Section B.2. For this test, the set of $J = 14,960$ N-dimensional vectors was used to form a size $M = 1,496$ set of $NK = 330$ dimensional vectors (i.e., $K = 10$), with each element representing a different realization $\tilde{\mathbf{w}}^{(m)}$. Each $\tilde{\mathbf{w}}^{(m)}$ was formed using K temporally adjacent 1/3 octave vectors containing noise alone to ensure this test measures independence within a sequence of $K = 10$ vectors. Since the coherence test involves calculating a measure using only the data, as opposed to making a comparison between the data and an independent set as in the energy distance test in (B.7), only a single trial was performed to obtain a measure of coherence between noise vectors in the KEFJ004 data set. This value is shown by the dashed vertical line in Fig. B.2. Though multiple trials could potentially be performed, this would require an abundance of data, and using more samples to form $\hat{\mathbf{R}}_{\tilde{\mathbf{w}}}$ in (B.8) improves the accuracy of the test statistic. However, as with the Gaussianity test, for benchmarking purposes 100 trials were performed to generate a histogram of the coherence measures obtained for a set of vectors that are approximately $\overset{\text{IID}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$, which is also shown in Fig. B.2. As before, these benchmark vector sets were generated using the `'mvnrnd'` command in MATLAB.

As can be seen from Fig. B.2, the coherence measure between KEFJ004 noise vectors is close to the sample mean of the coherence measures obtained from the 100 trials using data that was specifically generated to be IID. The value of $\left\|\hat{\mathbf{C}}\right\|_F$ obtained for the KEFJ004 noise is also reasonably close to $\left\|\mathbf{I}_{NK}\right\|_F = \sqrt{330}$, which is the theoretical value of the coherence measure for IID Gaussian random vectors. Therefore, considering the results of the Gaussianity test above in addition to these independence results, it is indeed reasonable to accept $\mathcal{H}_0$ in (B.9) and declare $\mathbf{w}_k \overset{\text{IID}}{\sim} \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{w}}, \mathbf{R}_{\mathbf{w}}\right)$.

APPENDIX C

One-Third Octave Representation of Doppler-Shifted Waveforms

This appendix draws a link between the behavior of sources in motion that we wish to characterize (discussed in Section 2.2.2) and their 1/3 octave signatures [62] that must be used for such analysis. In particular, it is shown how Doppler impacts the 1/3 octave signatures of such sources by causing the frequency components to vary nonlinearly with time. The intension is to provide insight regarding what is perhaps the primary cause of variable and nonstationary signatures produced by extrinsic sources, which necessitates the development of sophisticated detection and classification approaches capable of tracking such signatures. First, a model for the received continuous time waveform corresponding to the signatures of a single narrowband source are presented. The implications of this model for the 1/3 octave signatures of a source are then discussed, together with simulations that directly show the impacts of source motion.

C.1. Signal Model

Consider the data collection scenario presented in Fig. C.1 where a fixed monitoring station (receiver) records the acoustical signatures of a single source in motion. For simplicity, the source is assumed to emit a single narrowband tone at frequency f_s, and maintain a constant velocity v throughout the observation period. Additionally, the source is assumed to have a constant linear trajectory so that the velocity and angle of the source w.r.t. the receiver, denoted by r_t and θ_t at time t, respectively, are predictable albeit nonlinear. Finally, only 2D motion of the source relative to the receiver is considered. Denoting f_d as the Doppler-shifted frequency, the received signal at

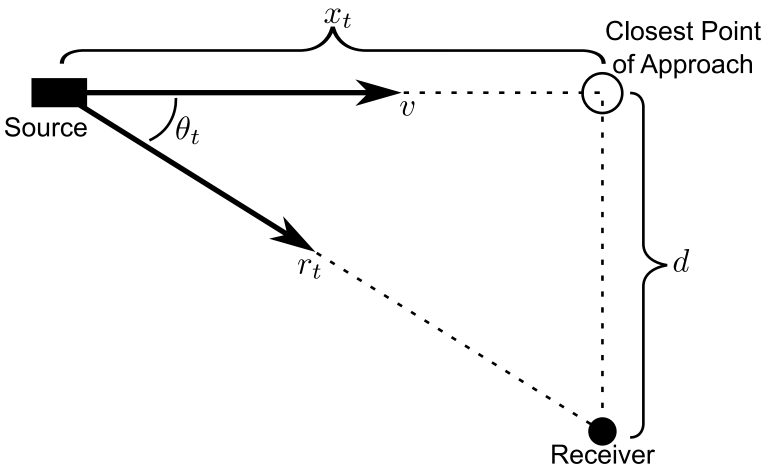


FIGURE C.1. Geometry of the spatial relationship between a narrowband source and receiver for the problem considered in this study.

time t can be written as

$$\begin{aligned} u_t &= A_t \cos(2\pi f_d t) \\ &= A_t \cos\left(2\pi \left(\frac{c}{c - r_t}\right) f_s t\right) \\ &= A_t \cos\left(2\pi \left(\frac{c}{c - v\cos\theta_t}\right) f_s t\right) \\ &= A_t \cos\left(\frac{2\pi f_s t}{1 - (v/c)\cos\theta_t}\right) \end{aligned} \tag{C.1}$$

where A_t is amplitude of the received waveform (dependent on source distance at time t) and c is the medium velocity, which is approximately equal to 340.29 m/s in air at sea level. Using the geometry suggested by Fig. C.1 where, without loss of generality, the motion of the source w.r.t. the receiver is parameterized by a single dimension, the angle of the source w.r.t. the receiver may be written as

$$\theta_t = \frac{\pi}{2} - \tan^{-1}\left(\frac{x_t}{d}\right)$$

where x_t is x-axis position of the source relative to the receiver at time t, and d is the minimum distance achieved between the source and receiver.

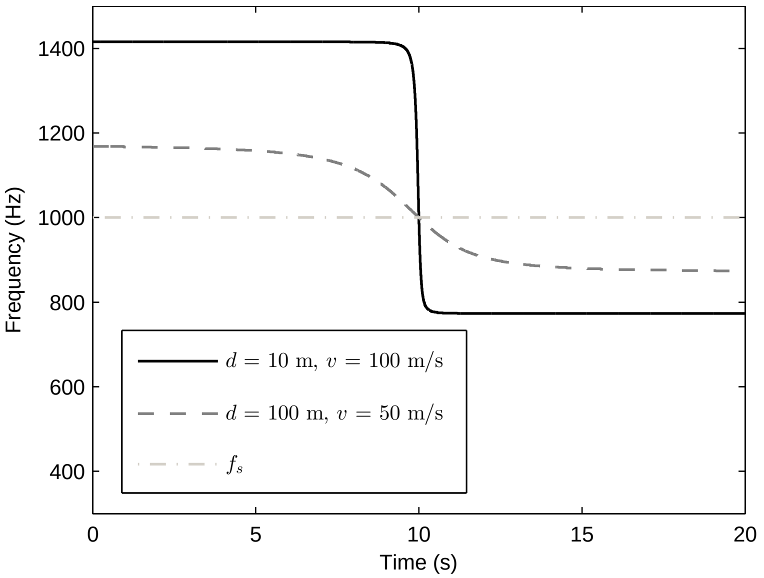
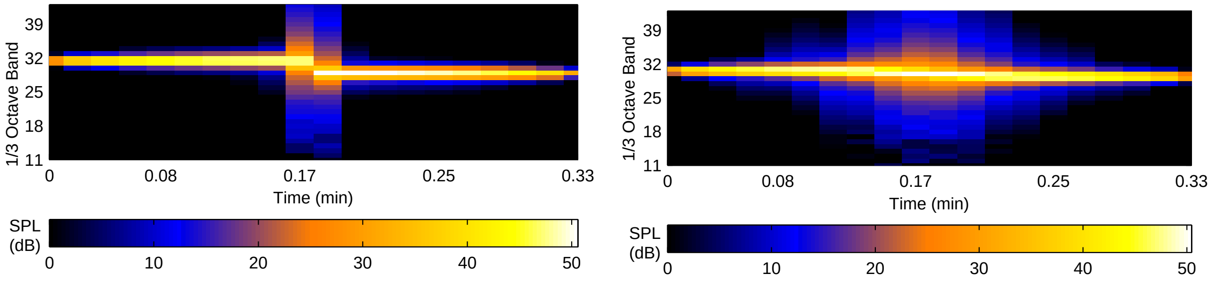


FIGURE C.2. Relationship between Doppler-shifted frequency and time for a source with the characteristics assumed in this study.

The multiplier $(1-(v/c)\cos\theta_t)^{-1}$ in (C.1) causes u_t to deviate from a single tone, to something that has nonlinear time-frequency characteristics that are similar to those displayed in Fig. C.2, for two $f_s = 1$ kHz sources with different motion characteristics. This is a direct consequence of r_t being a nonlinear function of time. In this figure it is assumed that a given source begins moving toward the receiver until it reaches its closest point of approach, at a distance d from the receiver, where it then moves away from the receiver. The resulting Doppler-shifted frequencies are monotonically decreasing in this scenario, being larger than f_s as it is moving towards the receiver, equal to f_s at its closet point of approach (where $t = 10$ s and $r_t = 0$ m/s), and smaller than f_s as it moves away from the receiver. As can be seen, the frequency/time slope and the consequent impact of Doppler is more severe for larger v and smaller d.

The consequences of these nonlinear frequency shifting characteristics cannot be understated, as they have a severe impact on the 1/3 octave vectors extracted from such signatures. Although finding an explicit form of 1/3 octave vectors extracted from Doppler shifted waveforms is possible, it is not done here since the results are unwieldy and seldom rewarding due to the fact that the motion parameters of sources are unknown in practice. Instead, simulations are presented that



(a) $f_s = 1$ kHz, $d = 10$ m, and $v = 100$ m/s.

(b) $f_s = 1$ kHz, $d = 100$ m, and $v = 50$ m/s.

FIGURE C.3. 1/3 octave vector sequences extracted from simulated waveforms corresponding to sources with different motion characteristics.

demonstrate the aforementioned effects in the 1/3 octave domain, in order to clarify the underlying causes of variations in the structure of the data to be analyzed.

C.2. SIMULATIONS

To illustrate the effects of Doppler in the 1/3 octave domain, two waveforms were simulated according to (C.1), that correspond to the received signatures of two sources with different motion characteristics. These sources have the same time-frequency characteristics as those shown in Fig. C.2, meaning they emit a single tone at $f_s = 1$ kHz, but one is parameterized by $d = 10$ m and $v = 100$ m/s, while other by $d = 100$ m and $v = 50$ m/s. These source velocities are reasonable values for some aircraft, e.g., propeller planes, though the distances were made to be overly small to exaggerate the effects of Doppler. Corresponding 1/3 octave vector sequences were extracted from these simulated time domain waveforms, and are shown in Figs. C.3(a) and C.3(b).

As can be seen in both cases, the energy present in the vector sequences is confined to the 29th – 31st 1/3 octave frequency bands (which contains f_s) when the velocity of the source w.r.t. the receiver r_t is near constant. However, during the times when the observed frequency changes most rapidly the energy become broadband in the 1/3 octave domain. This result is due to the fact that the time-frequency characteristics of a source with the assumed characteristics are nearly linear within the one second observation period used to extract a vector. Consequently, the approximate linear FM chirp in the time domain becomes a scaled linear chirp in the frequency domain [106], meaning the energy in the frequency domain is spread out over a wider range of the spectrum than Fig. C.2 might indicate. This is why the 1/3 octave vector sequence in Fig. C.3(a) has broadband signatures only during the two seconds in the middle of the sequence, while the change in frequency occurs more gradually over time for the source represented in Fig. C.3(b), leading to a larger number of 1/3 octave vectors with wideband signatures. The latter figure also demonstrates that the bandwidth of the energy is related to the magnitude of the time-frequency slope of the time domain source signatures shown in Fig. C.2. This momentary increase in perceived bandwidth is in addition to the shift in overall received frequency of each tone from the beginning to the end of a source's signatures.

As mentioned in Section 2.2.2, there are many other factors that contribute to nonstationary source signatures that are a mainstay of the present problem, though many of them are unique to a given source type, e.g., blade-vortex interaction for helicopters. When it comes to designing algorithms to detect and classify sources with such signatures, it is clear that they must be based on a flexible measurement model, that is capable of accommodating unpredictable source characteristics. For instance, since there are essentially no restrictions on the trajectory of a source w.r.t. the receiver, source motion parameters and signatures tend to vary wildly. Nonetheless, a successful approach to the present problem must be able to assign the same class label to acoustical events associated with the same source type, but with with largely different parameters, as with the simplified case in Fig. C.3.

APPENDIX D

Review of Matched Subspace Detectors

This appendix provides a review of the matched subspace detector (MSD), which uses the principle of the generalized likelihood ratio test (GLRT) [44] to yield a uniformly most powerful invariant detector. The MSD is can detect the presence of a signal in a vector observation corrupted with structured interference and additive Gaussian noise. The MSD is discussed here since it is a simple but powerful method for performing detection and classification separately on individual observations in a sequence, as mentioned in Section 3.2.3. Such decisions may then be combined using, e.g., the HMM-based sequential decision fusion in Section 3.3. Additionally, Appendix E demonstrates the relationship between the MSD and the sequential random coefficient tracking (SRCT) method introduced in Chapter 4. Many of the concepts outlined in this appendix are drawn from [40, 44, 107].

The exact formulation of the MSD varies depending on the assumed composition of the observation $\mathbf{y} \in \mathbb{R}^N$ under each hypothesis in terms of signal, interference, and noise, as well as the structure of each of these components [44]. Here, the following hypotheses testing problem is considered

$$\begin{aligned} \mathcal{H}_0 &: \mathbf{y} = \mathbf{h} + \mathbf{w} \\ \mathcal{H}_1 &: \mathbf{y} = \mathbf{s} + \mathbf{h} + \mathbf{w} \end{aligned} \tag{D.1}$$

where $\mathbf{s}$ and $\mathbf{h}$ are signal and interference vectors, respectively, and $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_\mathbf{w})$ is a zero-mean Gaussian noise vector with full rank covariance matrix $\mathbf{R}_\mathbf{w} \in \mathbb{R}^{N \times N}$. The goal of the MSD is then to detect the presence of $\mathbf{s}$ in $\mathbf{y}$. In this formulation only one class of signal and one class of interference are considered, meaning several MSDs would be required to detect and classify sources for the soundscape characterization problem considered in this thesis (see Appendix E for an elaboration of this concept).

As with the SRCT method in Chapter 4, the MSD is applied to transformed observations $\mathbf{z} = \mathbf{R}_{\mathbf{w}}^{-\frac{1}{2}}\mathbf{y}$ with white observation noise $\boldsymbol{\omega} = \mathbf{R}_{\mathbf{w}}^{-\frac{1}{2}}\mathbf{w}$, meaning $E\left[\boldsymbol{\omega}\boldsymbol{\omega}^T\right] = \mathbf{I}_N$, i.e.,B the $N \times N$ identity matrix. This transformation allows for writing the test statistic in a simpler form by removing its dependence on $\mathbf{R}_{\mathbf{w}}$ [107]. The underlying concept of the MSD is assuming that the transformed signal vector $\mathbf{R}_{\mathbf{w}}^{-\frac{1}{2}}\mathbf{s}$ and transformed interference vector $\mathbf{R}_{\mathbf{w}}^{-\frac{1}{2}}\mathbf{h}$ lie in known low dimensional subspaces, $\langle\mathbf{S}\rangle$ and $\langle\mathbf{H}\rangle$, respectively, that are spanned by the columns of $\mathbf{S} \in \mathbb{R}^{N\times M_s}$ and $\mathbf{H} \in \mathbb{R}^{N\times M_h}$, respectively, with $M_s < N$ and $M_h < N$ being the dimensionality of these subspaces. The subspaces $\langle\mathbf{S}\rangle$ and $\langle\mathbf{H}\rangle$ are not necessarily orthogonal, but they are linearly independent, meaning no vector in $\langle\mathbf{S}\rangle$ can be written as a linear combination of vectors in $\langle\mathbf{H}\rangle$, or vice versa. It follows that $\mathbf{R}_{\mathbf{w}}^{-\frac{1}{2}}\mathbf{s} = \mathbf{Sa}$ and $\mathbf{R}_{\mathbf{w}}^{-\frac{1}{2}}\mathbf{h} = \mathbf{Hb}$ where $\mathbf{a} \in \mathbb{R}^{M_s}$ and $\mathbf{b} \in \mathbb{R}^{M_h}$ are deterministic but unknown vectors that respectively contain the signal and interference subspace coordinates.

Since $\mathbf{a}$ and $\mathbf{b}$ are deterministic and $\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$, the hypothesis test in (D.1) may now be written as

$$\begin{aligned} &\mathcal{H}_0 : \mathbf{z} \sim \mathcal{N}(\mathbf{Hb}, \mathbf{I}_N) \\ &\mathcal{H}_1 : \mathbf{z} \sim \mathcal{N}(\mathbf{Sa} + \mathbf{Hb}, \mathbf{I}_N). \end{aligned} \tag{D.2}$$

Therefore, the parameter of the density of $\mathbf{z}$ is $\Theta = \{\mathbf{a}, \mathbf{b}\}$, and the likelihood of Θ given $\mathbf{z}$ is

$$\ell(\Theta; \mathbf{z}) = \frac{1}{\sqrt{(2\pi)^N}} \exp\left\{-\frac{1}{2}\|\boldsymbol{\omega}\|_2^2\right\}$$

where the noise may be written as

$$\boldsymbol{\omega} = \mathbf{z} - \mathbf{Sa} - \mathbf{Sb}. \tag{D.3}$$

Since the MSD is a GLRT, the test statistic is based on a log-likelihood ratio (LLR) of the parameters under each hypothesis, i.e., Θ_i for hypothesis $\mathcal{H}_i$. To make the test generalized, each Θ_i is replaced by its maximum likelihood estimate (MLE) $\hat{\Theta}_i$ to yield the test statistic [44]

$$L(\mathbf{z}) = 2 \ln \left(\frac{\ell(\hat{\Theta}_1; \mathbf{z})}{\ell(\hat{\Theta}_0; \mathbf{z})} \right) = \|\hat{\boldsymbol{\omega}}_0\|_2^2 - \|\hat{\boldsymbol{\omega}}_1\|_2^2 \tag{D.4}$$

where $\hat{\boldsymbol{\omega}}_i$ is the MLE of the noise under $\mathcal{H}_i$, that is obtained by using the MLEs of the source coordinate vectors under the same hypothesis (elements of $\hat{\Theta}_i$), in (D.3). In particular, it can be shown [44] that

$$\hat{\boldsymbol{\omega}}_0 = \mathbf{z} - \mathbf{P}_{\mathbf{H}}\mathbf{z} = \mathbf{P}_{\mathbf{H}}^{\perp}\mathbf{z}$$

$$\hat{\boldsymbol{\omega}}_1 = \mathbf{z} - \mathbf{P}_{\mathbf{SH}}\mathbf{z} = \mathbf{P}_{\mathbf{SH}}^{\perp}\mathbf{z}$$

where $\mathbf{P}_{\mathbf{H}} = \mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T$ and $\mathbf{P}_{\mathbf{SH}} = [\mathbf{S}, \mathbf{H}]\left([\mathbf{S}, \mathbf{H}]^T[\mathbf{S}, \mathbf{H}]\right)^{-1}[\mathbf{S}, \mathbf{H}]^T$ are orthogonal projection matrices for the subspaces $\langle \mathbf{H} \rangle$ and $\langle [\mathbf{S}, \mathbf{H}] \rangle$, respectively, while $\mathbf{P}_{\mathbf{H}}^{\perp} = \mathbf{I}_N - \mathbf{P}_{\mathbf{H}}$ and $\mathbf{P}_{\mathbf{SH}}^{\perp} = \mathbf{I}_N - \mathbf{P}_{\mathbf{SH}}$ project onto the orthogonal complements of these subspaces, respectively. See Section 4.3.3 or [44] for a more detailed explanation as to why these are the MLEs of the noise in each case.

Finally, the test statistic in (D.4) may be written as

$$\begin{aligned} L(\mathbf{z}) &= \mathbf{z}^T\mathbf{P}_{\mathbf{H}}^{\perp}\mathbf{z} - \mathbf{z}^T\mathbf{P}_{\mathbf{SH}}^{\perp}\mathbf{z} \\ &= \mathbf{z}^T\left(\mathbf{P}_{\mathbf{SH}} - \mathbf{P}_{\mathbf{H}}\right)\mathbf{z}. \end{aligned}$$

The form of this test statistic demonstrates that, in general, MSDs operate by determining the energy of the components of the observation that lie in the signal subspace after the interference components have been removed. The hypothesis test is then implemented as

$$L(\mathbf{z}) \quad \mathop{\gtrless}_{\mathbf{z} \in \mathcal{H}_0}^{\mathbf{z} \in \mathcal{H}_1} \quad \eta$$

where η is a predetermined threshold.

APPENDIX E

Relationship between MSD and SRCT

In Chapter 4, a sequential random coefficient tracking (SRCT) method was introduced as a comprehensive solution to the source characterization problem considered in this thesis. This method is based on applying a hierarchy of log-likelihood ratio tests (LLRT) to each observation (see Section 4.3.4) to discover its composition in terms of signal, interference, and noise, and assumes basis coefficients associated with each source type follow a linear autoregressive model. This appendix examines the simplified forms of these LLRTs for the special case where source basis coefficients are deterministic but unknown. In particular, there exists an interesting relationship between the SRCT test statistics and those defined by the matched subspace detector (MSD) introduced in Appendix D. Note that the notation used in this appendix is the same as that in Chapter 4.

Recall that the SRCT method considers the following hypothesis testing problem

$$\begin{aligned}
\mathcal{H}_0 &: \mathbf{z}_k = \boldsymbol{\omega}_k \\
\mathcal{H}_1^{(p)} &: \mathbf{z}_k = \mathbf{S}^{(p)}\mathbf{a}_k + \boldsymbol{\omega}_k \\
\mathcal{H}_2^{(q)} &: \mathbf{z}_k = \mathbf{H}^{(q)}\mathbf{b}_k + \boldsymbol{\omega}_k \\
\mathcal{H}_3^{(p,q)} &: \mathbf{z}_k = \mathbf{S}^{(p)}\mathbf{a}_k + \mathbf{H}^{(q)}\mathbf{b}_k + \boldsymbol{\omega}_k
\end{aligned}$$

where $\mathbf{S}^{(p)}$ and $\mathbf{H}^{(q)}$ contain basis vectors that span the pth signal space and qth interference space, respectively, while $\mathbf{a}_k$ and $\mathbf{b}_k$ are the coordinates of the signal and interference components of the observation relative to these subspaces, respectively. In Chapter 4, $\mathbf{a}_k$ and $\mathbf{b}_k$ were considered random, but in this appendix they are assumed to be deterministic but unknown. Application of the SRCT method to an observation vector $\mathbf{z}_k$ involves calculating

$$\boldsymbol{\zeta}_k = \frac{1}{2}\ln\det\left(\boldsymbol{\Sigma}_k\right) + \frac{1}{2}\left(\mathbf{z}_k - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{z}_k - \boldsymbol{\mu}_k\right) \tag{E.1}$$

for each hypothesis listed above, where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ denote the mean vector and covariance matrix of a multivariate Gaussian distribution, respectively, that are unique for each hypothesis and observation.

The model under each hypothesis can be used to find a corresponding parameter set $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ for the observation $\mathbf{z}_k$. When source signatures are deterministic but unknown, the noise vector $\boldsymbol{\omega}_k \overset{\text{IID}}{\sim} \mathcal{N}\left(\mathbf{0}, \mathbf{I}_N\right)$ is the only random term in $\mathbf{z}_k$, and hence, $\boldsymbol{\Sigma}_k = \mathbf{I}_N$ under all hypotheses. Since no sources are present under $\mathcal{H}_0$, the corresponding test statistic remains unchanged as $\boldsymbol{\zeta}_k = \frac{1}{2}\mathbf{z}_k^T\mathbf{z}_k$. For other hypotheses, using the notation in Table 4.1, the mean vector $\boldsymbol{\mu}_k$ may be replaced by its maximum likelihood estimate (MLE) $\hat{\boldsymbol{\mu}}_k = \mathbf{C}\hat{\mathbf{x}}_k$, where $\hat{\mathbf{x}}_k$ is the MLE of the state vector that contains estimates of basis coefficients ($\hat{\mathbf{a}}_k$ and/or $\hat{\mathbf{b}}_k$), given by either (4.16) or (4.20). More specifically,

$$\begin{aligned}
\mathcal{H}_1^{(p)} &: \hat{\mathbf{x}}_k = \mathbf{S}^{(p)\dagger}\mathbf{z}_k \\
\mathcal{H}_2^{(q)} &: \hat{\mathbf{x}}_k = \mathbf{H}^{(q)\dagger}\mathbf{z}_k \\
\mathcal{H}_3^{(p,q)} &: \hat{\mathbf{x}}_k = \left[\left(\mathbf{S}^{(p)\dagger}\mathbf{E}_{\mathbf{S}}^{(p,q)}\mathbf{z}_k\right)^T,\ \left(\mathbf{H}^{(q)\dagger}\mathbf{E}_{\mathbf{H}}^{(p,q)}\mathbf{z}_k\right)^T\right]^T
\end{aligned}$$

where $\mathbf{E}_{\mathbf{S}}^{(p,q)}$ and $\mathbf{E}_{\mathbf{H}}^{(p,q)}$ are oblique projection matrices defined in Section 4.3.3, and † means Moore-Penrose inverse [76]. Now, (E.1) can be used to form the estimate $\hat{\boldsymbol{\zeta}}_k = \frac{1}{2}\left\|\mathbf{z}_k - \mathbf{C}\hat{\mathbf{x}}_k\right\|_2^2$ where $\left\|\cdot\right\|_2$ denotes the ℓ_2-norm. By assigning $\mathbf{D} = \mathbf{I}$ from Table 4.1 (of appropriate dimension), the explicit form of $\hat{\boldsymbol{\zeta}}_k$ under each hypothesis is then

$$\begin{aligned}
\mathcal{H}_1^{(p)} : \hat{\boldsymbol{\zeta}}_k &= \frac{1}{2}\left\|\mathbf{z}_k - \mathbf{S}^{(p)}\mathbf{S}^{(p)\dagger}\mathbf{z}_k\right\|_2^2 = \frac{1}{2}\mathbf{z}_k^T\mathbf{P}_{\mathbf{S}}^{(p)\perp}\mathbf{z}_k \\
\mathcal{H}_2^{(q)} : \hat{\boldsymbol{\zeta}}_k &= \frac{1}{2}\left\|\mathbf{z}_k - \mathbf{H}^{(q)}\mathbf{H}^{(q)\dagger}\mathbf{z}_k\right\|_2^2 = \frac{1}{2}\mathbf{z}_k^T\mathbf{P}_{\mathbf{H}}^{(q)\perp}\mathbf{z}_k \\
\mathcal{H}_3^{(p,q)} : \hat{\boldsymbol{\zeta}}_k &= \frac{1}{2}\left\|\mathbf{z}_k - \mathbf{P}_{\mathbf{S}}^{(p)}\mathbf{E}_{\mathbf{S}}^{(p,q)}\mathbf{z}_k - \mathbf{P}_{\mathbf{H}}^{(q)}\mathbf{E}_{\mathbf{H}}^{(p,q)}\mathbf{z}_k\right\|_2^2 = \frac{1}{2}\mathbf{z}_k^T\mathbf{P}_{\mathbf{SH}}^{(p,q)\perp}\mathbf{z}_k
\end{aligned}$$

where the matrices $\mathbf{P}_{\mathbf{S}}^{(p)\perp}$, $\mathbf{P}_{\mathbf{H}}^{(q)\perp}$, and $\mathbf{P}_{\mathbf{SH}}^{(p,q)\perp}$ project orthogonally onto the subspaces $\left\langle\mathbf{S}^{(p)}\right\rangle^{\perp}$, $\left\langle\mathbf{H}^{(q)}\right\rangle^{\perp}$, and $\left\langle\left[\mathbf{S}^{(p)},\mathbf{H}^{(q)}\right]\right\rangle^{\perp}$, respectively. These $\hat{\boldsymbol{\zeta}}_k$'s are identical to the square of the ℓ_2-norm of noise estimates used to generate test statistics for a MSD [44], as shown in Appendix D for general $\mathcal{H}_2^{(q)}$ and $\mathcal{H}_3^{(p,q)}$ hypotheses.

In essence, when sources may be modeled as having deterministic but unknown basis coefficients, the SRCT framework may be implemented as the application of a set of MSDs to $\mathbf{z}_k$ by replacing $\boldsymbol{\zeta}_k$'s with $\hat{\boldsymbol{\zeta}}_k$'s when calculating the LLRTs in Section 4.3.4. Consequently, there is no need for estimation of state vectors using a Kalman filter in this deterministic case, as the distribution of each observation is not dependent on previous observations. Furthermore, the estimate of the source signatures under each hypothesis is simply $\mathbf{C}\hat{\mathbf{x}}_k$.