Robust Anomaly Detection

with Applications to Acoustics and Graphs

by

Nash M. Borges

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

January, 2014

Abstract

> There are four sorts of men:
> He who knows not and knows not that he knows not:
> He is a fool—shun him;
> He who knows not and knows that he knows not:
> He is simple—teach him;
> He who knows and knows not that he knows:
> He is asleep—wake him;
> He who knows and knows that he knows:
> He is wise—follow him.
>
> —*Arabic proverb*

Our goal is to develop a robust anomaly detector that can be incorporated into pattern recognition systems that may need to be taught, but will never be shunned. The ability to *know what we do not know* is a concept often overlooked when developing classifiers to discriminate between different types of normal data in controlled experiments. We believe that an anomaly detector should be used to produce warnings in real applications when operating conditions change dramatically, especially when other classifiers only have a fixed set of bad candidates from which to choose.

Our approach to distributional anomaly detection is to gather local information using features tailored to the domain, aggregate all such evidence to form a global

ABSTRACT

density estimate, and then compare it to a model of normal data. A good match to a recognizable distribution is not required. By design, this process can detect the "unknown unknowns" [1] and properly react to the "black swan events" [2] that can have devastating effects on other systems. We demonstrate that our system is robust to anomalies that may not be well-defined or well-understood even if they have contaminated the training data that is assumed to be non-anomalous.

In order to develop a more robust speech activity detector, we reformulate the problem to include acoustic anomaly detection and demonstrate state-of-the-art performance using simple distribution comparison techniques that can be performed at high speeds. We begin by demonstrating our approach when training on purely normal conversational speech and then remove all annotation from our training data and demonstrate that our techniques can robustly accommodate anomalous training data contamination. When comparing continuous distributions in higher dimensions, we develop a novel method of discarding portions of a semi-parametric model to form a robust estimate of the Kullback-Leibler divergence. Finally, we demonstrate the generality of our approach by using the divergence between distributions of vertex invariants as a graph distance metric and achieve state-of-the-art performance when detecting graph anomalies with neighborhoods of excessive or negligible connectivity.

Primary Reader: Gerard G. L. Meyer

Secondary Reader: Hynek Hermansky

Acknowledgments

This research would not have been possible without several educational programs provided by the United States Department of Defense.

First, I would like to thank to my advisor, Gerard G. L. Meyer. He taught me how to structure long-term research and the importance of being able to explain my work to any audience. I am also grateful for his patience and understanding throughout my graduate research while I was balancing it with a demanding career.

To my first research mentor, Teresa Kamm, and my good friend, Melanie Rudoy, who inspired me to pursue a Ph.D. by showing me the way. To Allen Gorin and Dave Farris, who helped me discover the importance of this work. To Chuck Wooters, who helped turn some of these ideas into reality. To Glen Coppersmith and Bradley Skaggs, who have been true friends and inspiring colleagues that read several drafts of this work and provided valuable insights. To my supervisors, Jack Godfrey and Renee Burton, who helped me begin and navigate a career in research while giving me the freedom to pursue my graduate research.

To professors Hynek Hermansky, Sanjeev Khudanpur, and Carey Priebe, who

taught me about biologically-inspired systems, speech processing, random signal analysis, robust statistics, and graph theory that form the bedrock of this research. To other friends and colleagues at both The Center for Speech and Language Processing and the Human Language Technology Center of Excellence, who have been an invaluable sounding board for my research.

To the teachers who inspired me to love science, the arts, and computers over the years, Lois Bennett, Christine Conklin, and Linda Watson. To professors Karen Panetta, Amir-Homayoon Najmi, and Jack Carmody, who introduced me to digital logic, signal, and speech processing and encouraged me to continue down this path.

To my wife, Heather, who has supported and encouraged me throughout the last decade when there was rarely a time that I was not both working and going to school at the same time. To my cousin Cristina, who was the first (and only) person to describe me as “a scholar” inspiring me to pursue this path. To Granny, who always helped me with my homework, proofread my essays, and inspired me to pursue a higher education. I miss you dearly. To my father, who showed me that getting started is sometimes the only way to address a complex problem with no clear solution. To my mother, whose love and support has always been unconditional. Without you, none of this would have been possible.

To all the other friends, colleagues, mentors, and family who have supported me throughout this endeavor and to whom I am eternally grateful.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Imagine a girl growing up in the West learning about different types of produce. She knows that apples, oranges, and bananas are *fruits* and that lettuce, carrots, and celery are *vegetables*. Suppose that she is then presented with a noni (Figure 1.1) and asked whether it is a fruit or vegetable. Noni (*Morinda citrifolia*) are yellowish-green and “have an unpleasant odor resembling cheese” [6]. They are edible, but mostly tasteless, and have been used as a botanical remedy by Polynesians for over 2000 years to fight cancer [7], infection [8], arthritis, asthma, hypertension, and pain [9, 10].

If the girl has never seen a noni before, she might consider it an anomaly since it “deviates markedly” [11] from the produce that is familiar to her. She could guess the type of produce if pressured, but she would be aware of her uncertainty and could voice it. The noni is actually a multiple fruit called a syncarp formed from several flowers that combine into a single fleshy mass. This category would probably be new

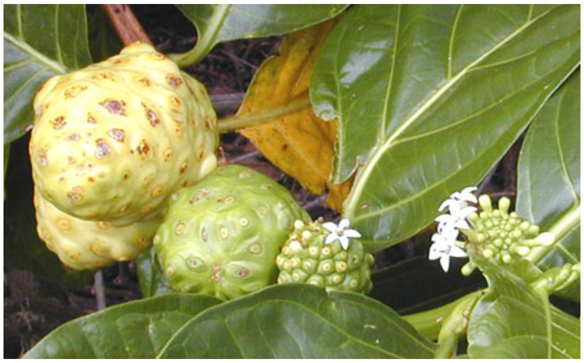

Figure 1.1: Noni fruit (*Morinda cirifolia*) developing by Eric Guinther, June 2, 2005, Creative Commons Attribution-Share Alike.

to the girl and she might actively request to see other multiple fruit (e.g. pineapples and figs). The child's ability to know what she does not know is a concept often overlooked when developing pattern recognition systems.

In addition to novel object detection, another marvel of the human visual system is that it enables the girl to recognize known objects despite changes in viewing conditions. Recent psychophysical experiments suggest that this is done through view combination, which is a form of evidence aggregation that uses information derived from multiple views [12–14]. These multiple views can come from object rotations or observer movements [15] and subsequent recognition likely follows after a normalization process that compensates for moderate changes in perspective [16–18].

If we consider each image of an object that enters the girl's visual system as a single complex unit of information, her motion or that of the object allows her visual system to acquire many such units of information from slightly different perspectives as a temporally bound sample derived from the same physical scenario [19]. This sample is a subset of the overall population of all images that could be amassed by observing all viewpoints of the object under all viewing conditions.

We want to mimic these robust capabilities that are inherent in human perception and develop automated methods of robust anomaly detection using evidence aggregation. This will allow us to perform a more holistic comparison between data that we want to classify and data that we know is normal[1]. In order to do this we will model the population of measurements from a finite sample and then estimate the statistical divergence between populations.

1.1 Machine Learning

Distinguishing fruits from vegetables provides an example of a task common to many fields of study including machine learning, data mining, pattern recognition, statistics, medicine, and operations research. The goal is to infer the type of produce given some measurements such as its size, shape, weight, color, or texture. The machine learning community would consider this a binary classification task, where the goal is to train a classifier that can decide between two labels. They would recommend that some *features* be chosen which characterize each piece of produce and hopefully are useful for the task at hand. Since vegetables are often green and fruit are often vibrant shades of red, orange, or yellow, we could begin by measuring

[1] To avoid confusion, we will use the term “normal” to mean non-anomalous and “Gaussian” when referring to the statistical distribution.

the color of each piece of produce. In order to get a machine to learn from this data, we need a set of *training data* providing each item along with its feature (color) and label (type) as shown in Table 1.1. Large data sets often used in machine learning tasks can have millions of such examples, but we will use this toy data set for the purpose of this introduction to key concepts and nomenclature.

After we have trained our classifier on the information in Table 1.1, we could present it with some *test data* that we want it to label. We would expect that our system would label a yellow lemon and green asparagus as fruit and vegetable, respectively. The two yellow items in its training data are fruit and seven out of eight green items are vegetables. However, the yellowish-green noni is a color that we have not explicitly seen before and therefore it is unclear how our classifier should proceed. If we had a metric to estimate the distance between colors, perhaps by using the color's hue instead of treating it like a categorical attribute, we might be able to better generalize from our training data and recognize that yellowish-green is "close" to both yellow and green. However, if we did so and determined that yellowish-green is equally distant from both yellow and green, how should we proceed? We have learned that yellow produce are fruit and green produce are mostly vegetables and we have an item to label that is somewhere in between. We can imagine that our *confidence* of a piece of produce being a fruit changes smoothly along the spectrum starting very high at yellow and ending up quite low at green (Figure 1.2). We know that somewhere along the path our confidence is exactly 50%, which represents the

Table 1.1: Produce colors and types

Data	Product	Color	Hue $\in [-60, 300]$	Type
Training	apple	red	-7.9	fruit
	cherry	red	-5.7	fruit
	raspberry	red	-3.9	fruit
	strawberry	red	0.0	fruit
	carrot	orange	15.2	vegetable
	orange	orange	25.0	fruit
	lemon	yellow	39.4	fruit
	banana	yellow	44.9	fruit
	broccoli	green	70.0	vegetable
	celery	green	75.4	vegetable
	artichoke	green	79.4	vegetable
	lettuce	green	82.2	vegetable
	lime	green	88.8	fruit
	asparagus	green	90.0	vegetable
	spinach	green	101.4	vegetable
	cucumber	green	109.7	vegetable
Test	noni	yellowish-green	59.3	???
	blue crab	blue	200.8	???

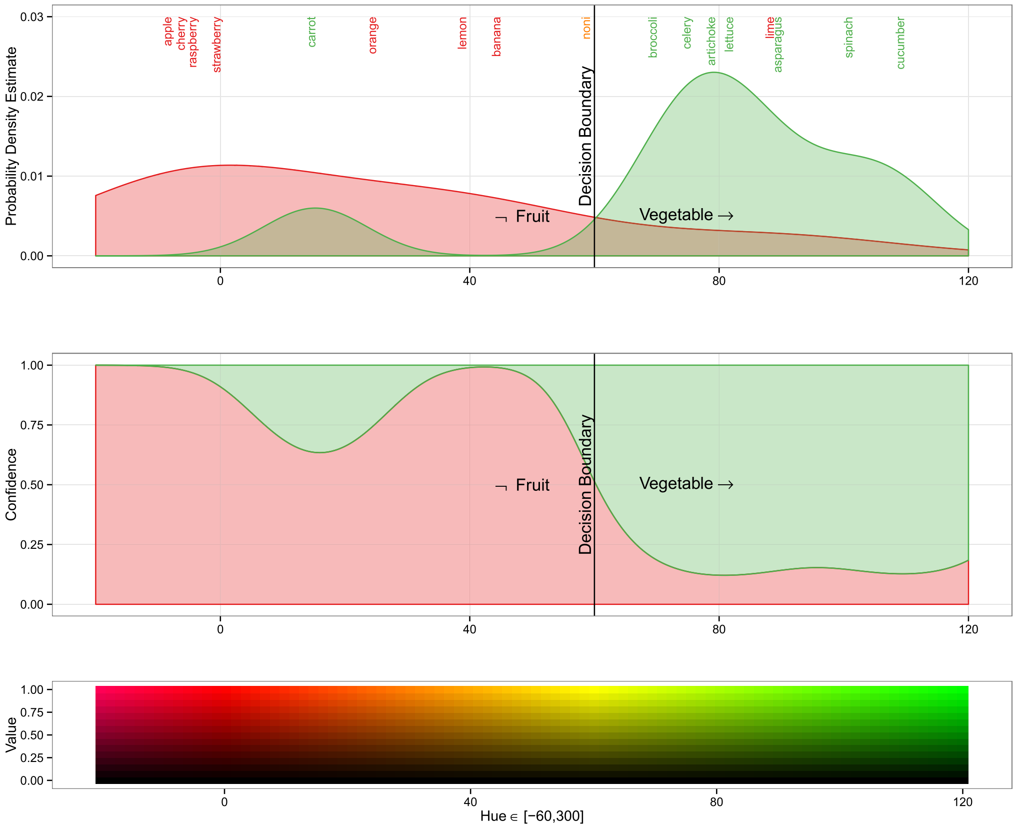

Figure 1.2: Kernel probability density estimates (Section 2.1.3) of color hue for fruits and vegetables (top), derived confidence estimate (middle), and color hue spectrum (bottom). Note that the noni is very close to the decision boundary leading to uncertainty about its class label.

decision boundary where the label that we would give to a piece of produce changes from fruit to vegetable.

Proximity to the decision boundary is a well studied area of pattern recognition [20], but it is not the only source of uncertainty when using machine learning algorithms in real applications. Uncertainty can also come from input that is unlike any of the training data. If our machine is asked to classify a Chesapeake blue crab as either a fruit or vegetable, what should it do? It comes from an entirely different *domain* (i.e. animals), which our produce detector was not trained to recognize or ignore. When we look outside the scope of our training data to the full range of color hues (Figure 1.3), we can see that due to its limited domain knowledge, our classifier will label all green objects as vegetables and everything else as fruit even if the item is not produce. While that may sound like a pleasant coincidence given our motivating reason for using color, it can lead to disastrous unintended consequences. Since the blue crab is far away from a decision boundary, our confidence estimation that assumed all input was produce led our classifier to be nearly certain that a blue crab is a fruit. The challenge is how to recognize this phenomena in higher dimensions that we cannot easily visualize.

To deal with that difficulty, we propose using an anomaly detector in parallel with any discriminative classifier trained on a finite amount of labeled examples (Figure 1.4). This anomaly detector should be able to learn from a random sample of unlabeled data, which is mostly normal assuming that anomalies are rare. Such an approach will ensure that other pattern recognizers can be relied upon and nothing drastic has changed in the data at-large. Viewing the problem this way naturally leads us to model the distribution of the entire population, even if we have taken great care to build other pattern recognizers whose job is to differentiate between various subpopulations. When we want to determine whether or not the data has drastically changed, we will estimate the statistical divergence between the current population and that on which it was trained.

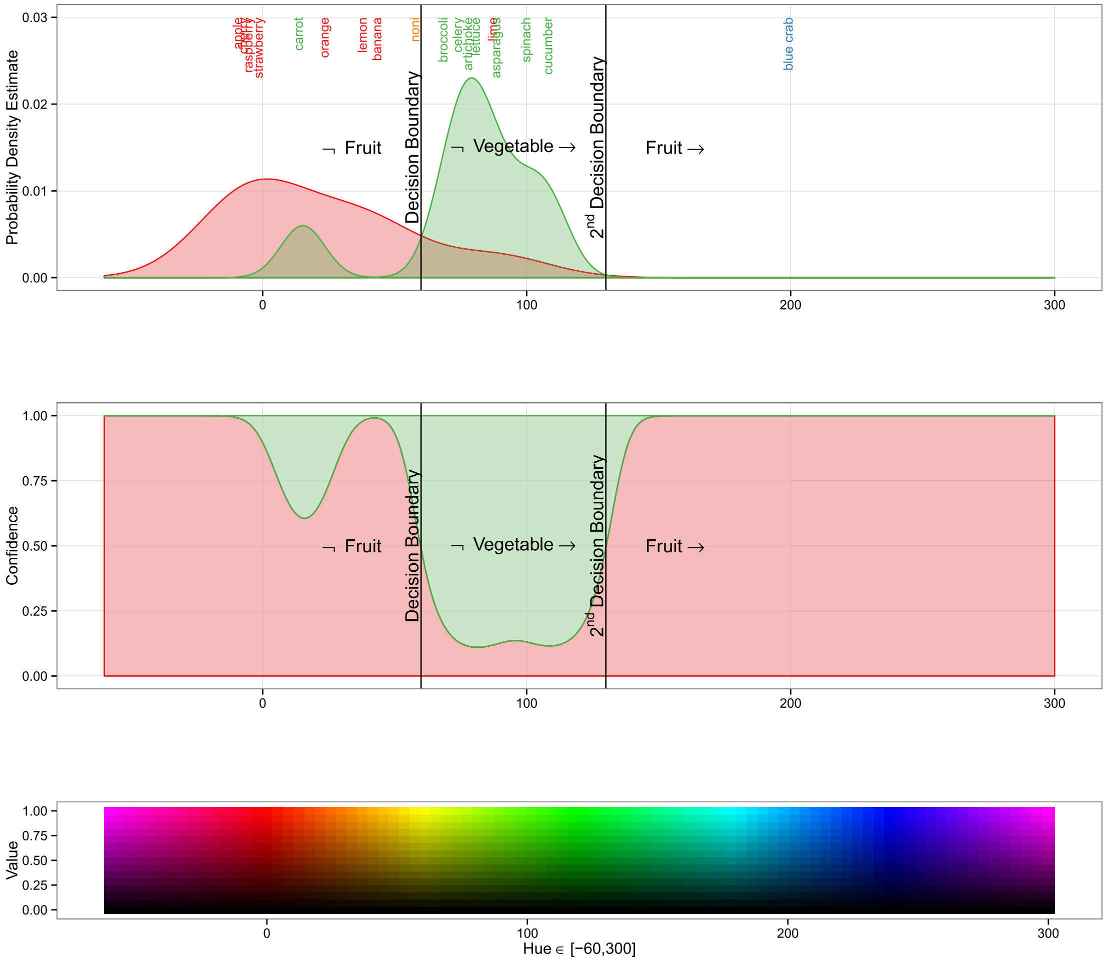

Figure 1.3: Kernel probability density estimates (Section 2.1.3) of color hue for fruits and vegetables (top), derived confidence estimate (middle), and color hue spectrum (bottom). Note that for this particular model, the "fatter tail" of the hue distribution for fruit led to a second decision boundary at a hue level outside of the support of the training data. Since the blue crab is far from both decision boundaries, this lead to a false sense of confidence that it should be labeled as a fruit.

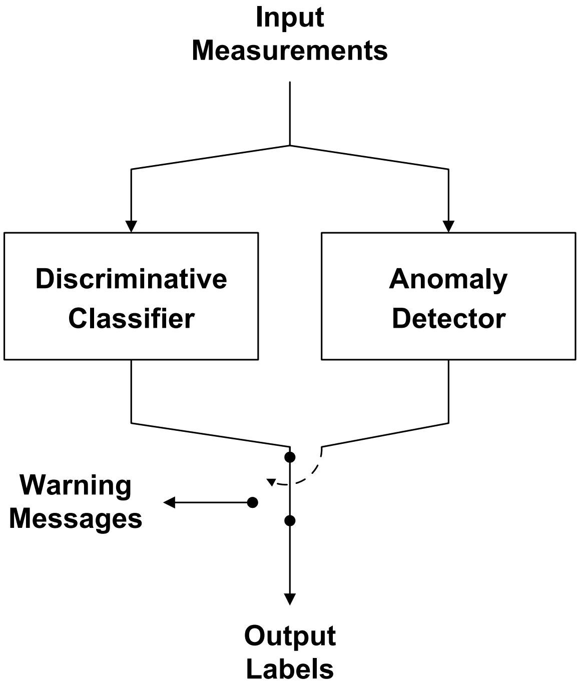

Figure 1.4: Robust pattern recognition system.

When operating in a streaming environment, where all the data cannot be saved and decisions must be made in near real-time despite statistical drift, such a system could periodically sample new data to re-learn what constitutes normal. Alternatively, the fixed set of training data used for other pattern recognizers could also be used for the anomaly detector. This would allow it to detect when the data has changed enough that the other pattern recognizers need to be retrained.

When classifying produce, we could estimate the distribution of color hue *for each item* using many measurements, rather than relying on a single sample which could be noisy. An anomaly detector could then estimate the statistical divergence between each item and other data that we know is normal for produce. We have done this for photographs of a strawberry, lettuce, noni, and blue crab in Figure 1.5. By aggregating evidence, we can see another attribute that could not be detected with a single measurement. While the strawberry, lettuce, and noni have a single peak in their distribution of color hue, the blue crab has two distinct peaks from its blue legs and orange claws. Such an attribute is unusual for a single piece of produce, which would be easy to detect with a distributional anomaly detector.

1.2 Controlled Experiments vs. Real Applications

Recently, there has been considerable excitement in the machine learning community about the use of deep neural networks. Hinton and others [21–23] have achieved unprecedented performance on challenging benchmarks in computer vision [24, 25] and speech recognition [26–30] using new methods to train these artificial neural networks with many hidden layers. Before deep neural networks came of age [31], other breakthroughs in statistical learning were achieved using support vector machines [32] and ensemble classifiers, such as boosting [33], bagging [34], and random forests [35]. Such highly sophisticated techniques have led to significant advances in pattern recognition theory and unprecedented performance in controlled experiments. However, Hand [36] notes that improvements seen in controlled experiments may not yield success in a real application if they are "swamped by other sources of variation". This

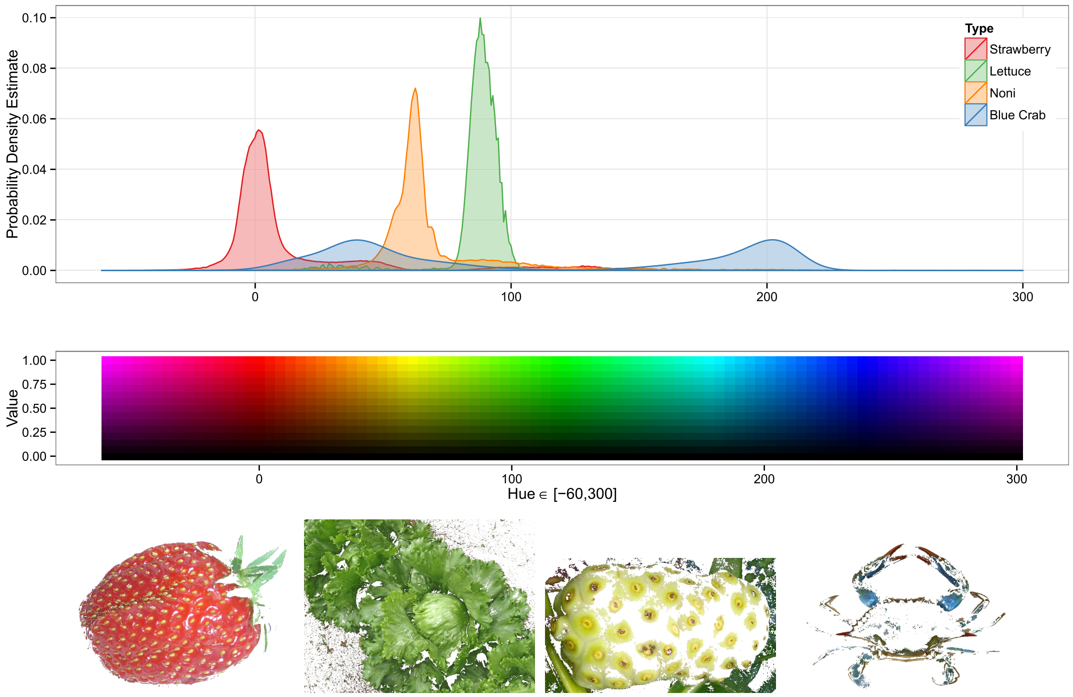

Figure 1.5: Kernel probability density estimates (top) of color hue (middle) for photographs of a strawberry, lettuce, noni, and blue crab (bottom from left to right). The pixels with low color saturation (below 0.25) or low value (below 0.25) in the HSV color space have been removed to help eliminate the background and highlight the vibrant colors. Strawberry (*Fragaria ananassa*) by David Monniaux, 7 May 2005, Creative Commons Attribution-Share Alike. Iceberg lettuce (*Lactuca sativa*) by Jan Bakker, 20 May 2007, Creative Commons Attribution-Share Alike. Noni (*Morinda citrifolia*) fruit by Wilfredo Rodriguez, 19 October 2008, Creative Commons Universal Public Domain Dedication. Chesapeake blue crab (*Callinectes sapidus*) by Wendy Kaveney, September 13, 2004, Creative Commons Attribution-Share Alike.

variation can be the result of commonly made implicit or explicit assumptions that may not hold in real applications [37]:

1. Training and test data are drawn randomly from the same population.
2. The distribution of the data does not change over time.
3. The classes are well defined.
4. There are no errors in the labels of the training or test data.
5. The costs of making different types of classification errors are known accurately.
6. Measurement error, missing data, and mislabeled examples are equally represented in the controlled experiment and real application.
7. The criteria used to evaluate the performance of the classifier in the controlled experiment are derived from requirements known to produce operational impact in the real application.

These assumptions are well-intentioned and usually made to improve the tractability of difficult problems. As Box famously wrote, “essentially, all models are wrong, but some are useful” [38]. This work is devoted to verifying the first two assumptions, while being robust to the third, fourth, and sixth, in order to address the fifth and seventh.

1.2.1 Robust Statistics

The term “robustness” was first coined by Box when he noted that statistical tests of equal means work well for non-Gaussian data, unlike tests of equal variance [39]. Robust statistics provide alternatives to classical statistical methods without being severely impacted by outliers or slightly inaccurate assumptions. For robust estimates of central tendency, Tukey [40] and others proposed the trimmed and Winsorized means, which remove a fraction of the smallest and largest samples and either discard them or replace them with the maximal remaining values. Shortly thereafter, Huber observed that classical estimators were not robust due to their inherent Gaussian assumptions and reliance on least squares estimation [41]. He developed a general theory of robust statistics and showed the mean, median, and maximum likelihood (ML) estimates all to be special cases of M-estimators that minimize some function of the error between the samples and the estimator.

In order to quantify robustness, Hampel defined the “breakdown point” of an estimator to be the smallest proportion of *contamination* that can cause it to take on “arbitrarily large aberrant values” [42]. The *mean* is not robust with breakdown point 0, since a single outlier can arbitrarily affect the estimate, whereas the α-*trimmed mean* is robust with breakdown point α, and the *median* with breakdown point $\frac{1}{2}$. We will use robust statistics to ensure that our approaches to anomaly detection can learn from normal data that might be contaminated with some anomalies (Section 1.2.2). We will also test to see if our methods can robustly detect anomalies that were not observed in the training data, but should be expected when deploying a pattern recognition system in the real world.

1.2.2 Noisy Labels

The fourth assumption in our list (Section 1.2) is that the labels in the training data are correct. While this is usually the case for the vast majority of data carefully annotated by expert professionals [43], developing such corpora is an expensive and time consuming process. To avoid this, researchers often use methods that require less labeled data or seek alternative sources of annotation.

One popular method of getting labeled data at little cost is relying on some underlying structure or metadata that is typically outside the scope of the machine learning algorithm. Google's PageRank [44, 45] is one famous example that does not require additional annotation. Rather than collecting human judgments about the subjective importance of each Web page in order to help rank search results, PageRank uses the underlying structure of the Web to infer the importance of each page objectively based on how other pages link to it. Another common example of cheap labeled data on the Web is using the URL [46] or ISO 639-3 language code [47] for training text language identification systems.

Games that produce annotations as a side effect are another innovative source of free labeled data. Von Ahn pioneered some of the early work in this area of "gamification", developing the ESP Game to collect image descriptions [48], Verbosity for common-sense facts [49], and DuoLingo for web translations [50]. Yahoo! Answers[2] takes a slightly different approach to human computation in a question-answering forum and rewards members who provide the best answers with positive reinforcement and points that increase their status in the community.

Monetary-based crowdsourcing platforms such as Amazon's Mechanical Turk[3] and CrowdFlower[4] are increasingly being used [51] as cheap, fast, but *noisy* sources of annotation [52] for a variety of media including text, image, audio, and video [53–56]. While these platforms provide a low-cost, scalable labor force ideal for performing small tasks that require human intelligence, crowdworkers typically do not have any specialized training and have higher disagreement rates than professional annotators for a variety of reasons [57]. Although quality control should be carefully monitored and cheating addressed, the majority of crowdworkers complete their effort in good faith and for some tasks [53,57] their annotations can be nearly as effective for training machine learners as those produced by experts.

Returning to our produce example, if we were to decide on the type of each piece of produce using labels provided by crowdworkers, then it is likely that we would see a fair amount of noisy labels. *Ambiguities* are one source of noisy labels that might lead to a low rate of inter-annotator agreement, such as the differences in the botanical and cultural definitions of fruit and vegetables. For example, we might find that the majority of the crowdworkers consider tomatoes to be vegetables, while

[2] http://answers.yahoo.com
[3] http://mturk.com
[4] http://crowdflower.com

those in the scientific community consider them fruits. Though this example was specifically chosen to illustrate a challenging task without a correct answer, human-produced labels are rarely in unanimous agreement [58,59]. This type of error, where data is close to a decision boundary, is difficult for even humans to resolve and is not the focus of our work. When this does occur, we choose to let the data "speak for themselves" [60] and we will not concern ourselves with the age-old debate about the proper classification of a tomato[5].

On the other hand, *unambiguous* mistakes in labeling can be due to a variety of reasons. For example, a poorly designed human intelligence task (HIT) can be exploited by crowdworkers for monetary gain. Cheating and collusion [62] have been seen on Mechanical Turk and are usually combated with pre-screening or occasional performance verification tests [57,63]. If we are not so careful, we might find that some crowdworkers cheat and make purposefully erroneous decisions to finish the task as quickly as possible, such as labeling all produce as fruit. If we are not using crowdsourcing, we could rely on assumptions about the data or natural partitions thereof, such as assuming that all produce at a salad bar are vegetables. We could similarly assume that any sweet produce eaten as a desert are fruit. While these assumptions will not be *entirely correct*, they would allow us to quickly attain many labeled examples at minimal cost.

Our focus will be on these types of unambiguous mistakes. They are likely to be

[5] The United States Supreme Court ruled unanimously in Nix vs. Hedden [61] that the tomato is a vegetable for the purposes of taxation, while granting that it is a botanical fruit.

correctable at some cost, but we assume that we are unable or unwilling to pay it and we will use robust machine learning techniques to accommodate any assumptions that are only *mostly correct*.

1.2.3 Statistical vs. Egregious Errors

When the costs of different types of errors are discussed in the literature, it is often in reference to misses versus false alarms, but there are other subtle and more subjective error conditions in real applications. For example, when using an algorithm to place advertisements on websites, it is easy to imagine how inappropriate advertisements could end up on a website about sexual harassment in the workplace. What is the cost of harming a company's reputation and how could that be factored into an optimization technique without knowing all the different ways that it could happen?

Returning again to our example in Section 1.1, if our classifier is presented with produce that is red, orange, yellow, or green, then we can hopefully rely on it to make reasonable decisions. The classifier will probably still make some *statistical errors* such as incorrectly labeling a lime as a vegetable, but this can be studied ahead of time allowing for appropriate performance expectations to be set. There are many reasons for this type of error, such as the inherent difficulty of the task or biases in the training data. Most users of automated systems will accept such errors, especially if they appreciate the difficulty of the task and gain some benefit from the automated solution.

However, *egregious errors* are often viewed more harshly and quickly lead to the distrust of pattern recognition systems. If our produce classifier repeatedly decides that a Chesapeake blue crab is a fruit, then it may be considered unreliable and might stop being used altogether. While our classifier was not trained to recognize animals or distinguish them from plants, an intelligent system should be able to recognize such a drastic abnormality and act accordingly. We want to develop a robust anomaly detector that can flag such items as unusual and warn us when other pattern recognizers are likely to make egregious errors because we are operating in a *region of uncertainty* [64].

Even though we want to detect anomalies, we do not want to learn about them by studying specific examples for several reasons. First, anomalies are rare and difficult to find, so collecting data about them can be prohibitively expensive. Second, unlike normal data, anomalies can result from many different causes and learning from a few easy to find examples could lead to poor generalization and result in egregious errors. Therefore, we seek to learn only from mostly normal data and identify anything that deviates markedly from that as anomalous.

1.2.4 Criteria Other Than Accuracy

As each new machine learning technique improves on the previous state-of-the-art, there are fewer errors to correct. This is especially true for problems where the best systems are approaching the Bayes error rate, which is a theoretical bound on how well any method can perform given a certain set of measurements [65]. Often, the biggest strides are made early on in the application of machine learning to a given problem. Hand found that for many different problems, 90% of the predictive power of the best methods can be achieved using the simplest models [37]. After some initial progress has been made, “if large improvements are [still] possible, they are more likely to come from a reformulation of the problem” [66]. Common ways of reformulating a problem include developing new features and using large amounts of unlabeled data. Criteria other than accuracy must also be considered when choosing a classification method for a real application [66]:

1. What prior knowledge does the researcher already have about the technical and application domains?[6]

2. Is a validated implementation of the algorithm readily available?

3. What expertise is required to use the algorithm and tune its parameters?

4. Is parameter tuning automatic? Does it require separate data?

5. How much labeled data is available? Can the algorithm take advantage of large amounts of unlabeled data?

6. Does the algorithm automatically increase model complexity as additional training data is available?

[6] Both are critical!

7. Can the algorithm handle different types of data? Can it cope with measurement error? missing data? mislabeled examples? statistical drift?

8. Do complex feature combinations need to be supplied as input or can those be learned by the algorithm?

9. What is the speed and cost of making each measurement, pre-processing the data, training the algorithm, and applying the classification rule?

10. Can the learning and classification be done on-line in a streaming fashion or does the data have to be processed in batches?

11. Can the researcher interpret or visualize what the model learns about the training data or why a decision is made for a particular example?

In summary, the real question a researcher generally wants to ask is "which classification method is best for *me* to use on *my* problem with *my* data" [67]. The only answer is that "different classifiers suit different problems" [66] and the best approach is simply to get started. Our goal is to develop an anomaly detector that can learn from noisy labels and be used as part of a robust pattern recognition system to avoid making egregious errors while remaining computationally efficient.

1.3 Dissertation Contributions

- In the interest of developing a robust speech activity detector, we reformulate the problem to include acoustic anomaly detection and demonstrate state-of-the-art performance using simple distribution comparison techniques that can be performed at high speeds.

- In the interest of robustly comparing two distributions in a high dimensional space, we develop a novel method of discarding portions of a semi-parametric model when estimating the Kullback-Leibler divergence.

- In the interest of robustly detecting any type of anomalous graph, we reformulate the problem of assessing graph similarity by comparing distributions of local measurements. We demonstrate superior performance to available state-of-the-art approaches against a specific type of anomaly and further demonstrate superior generalization to entire classes of graph anomalies.

1.4 Structure of Dissertation

In Chapter 2, we discuss the preliminaries necessary for our approach of estimating the statistical divergence between populations. We begin our focus on acoustics in Chapter 3 and demonstrate our approach when training on purely normal conversational speech. Then, in Chapters 4 and 5 we remove all annotation from our training data and demonstrate that our techniques using histograms and trimmed Gaussian mixture models, respectively, can be used to robustly accommodate anomalous training data contamination. In Chapter 6, we demonstrate how our approach can generalize to other domains by introducing our second application focus on graphs. We discuss the change from temporally focused acoustic features to vertex-centric graph features and evaluate our best performing histogram techniques. Finally, Chapter 7 summarizes our findings and proposes ideas for future research.

Chapter 2

Statistical Inference

In Chapter 1, we described our goal of developing an anomaly detector using robust statistics (Section 1.2.1) which can accommodate noisy labels (Section 1.2.2) and be used as part of a robust pattern recognition system (Figure 1.4) that can avoid making egregious errors (Section 1.2.3) while remaining computationally efficient (Section 1.2.4). In this chapter, we will discuss the statistical inference paradigms of density estimation and machine learning required to do so.

Probability distributions provide a framework to succinctly characterize observations and enable statistical inference. Their use abounds in science and mathematics, from the measurements and errors that we assume are Gaussian[1] to the Poisson approximation[2] of the sum of independent Bernoulli trials. Even when empirical ob-

[1] Like many modeling assumptions this is often wrong, but mathematically useful [38].

[2] In a classic twist of eponymy, while Poisson invented the Cauchy distribution, it is likely that credit for the general formula bearing his own namesake actually belongs to de Moivre, although it was Bortkiewicz who first recognized its significance [68].

servations are not characterized by a typical size or “scale”, their distribution can still be modeled using power laws [69] such as Pareto’s distribution of wealth [70], Zipf’s law of word frequencies [71], Lotka’s distribution of scientific productivity [72], and Mandelbrot’s study of price changes [73] with infinite variance[3]. With experience and enough exposure to samples drawn from these distributions we can develop an intuition about which observations seem reasonable and which do not. Thus, we “know” that the probability of a man being nine feet tall[4] is nearly zero, and find it reasonable that the word “the” is found about once every twenty words in English [76]. While no finite set of real data obeys any of these laws exactly, they provide useful mechanisms to succinctly characterize the underlying processes.

Unlike many pattern recognition systems which make *local* decisions, we will explore methods of recognizing deviations from these *global* distributions. Local measurements are those nearby in some space which will depend on the application. We will use time for acoustic processing and vertex-centric neighborhoods when working on graphs. We will show that our global approach can detect the Rumsfeldian“unknown unknowns” [1] and Talebian “black swan events” [2] that can cause pattern recognition systems to produce *egregious errors* whose cost can far exceed the expected *statistical errors* when operating in new domains.

[3] Price changes over a fixed time period may follow a Lévy distribution with infinite variance, which is not necessarily incompatible with price changes being Gaussian for a fixed number of transactions since the number of transactions in any given period of time is random [74].

[4] Robert Wadlow was the world’s tallest man standing 8 feet 11.1 inches tall as a result of an overactive pituitary gland [75].

2.1 Density Estimation Paradigms

Density estimation is the act of estimating the probability that a member of a certain category will be found to have particular features [77]. In our produce example from Chapter 1, this could be used to find the most likely color hue observed for a piece of produce (Figure 1.3) given a distribution estimate derived from many measurements (Figure 1.5). Only after modeling the entire distribution of color hue for individual items, did we discover the difference between the unimodal color hue distributions of produce and the anomalous bi-modal distribution of a blue crab with orange claws.

We will generally rely on Wasserman [78] and Bickel and Doksum [79] for notation and theory. Formally, let X be a random variable yielding observation $x = X(\omega) \in \mathbb{R}$ for a particular experiment ω in sample space Ω. If X has distribution F, denoted $X \sim F$, then $F(x) = P(X \leq x)$ represents the probability that $X \leq x$. If its density p is absolutely continuous, then $F(x) = \int_{-\infty}^{x} p(t)dt$ for all x. In our multivariate setting, $\boldsymbol{X} = (X_1, \ldots, X_D)^T$ is a random vector made up of multiple random variables $X_1, \ldots, X_D$, where a particular experiment $\omega \in \Omega$ yields an observation vector $\boldsymbol{x} = \boldsymbol{X}(\omega) \in \mathbb{R}^D$.

2.1.1 Parametric Models

Density estimates can be described as either parametric, semi-parametric, or non-parametric [80]. Parametric models make strong assumptions about the data generating process, and are so named because they use a fixed number of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$ to characterize a family

$$\mathcal{P} = \left\{ \hat{p}(\boldsymbol{x}; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta} \right\}, \tag{2.1}$$

where $\boldsymbol{\Theta} \subset \mathbb{R}^k$ is the parameter space. We say that the model is *correct* when the true density p is an element of the family $\mathcal{P}$. When this is true we can accurately estimate the parameters with relatively little data. Perhaps the most well-known multivariate parametric model is the multivariate Gaussian distribution where the probability of an input $\boldsymbol{x} \in \mathbb{R}^D$ is

$$p(\boldsymbol{x}; \boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\exp\left(-\frac{1}{2} \left(\boldsymbol{x} - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{x} - \boldsymbol{\mu}\right) \right)}{\left(2\pi\right)^{d/2} \left|\boldsymbol{\Sigma}\right|^{1/2}}, \tag{2.2}$$

where the model parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are comprised of a mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

2.1.2 Semi-Parametric Models

Semi-parametric density estimation “enlarges” the model family for greater flexibility, while maintaining a fixed number of parameters regardless of sample size. One common form of this is a mixture model

$$\mathcal{P} = \left\{ \hat{p}(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{i=1}^{m} w_i \cdot p(\boldsymbol{x}; \boldsymbol{t}_i) : \boldsymbol{t}_i \in \boldsymbol{T}_i \right\}, \tag{2.3}$$

where each component $p(\boldsymbol{x}; \boldsymbol{t}_i)$ has its own parameter space $\boldsymbol{T}_i \subset \mathbb{R}^{k_i}$ and weight $w_i > 0$, such that $\Sigma_{i=1}^{m} w_i = 1$ for $i = 1, \ldots, m$. The mixture model parameters $\boldsymbol{\theta} = (w_1, \boldsymbol{t}_1, \ldots, w_m, \boldsymbol{t}_m)$ are the superset of each component's parameters, which often belong to the same sub-family. One classic example presented in Dempster, Laird, and Rubin's seminal paper on the expectation-maximization (EM) algorithm [81] is the Gaussian mixture model (GMM)

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{i=1}^{m} w_i \cdot \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{2.4}$$

with parameters $\boldsymbol{\theta} = (w_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \ldots, w_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$.

2.1.3 Non-Parametric Models

Non-parametric models enable even greater flexibility, although it is hard to precisely define what constitutes a non-parametric density estimate. Scott [80] states that a heuristic definition is that non-parametric density estimates "work"[5] for a "large" class of true densities. Silverman says that they make "less rigid assumptions" about the distribution and let the data "speak for themselves" [60]. Terrell and Scott [82] have an elegant definition which states that the influence of an individual example should vanish asymptotically for non-parametric estimates, which is often not true

[5] More precisely, they must be consistent in the mean square.

for parametric or semi-parametric models[6]. In other words, non-parametric density estimates should be asymptotically local and robust to outliers, which we discussed in Section 1.2.1. Technically, it is a slight misnomer to call them "non-parametric" because they do have parameters, but their number often grows with the sample size or is theoretically infinite [83].

One common non-parametric model is the kernel density estimate [84, 85]. Given a sample of independent and identically distributed draws $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, where $\boldsymbol{x}_i = (x_i^{(1)}, \ldots, x_i^{(D)})^T \in \mathbb{R}^D$, the multivariate product kernel density [80]

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{nh_1 \cdots h_D} \sum_{i=1}^{n} \left\{ \prod_{j=1}^{D} K\left(\frac{x^{(j)} - x_i^{(j)}}{h_j} \right) \right\} \tag{2.5}$$

is essentially a mixture model with "kernel" function K centered at each sample. Therefore, the parameters $\boldsymbol{\theta} = (h_1, \ldots, h_D, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ include all the samples and smoothing parameters $h_j > 0$ for each dimension $j = 1, \ldots, D$. While a Gaussian or uniform kernel is often used for visualization purposes, Epanechnikov [86] demonstrated that the optimal kernel is

$$K(y) = \frac{3}{4}(1 - y^2)\mathbb{1}\left(|y| < 1\right) \tag{2.6}$$

regardless of the true probability density, sample size, and dimensionality.[7]

A histogram[8] is another non-parametric density estimate where a continuous space is sub-divided into regions within which the probability distribution is assumed to be

[6] We have devoted Chapter 5 to overcoming this limitation for GMMs.

[7] $\mathbb{1}(\cdot)$ is the indicator function, which is 1 or 0 depending on whether the condition is true or false.

[8] The term *histogram* was originally coined by Pearson [87] to describe a "historical diagram" that was useful for visualizing the "reigns of sovereigns or periods of different prime ministers".

uniform. To this day, histograms “remain the most widely applied and most intuitive non-parametric estimator” [80]. Histograms can use adaptive bins with non-uniform sizes or they can use fixed bin sizes over the entire space.

Techniques using fixed bins come in two varieties: The entire feature space can be divided up into regular intervals or the saved bin locations can be derived after clustering the data. A histogram with fixed bins at regular intervals can be defined by parameters $\boldsymbol{\theta} = (\boldsymbol{b}_1, c_1, \ldots, \boldsymbol{b}_m, c_m)$ comprised of bin centroids $\boldsymbol{b}_j$ with corresponding counts

$$c_j = \sum_{i=1}^{n} \mathbb{1} \left(j = \arg\min_k d(\boldsymbol{b}_k, \boldsymbol{x}_i) \right) \tag{2.7}$$

using the Euclidean distance d, for samples $\boldsymbol{x}_i$ for $i = 1, \ldots, n$. We convert this to a density estimate using add-one smoothing,

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{c_j + \frac{1}{m}}{(\sum_{k=1}^{m}(c_k)) + 1}, \tag{2.8}$$

where $j = \arg\min_k \; d(\boldsymbol{b}_k, \boldsymbol{x})$ for a histogram with m cells.

2.2 Learning Paradigms

The goal of pattern recognition is to learn a mapping from input measurements to output labels. In the 1920’s, Fisher [88] formulated the classical solution to this problem, which begins by using input measurements to estimate the most likely parameters of a parametric model that could have generated the data. The equations governing these data generating processes can then be used to derive a decision rule, but there are two drawbacks to this approach.

First, it requires *a priori* knowledge about the type of distribution that generated the data. While this was often the case when Fisher was designing small sample experiments in agricultural and genetic studies, a century's increase in computational power has led to the common practice of opportunistic data mining in high dimensional, heterogeneous measurement spaces. These complex measurement spaces are often not well understood making accurate model creation difficult.

Second, solving the intermediate problem of explaining the data generating processes can be more difficult than addressing the original goal of learning a decision rule to apply labels. In statistical learning theory, Vapnik [32] took a more direct approach. Through a process known as structural risk minimization, he suggests learning how to best discriminate between different types of input while balancing model complexity. However, since these discriminative classifiers (discussed further in Section 2.2.5) lack a model of the data generating processes, they also lack the ability to detect if those processes have fundamentally changed. For pattern recognition systems that will be used in real applications, we believe a crucial secondary goal is recognizing if such a change has occurred[9].

It is both Fisher's *quest for fundamental understanding* and Vapnik's *consideration of use* that form the cornerstone of the *use-inspired basic research* [89] that

[9] Verifying Assumption 1 in Section 1.2.

we will conduct in this dissertation. Our overall strategy is to employ both methodologies in parallel (Figure 1.4), allowing us to use powerful discriminative classifiers when operating on normal data while also having an anomaly detector warn us when anything is out of the ordinary.

2.2.1 Supervised Classification

Supervised binary classification is the prediction of binary class labels from a set of training data. Formally, let $\boldsymbol{X} : \Omega \to \mathbb{R}^D$ be a random vector and $Y : \Omega \to \{0, 1\}$ be a random variable yielding observations $\boldsymbol{x} = \boldsymbol{X}(\omega) \in \mathbb{R}^D$ and $y = Y(\omega) \in \{0, 1\}$ for a particular experiment ω in the sample space Ω. $\boldsymbol{X}$ represents the input measurements and Y represents the output label. Depending on the field of study, the input can also be referred to as a feature vector, observations, or explanatory variables and the output as a decision, class, category, or dependent variable.

We are interested in learning a decision rule $g : \mathbb{R}^D \to \{0, 1\}$ so we can accurately categorize unlabeled data. When performing supervised learning, training data is required which consists of a set of paired observations and labels $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ which we assume are independent and identically distributed observations drawn from $p(\boldsymbol{x}, y)$. We will refer to $\boldsymbol{x}$ as a positive or negative example when $y = 1$ or $y = 0$, respectively.

To constrain the problem, statistical learning theory suggests [32] specifying the family of mapping functions $\mathcal{G}$ to be investigated using a parameter vector $\boldsymbol{\alpha} \in \boldsymbol{\Lambda}$

so that $\mathcal{G} = \{g(\boldsymbol{x}, \boldsymbol{\alpha}), \boldsymbol{\alpha} \in \boldsymbol{\Lambda}\}$. We also must choose a loss function, $L\left(y, g(\boldsymbol{x}, \boldsymbol{\alpha})\right)$ between the label y and our mapping $g(\boldsymbol{x}, \boldsymbol{\alpha})$. Vapnik suggests [90] using the indicator function $\mathbb{1}(y \neq g(\boldsymbol{x}, \boldsymbol{\alpha}))$ when performing binary classification. The goal of supervised classification can now be formally stated as choosing the mapping function to minimize the risk, or expected value of the loss function,

$$g(\boldsymbol{x}, \boldsymbol{\alpha}_0) = \arg\min_{\alpha} \int L\left(y, g(\boldsymbol{x}, \boldsymbol{\alpha})\right) dp(\boldsymbol{x}, y). \tag{2.9}$$

However, because learning $p(\boldsymbol{x}, y)$ is a challenging problem which often requires large amounts of labeled data that may not be available, Vapnik suggests [32] choosing the mapping function which minimizes the empirical risk,

$$g(\boldsymbol{x}, \boldsymbol{\alpha}_{emp}) = \arg\min_{\alpha} \frac{1}{n} \sum_{i=1}^{n} L(y_i, g(\boldsymbol{x}_i, \boldsymbol{\alpha})). \tag{2.10}$$

For small sample sizes, the goal of structural risk minimization is to minimize empirical risk while controlling complexity by defining a structured set of mapping functions [91]. In order to achieve this, support vector machines (SVMs) employ a strategy of keeping the empirical risk fixed while minimizing complexity [90].

In practice, SVMs project the data into a higher-dimensional space and then find an optimal separating hyperplane. When only considering classification accuracy (Section 1.2.4), properly tuned SVMs perform quite well, but they often require non-trivial tuning and a fair amount of expertise on the part of the researcher. They also lack the ability to detect if the data generating processes have changed.

2.2.2 Unsupervised Learning

Unsupervised methods are used to learn from unlabeled observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, and these methods often involve clustering data into groups by similarity. One reason to do this is to infer an unknown naturally occurring structure, or in other words to "carve nature at the joints" [66] for the purpose of better understanding. The interest may lie in estimating the number of classes present in the data, assigning each observation to a class, or both. Hierarchical clustering is a common example used to better understand complex data and its natural, but hidden, structure.

Another reason for unsupervised learning is to divide up data for the sake of convenience [66], often for dimensionality reduction or density estimation. An example of this that we will use heavily in this work is mixture modeling. While the induced class labels may not be important to the task at hand, the overall divide-and-conquer strategy is useful for semi-parametric density estimation (Section 2.1.2).

We will use unsupervised learning in Chapters 4 and 5 to train a binary classifier by assuming that most of the training data is normal and labeling anything unusual as anomalous. A similar example in Section 1.2.2 would be assuming that most produce at a salad bar are vegetables and labeling anything unusual as fruit. This anomaly detection strategy will be explored further in Section 2.3.

2.2.3 Semi-Supervised Classification

Semi-supervised learning [92] is often used when the process of labeling data is difficult, expensive, or time consuming, but a large amount of unlabeled data is readily available. In semi-supervised binary classification, a small amount of labeled data, typically consisting of both positive and negative examples, is available for training and the unlabeled data is used to identify low density regions of the feature space for optimal placement of the decision boundary [93, 94].

2.2.4 Partially Supervised Classification

If the cost or difficulty of finding different types of input is highly imbalanced, we may be forced to learn from labeled examples of only one of the classes. Such learning is referred to as *partially supervised classification* [95]. Learning from positive and unlabeled data recently became an active area of research [96–103] after it was first demonstrated in the context of the probably approximately correct (PAC) learning framework [104], where theoretical results show that it is sometimes sufficient to consider the unlabeled data as negative examples [105].

In document classification, if the number of both positive and unlabeled documents is large and the proportion of unlabeled documents that are positive is known, then a positive-naïve Bayes (P-NB) classifier [106] can be trained by statistically removing the effect of positive examples *from the model* of unlabeled data. This approach has been extended through the use of co-training [105] to better cope with a smaller number of positive documents.

Another common approach to partially supervised classification is to begin by automatically labeling some "reliable" negative examples from the unlabeled set and then iteratively applying "soft" labels to the unlabeled data used to train the classifier. Spy-expectation maximization (S-EM) [95] holds out some of the labeled positive examples to use as "spies" to identify likely negative examples from the unlabeled set using a NB classifier. This is then followed by iteratively applying the EM algorithm to estimate the hidden labels while training the classifier and keeping the positive labels fixed. When the NB assumption is not well-satisfied, logistic regression has been used on weighted examples yielding better performance [101]. When data sparsity is a concern, a support vector machine (SVM) can been used, which is the approach undertaken by PEBL [100] and Roc-SVM [102]. The latter is so-named because it uses the Rocchio algorithm as an initial classifier relying on the assumptions that positive examples are rare in the unlabeled set and negative examples cover a broader region of the feature space.

Others have simplified partially-supervised classification by ignoring the unlabeled data completely and performing *one-class classification* [107]. This is often done with an SVM trained to estimate the support of the positive examples [108]. While this approach can detect extreme outliers, it is highly dependent on the input features and choice of kernel in a manner that is not well understood [109]. This approach would also not work on data sets where the anomalies are in the same region of the feature space as normal data, but their distributions are anomalous (Figure 3.2)

One of many assumptions made in these approaches is that the unlabeled set is comprised mostly of the opposite class of examples from the labeled set. When retrieving relevant documents from a set that is mostly irrelevant such an assumption is valid, but the opposite is usually true for anomaly detection where the probability of occurrence is consistently small.

2.2.5 Generative vs. Discriminative Classification

Machine learning algorithms typically use generative models, discriminative classifiers, or a hybrid of both. Generative models estimate the joint distribution $p(\boldsymbol{x}, y)$ and choose the label which maximizes that probability [110]. This approach is so named because it models a distribution that could be used to generate novel feature vectors, although this is rarely done in practice. Dawid refers to this as the "sampling paradigm" [111] common in statistical theory, which concentrates on the distribution $p(\boldsymbol{x}|y)$ regarding the label y as an unknown causative parameter to the feature vector x. The drawback to this approach is that convenience or opportunity sampling, which is a common scenario when data mining [112], should not be used unless the selection bias is carefully corrected [113].

The knowledge stored in generative models is not strictly necessary for classification [90], but there are several uses that can increase robustness when input data to be classified may not have come from the same domain as the training data. Some forms of adaptation, such as feature mapping [114] and maximum likelihood linear regression [115], use generative models to develop linear transforms from the observed feature space into one that better matches trained models. We will explore the use of generative models to enable distributional anomaly detection. In the produce example from Chapter 1, we could have used a generative model of color hue for all produce to recognize that the blue crab was unlike anything we had seen before. Further, if the input distribution is anomalous like the blue crab's bi-modal color hue, a robust pattern classification system could produce warnings instead of labels. While confidence estimation has a similar goal, we saw how easily it can fail when anomalies which are far away from the decision boundary lead to false confidence (Figure 1.3).

Alternatively, in the "diagnostic paradigm" [111] discriminative classifiers model the posterior $p(y|\boldsymbol{x})$ estimating parameters for the decision boundary directly [116]. This approach has lower asymptotic classification error [110] when performing fully-supervised tasks, but lacks the additional knowledge required to recognize anomalous data that should not be processed normally.

2.3 Anomaly Detection

Anomaly detection has been studied in many different fields [117] such as network intrusion detection [118], monitoring systems [119], fraud detection [120], topic detection and tracking [121], data mining [122], and stylistic inconsistency detection [123]. In Hodge and Austin's survey of the subject [117], they categorized most research as either supervised classification, partially supervised novelty detection, and unsupervised clustering.

Anomaly detection techniques that require labeled instances of both normal and anomalous data fit into the traditional machine learning paradigm of supervised classification (Section 2.2.1). This technique is appropriate when the types of anomalies are fixed and obtaining labeled examples of them is easy. Partially supervised (Section 2.2.4) anomaly detection techniques that learn from only normal data are often referred to as novelty detectors. These methods are appropriate when it is relatively easy to obtain normal data, but anomalous examples are either difficult or expensive to obtain or are likely to change over time. Unsupervised clustering (Section 2.2.2) methods are appropriate when it is difficult to obtain any labeled examples. The classical example of unsupervised anomaly detection is the box plot (Figure 2.1), which can be used without any prior knowledge of the data. Unsupervised techniques commonly employ either *self-diagnosis* or *robust accomodation* of the unlabeled data in order to "cope with a sizable fraction of contamination" [124].

Now that we have described the statistical inference paradigms and nomenclature, we will evaluate the performance of distributional anomaly detection when we apply it to applications in speech processing and graph theory. Chapters 3 and 6 will investigate partially supervised anomaly detection in each of those applications, and

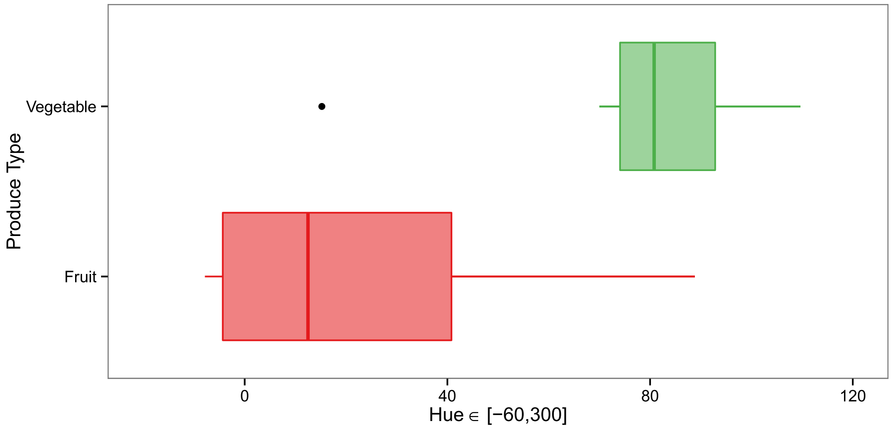

Figure 2.1: Box plots of color hue from the training data in Chapter 1. Note that the carrot has been identified as an outlier from the mostly green vegetables.

Chapters 4 and 5 will focus on unsupervised acoustic anomaly detection techniques that can accommodate noisy labels.

Chapter 3

Acoustic Anomaly Detection via Partially Supervised Learning from Normal Data

In Chapter 2, we introduced the statistical inference paradigms necessary for distributional anomaly detection. Here, we will investigate methods of computing the divergence between various density estimates to improve the robustness of speech activity detection by developing an acoustic anomaly detector that is trained using partially supervised learning from normal conversational speech.

The expressive potential of the human voice is enormous [125]. Spoken language contains both *explicit* information in its content as well as a variety of *implicit* information about the speaker [126]. It allows for recognition of the speaker's identity [127], provides evidence of their age [128], and conveys some information about their current health. It can also encode other demographic information about the speaker, such as their gender, geographical dialect, level of education [129], and social class in certain cultures [130]. It can even convey information about their current attitudes and emotions, medium term moods, and lifetime personality traits [126].

Automatic speech processing is the extraction of this explicit or implicit information by machine. While significant computational and algorithmic advances have been made in many speech processing systems over the last several decades, most still suffer from a lack of robustness with respect to noise, reverberation, and interfering speech [131] which makes them brittle under many natural conditions. In contrast, humans have an incredible ability to adapt to variations in acoustic conditions and their recognition performance degrades gracefully with increasing levels background noise and reverberation [132]. While speech and hearing may have evolved together to overcome many of these challenges, human comprehension is also robust to many synthetic signal degradations. For example, when designing analog scramblers for speech privacy, engineers have found it difficult to synthetically degrade speech beyond comprehension in a way that allows the intended recipient to reconstruct the signal. Systems that perform sample permutation, block permutation, and frequency inversion have all been found to be reasonably intelligible by humans [133] unless multiple techniques are used in tandem [134].

Humans also have an uncanny ability to exploit spectro-temporal redundancies in speech. Licklider and Pollack [135] showed that human intelligibility scores remained above 85% for speech with infinite clipping, a technique which only preserves the spectral aspect of the zero crossing rate and discards all temporal information about the speech envelope. If the speech was high pass filtered before being clipped, intelligibility exceeded 97%. Other experiments have shown that intelligibility also remains high when temporal information is preserved but spectral information is greatly reduced. Shannon et al. [136] found that by modulating the temporal envelope of band-limited white noise to mimic speech in as few as four broad frequency bands, words could be recognized nearly perfectly by human subjects.

Both high and low frequencies are important for recognizing different phonetic sounds. However, Fletcher and his team at Bell Labs showed that intelligibility remained above 97% when speech was high- or low-pass filtered [137] thereby demonstrating its spectral redundancies. Around the same time, Miller and Licklider [138] demonstrated temporal redundancies by masking speech intermittently with silence or noise of equal duration. They showed that this had little effect on understanding as long as it was done at least 10 times per second so that long contiguous segments were not removed.

These findings demonstrate that speech is an error correcting code with complex embedded redundancies enabling robust communication through a variety of noisy channels. If we are interested in improving the robustness of automated systems, we need to better exploit these redundancies so we can still recover information in the presence of missing or corrupted data. When processing audio to determine if it contains speech or not, we believe one method of doing so is to aggregate evidence over time and change the paradigm from short-term speech activity detection to long-term acoustic anomaly detection. Rather than assessing whether each fraction of a second of audio contains speech based on a single feature vector, in this chapter we will examine the features from five minutes of audio to determine if the entire segment as a whole is comprised of conversational speech worthy of further processing. This will enable us to make a more informed decision when presented with a potentially anomalous signal by examining its entire distribution instead of trying to aggregate many poorly made decisions *post hoc*.

3.1 Speech Activity Detection?

Speech processing systems are generally used to perform either *tokenization* or *classification* tasks. Tokenization is the process of converting an acoustic signal containing speech into the explicit content of the words, syllables, or phonemes being spoken. As the amount of input speech gets longer so does the number of output tokens. On the other hand, classification systems process a variable amount of speech and yield an output with a fixed size, often indicating their confidence associated with each decision. They typically extract some aspect of the implicit information about the speaker, such as their identity, language, gender, or emotional state. These speech classification tasks are often formulated as binary *detection* tasks, where the goal is to determine whether some implicit aspect of the speech is true or not.

Most speech processing systems depend on speech activity detection (SAD) as a pre-processing step. This enables the downstream tokenization or classification systems to know which part of the signal to analyze. This can save computational resources as well as avoid costly mistakes that can arise when presenting downstream processing systems with unusual non-speech signals that they were not trained to ignore. SAD also has other applications, such as saving bandwidth in cellular and Internet communications via discontinuous transmission [139], reducing environmental noise in hearing aid devices [140, 141], and improving intelligibility via echo cancelation [142, 143].

While SAD is a detection task, this definition refers to the decision made on each acoustic frame, which can be as short as 20 milliseconds. These frame-level decisions are frequently smoothed over time and the final output is more typical of tokenization, yielding a sequence of start and stop times indicating when speech is purportedly present in the acoustic signal.

Depending on the type and amount of background noise, the difficulty of SAD can range anywhere from incredibly easy to impossibly difficult. If the speaker is in a quiet indoor environment and the only likely sound is them talking, then a simple adaptive energy threshold [144] has been shown to perform quite well. This is also true for some telephony speech corpora such as Switchboard [129], which contains high quality landline telephone speech.

However, many SAD systems fail when the level of background noise increases [145] or if there is unexpected channel variability or degradation [146]. While significant progress has been made on many speech processing tasks over the last several decades, many automated systems degrade catastrophically under adverse acoustic conditions [147]. SAD and other speech processing systems based solely on processing short 20-30 millisecond frames typically lack robustness to noise and reverberation, conditions which do not significantly affect human comprehension.

3.2 Features

The speech signal is rich with information and one fundamental challenge of speech processing is distilling it into a manageable feature space for pattern recognition while preserving its explicit and implicit information. Consider the five minute audio segments that will be used in this chapter. Sampled at 8000 Hz, each segment consists of 2.4 million samples that are each discritized to one of 256 values, which is far too large to model directly. Furthermore, learning directly from these raw samples would require an immense amount of training data and would not generalize to new words, speakers, styles, or acoustic environments. Luckily, a temporal signal like speech is easy to *divide up meaningfully in time* allowing us to analyze each frame individually. This leaves us with three remaining questions:

- How long should each frame should be?
- What should we do with each frame?
- How should we aggregate evidence across many frames to summarize a segment?

3.2.1 Frame Length

To address the first question, we look to the human auditory system which inspires most dimensionality reduction techniques in speech processing. In the quest for the ideal time constant over which to analyze each frame of speech, a variety of psychoacoustic experiments have produced results ranging three orders of magnitude from 250 μs to 200 ms. Eddins and Green [148] believe that it is overly simplistic to assume that a single temporal window is used in the auditory cortex to analyze the incoming waveform. Several well-known experiments support this assertion and have demonstrated that perceived loudness increases with signal duration [149–152]. This time-intensity trade-off provides humans with high temporal acuity for loud sounds while maintaining the ability to hear quieter sounds via temporal integration [148].

Temporal acuity is often measured in gap detection experiments, which have found time constants ranging between 250 μs to 30 ms with many estimates clustered around 2 to 3 ms [153–155]. When performing temporal integration of quiet sounds, psychophysical and neurophysiological experiments suggest that the maximum time constant is between 100 and 200 ms [156]. Other experiments assessing the temporal order of two tones at different frequencies suggest that the spacing needs to be between 20 to 30 ms. This same delay is also required when discriminating between which of two lights came on first in experiments involving both sight and sound [157].

It is perhaps no accident that after decades of tuning and experimentation much of the speech community has settled on an analysis window of 20 to 30 ms. The most often cited reasoning for this window length is in some relation to the length of a phone, which is the smallest linguistic unit that can change the meaning of a word. However, the median duration for most phonetic classes is considerably longer, ranging between 60 to 100 ms [158]. Segmenting speech on the basis of phones is also difficult because adjacent phones often overlap in time, "blurring" together in an effect known as co-articulation, which is imposed by the biomechanical constraints of the vocal tract [159].

Greenberg, who has done a wide range of studies in linguistics, neuroscience, and psychoacoustics [158–164], states that there is increasing evidence that the "syllable, rather than the phone, is the basic unit of speech perception." Many studies have demonstrated that slow modulations in the acoustic envelope reflecting the syllable rate are probably just as important as spectral variation [136, 164–167]. Temporal changes in the speech envelope convey information about consonants, stress, voicing, phoneme boundaries, syllable boundaries, and phrase boundaries [164, 168, 169]. Listeners also appear more sensitive to syllabic [170] and phrasal [171] boundaries than those of phonemes [159].

Syllable rate modulations result from the alternating high intensity levels of vowels and low intensity levels of consonants [164]. Most syllables last between 107 and 260 ms [158], which corresponds to a syllable rate between 3.8 and 9.3 Hz. The importance of these modulations has been demonstrated by measuring intelligibility after "smearing" the speech envelope [165]. Removing amplitude modulations between 4 and 16 Hz significantly decreases intelligibility, while removing modulations outside of this range has little effect [172]. If we are interested in estimating syllable rate, a good statistical rule-of-thumb is to use at leave five examples, which would require between 535 ms and 1.3 seconds of audio.

3.2.2 Syllable Rate Estimation

Viemeister [173] accurately modeled the human ability to detect amplitude modulations below 50 Hz with a three stage model: (1) initial bandpass filtering, (2) half-wave rectification, and (3) temporal integration. Tuning the parameters in each stage yielded an initial bandpass filter between 4 kHz and 6 kHz, and a final low-pass cut-off frequency of 65 Hz for temporal integration. Forrest and Green [174] conducted similar experiments and arrived at slightly different parameter settings, but found that the same model fit the psycho-acoustic data quite well. These approaches closely model the feature extraction of a syllable rate speech activity detector (SRSAD) [175] that we were interested in using for our research.

SRSAD's features can be computed several hundred times faster than real-time, and our interest was in increasing the system's robustness to acoustic anomalies while not comprising on speed. The calculation of its features (Figure 3.1) begins by low pass filtering a 500 ms window of squared audio samples to get the frame's envelope. The mean of the envelope is then subtracted and the discrete Fourier transform of the resulting signal is computed. The components representing frequencies between DC and 60 Hz are normalized so they sum to one and are treated as a discrete probability distribution. The first feature is the expected value of the distribution characterizing the frequency of energy concentration. The second feature is the ratio of max to average component probability mass representing the peakedness of the distribution. Since speech has a syllabic rate between 3.8 and 9.3 Hz [158], the expected value of its envelope modulation should be lower than it is for noise and the peakedness should be higher. The window is shifted by increments of 100 ms for the length of the audio to better resolve the times of speech onset and offset. A typical distribution of these features for a large amount of conversational speech collected on landline telephones is shown in the top-left of Figure 3.2.

In general, SAD is a relatively well-understood problem capable of good performance using a variety of features such as pitch, zero-crossing rate, adaptive energy thresholds [144], signal to noise ratio [176], formant shape [177], and wavelet coefficients [178]. However, our goal with acoustic anomaly detection is to decide whether or not an unknown signal has long-term modulation energy distributed in a way that is consistent with conversational speech. We therefore aggregate our estimates

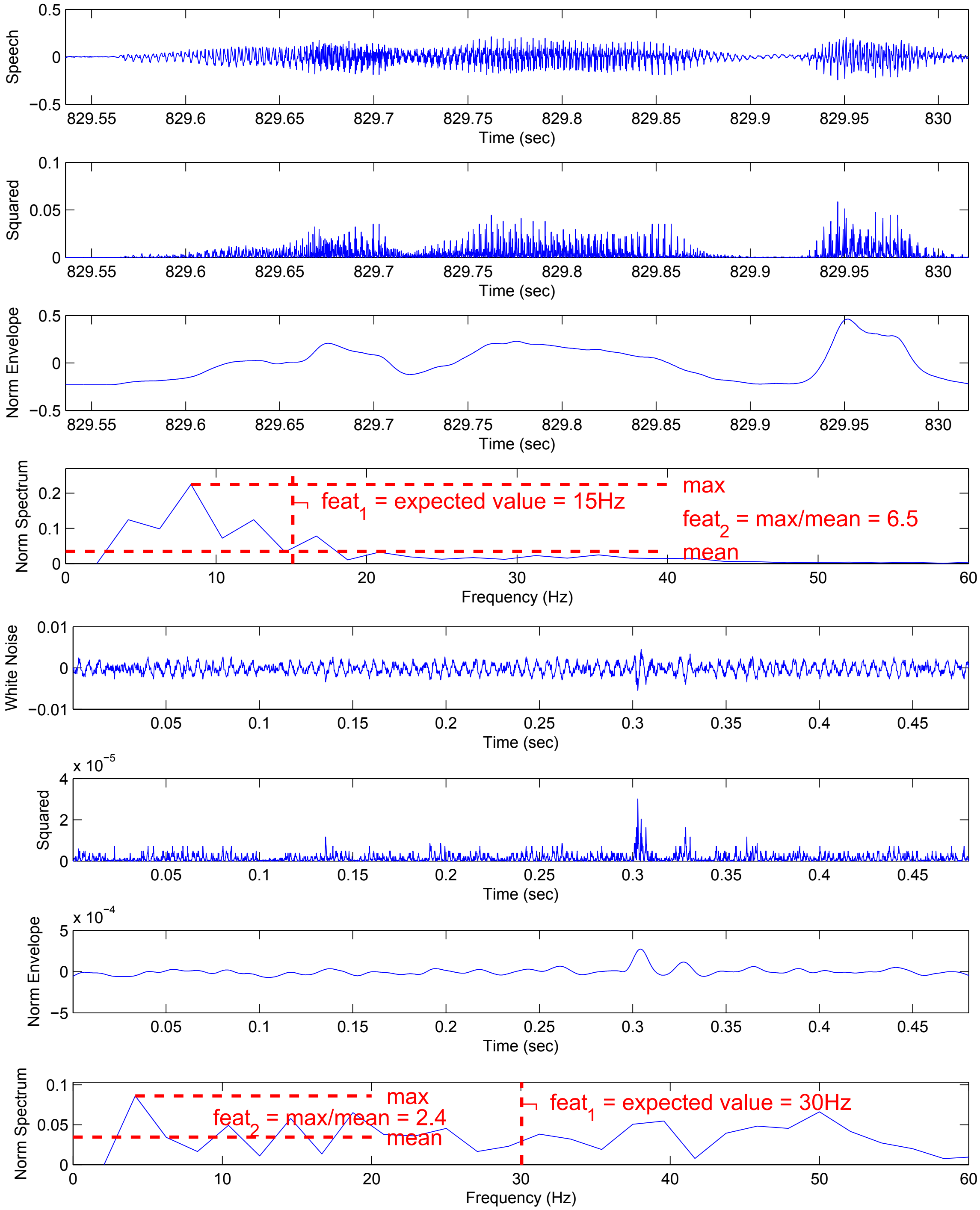

Figure 3.1: Syllable rate feature computations for speech (top four) and white noise (bottom four).

of syllabic rate and estimate the distribution of syllable rate for each five minute audio segment. Then we will decide whether or not each segment is anomalous by comparing it to a model of normal conversational speech.

3.2.3 Evidence Aggregation

Acoustic anomaly detection (AAD) takes the first step of identifying when acoustic conditions have decayed so drastically that statistical inference performed by downstream speech processing systems is likely to be unreliable. We want to distinguish conversational speech from *any anomalous sound* whether or not it was present in our training data. Such a sound could be synthetic or natural, anything from electronic music to animal vocalizations or something as simple as incorrectly decoded audio.

While we are interested in this wider scope of AAD, there is obviously significant overlap with speech activity detection. However, we have reformulated the tokenization task of SAD into a detection task making a single decision for a long acoustic segment. This allows us to aggregate evidence over time and make more informed decisions, but it comes at the cost of temporal localization. We envision using the best of both approaches by operating them in parallel (Figure 1.4). SAD should continue to be used for accurate temporal localization of speech, while AAD can monitor the conditions over longer time spans and make more informed judgements about the reliability of SAD output. For example, when operating on audio files collected from the Web, AAD can be used to decide which files are mostly comprised of conversa-

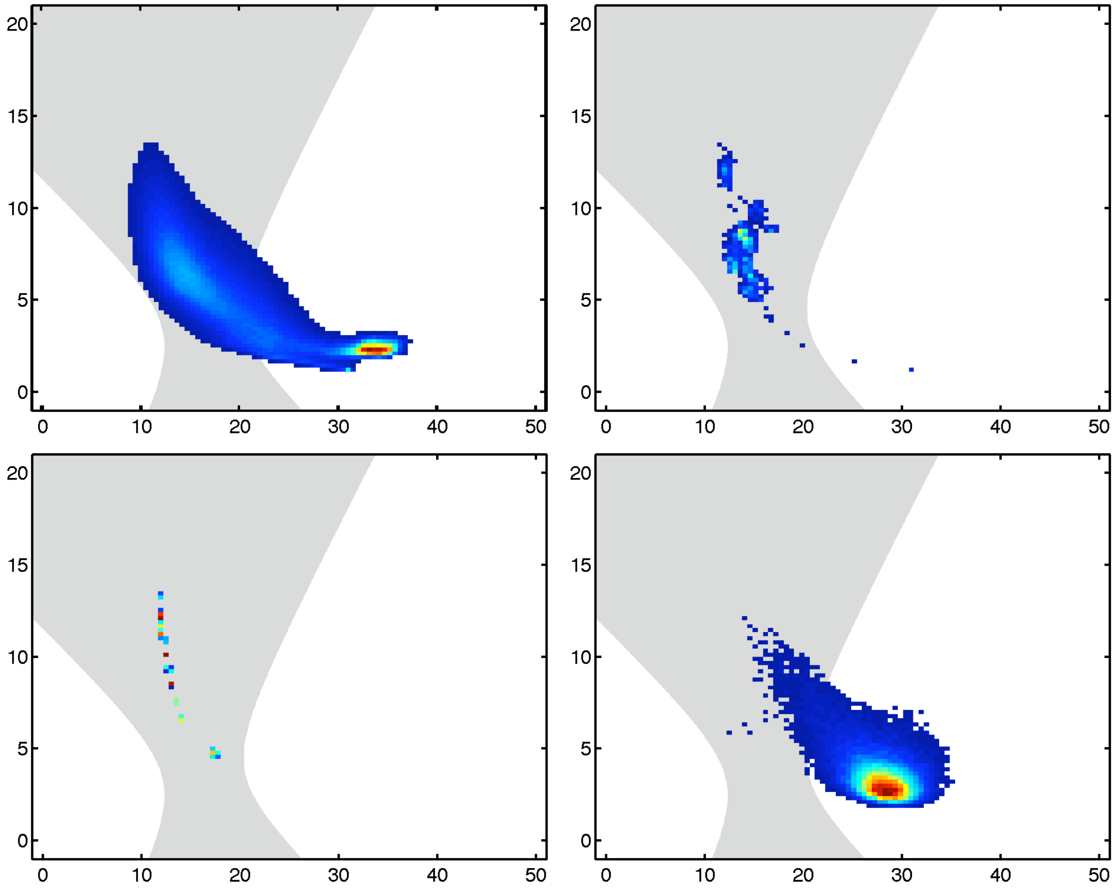

Figure 3.2: Histograms of syllabic rate features for CallHome English $train_{init}$ (top left), international Morse code at 2kHz using a 100ms unit time (top right), dual-tone multiple-frequency (DTMF) signaling keyed on and off for 400ms (bottom left), and a classical violin piece (bottom right) showing that anomalies can occupy the speech region (gray) when using log-likelihood ratio testing for a single Gaussian per class with $p(\text{speech}) = 0.4855$. The x-axis represents the expected value of the frequency of envelope modulation (in Hz) and the y-axis is the ratio of max to average Fourier component of the envelope.

tional speech. SAD can then be used to determine when the speech starts and stops within each of those files.

It is important to note that when training our AAD system, we never rely on any finite set of acoustic anomalies, because doing so would limit our ability to detect new acoustic anomalies in the future. Instead, we aim to model the distribution of acoustic features from normal conversational speech and then assess how anomalous an acoustic test segment is by estimating the divergence between it and normal data.

3.3 Data

The Defense Advanced Research Projects Agency (DARPA) has sponsored several programs in the last decade calling on participants to address the challenge of increasing the robustness of speech processing systems. The Naval Research Laboratory conducted an evaluation of Speech in Noisy Environments (SPINE) [179] with DARPA sponsorship from 2000-2001, which sought to improve word recognition of military-style speech in simulated military environments. Recently, DARPA sponsored the Robust Automatic Transcription of Speech (RATS) program to advance state-of-the-art speech processing in challenging push-to-talk (PTT) communications channels [180, 181]. State-of-the-art systems have demonstrated impressive performance in these and other challenging acoustic conditions [182–184], but in order to achieve these goals, participants typically develop tailored solutions that are trained on very specific communications channels to mirror the testing conditions.

While these and many other data sets were carefully constructed to simulate real applications, they do not represent the broader challenge of speech processing across the wide variety of communications channels available today. When considering large, diverse, multi-genre media like the BBC's television [185] and radio broadcast archives [186], podcasts and other Web audio [187], and YouTube videos [188], simply identifying which files contain speech can be a non-trivial task [189]. During the 2008 election season, Google developed a system to transcribe material posted to YouTube by presidential campaigns in order to make them searchable. Even though this was a small and rather homogeneous slice of YouTube's overall diversity, Alberti et al. [146] noted that their SAD system aggressively removed noisy parts of videos containing conversations with people on the street. They also acknowledged that videos with mismatched recording conditions had much higher error rates compared to the close-talking microphone used in the training data. They confirmed that this task was "much less controlled than a typical DARPA corpus" and other authors report that word recognition error rates on general YouTube videos are higher than 50% [189], which is five times the error rates reported on English broadcast news corpora [190].

Robustness to channel variability and acoustic anomalies is a fundamental challenge in automatic speech processing systems. When presented with difficult acoustic conditions, such as noisy PTT in the DARPA RATS program, researchers insist on similar training material because their algorithms require it to achieve optimal performance. As we discussed earlier, humans do not have this requirement and adapt well to new acoustic conditions. Perhaps it is the matched training and testing paradigm in controlled experiments that guides the research away from robust solutions to real applications.

We consider the unscripted conversations between family and friends in CallHome English [191] to be normal conversational speech. The *train* set consists of 80 two-sided conversations at most 30 minutes long, yielding approximately 77 hours of audio after both conversation sides are separated. We further split this into two random subsets of 54 and 26 conversations referred to as $train_{init}$ and $train_{heldout}$ respectively. The English *devel* and *eval* set each contain another 20 conversations with 18 and 19 hours of audio respectively. Since our algorithms do not require any annotation beyond the fact that each file is known to contain conversational speech, we were not restricted to the subset of audio with associated transcripts.

For testing out-of-domain data, we also experimented with Switchboard-2 Phase III [192] and Switchboard Cellular Part 1 [193]. Both consist of unscripted calls between people who do not know each other where the participants are given a topic for discussion. The former consists of 2,657 five minute land line calls yielding approximately 422 total hours of audio after separating both sides of the conversation. The latter is 250 mostly GSM cellular calls in a variety of environmental conditions, each approximately 6 minutes long, totaling nearly 50 hours of audio. Both have metadata about each speaker and call including some audit information about the amount of echo, background noise, and distortion reported in each.

We also wanted to experiment with wildly anomalous audio totally unlike conversational speech. The easiest data to simulate was incorrectly decoded audio. This was done by taking the μ-law companded CallHome English *devel* files and treating them as if they were a-law, reversed bit-ordered μ-law, reversed bit-ordered a-law, 8-bit linear, and 16-bit linear audio.

Since SRSAD looks for the slowly modulating audio envelope characteristic of speech, it has a tendency to false alarm on tones and noises of short duration and on certain kinds of muzak [194]. In the absence of any corpora more appropriate for anomaly detection, we designed a synthetic corpus that we knew would be challenging for SRSAD. Audacity plug-in effects[1] were used to generate various types of noise such as dual-tone multiple-frequency (DTMF) signaling, Morse code, buzzing, explosions, fires, guitar plucks, surf, tuning forks, and wind. One could naïvely hope that a SAD algorithm could easily label all of these signals as non-speech, but the problem is non-trivial especially when the only algorithms considered are those that have not been specifically trained to recognize these sounds as non-speech. We first reported these results in [195].

[1] http://audacity.sourceforge.net/download/plugins

3.4 Methods

We began by training a Gaussian mixture model (GMM) to characterize typical conversational audio using the CallHome English *train* set. This is similar to the universal background model (UBM) described in [196] except that it includes both speech and non-speech as well as using full covariance matrices because the syllable rate features are highly correlated. Significant deviation from this *normal world model* (NWM) was considered to be indicative of anomalous audio.

The probability of an observation $\boldsymbol{x} \in \mathbb{R}^D$ for a GMM (Section 2.1.2) is

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{i=1}^{m} w_i \cdot \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{3.1}$$

where m is the total number of mixtures, each having a weight w_i, mean vector $\boldsymbol{\mu}_i$, and covariance matrix $\boldsymbol{\Sigma}_i$. The mixture model $\boldsymbol{\theta}$ comprises a set of parameters $(w_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \ldots, w_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$. For a set of n observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^D$, we can estimate the average log probability,

$$\log p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \log p(\boldsymbol{x}_i; \boldsymbol{\theta}), \tag{3.2}$$

by assuming that they are independent and identically distributed.

SRSAD uses quadratic discriminant analysis [197] to separate speech from non-speech by assuming each are generated by a single Gaussian with different covariance matrices. Since we want the NWM to accurately represent the distribution of all the input, we did not impose this limitation on the number of Gaussians. In this chapter, we will report results when using 8 components which we found to be a good trade-off between performance and model complexity. In a scenario with available class conditional models for speech and non-speech, a NWM could also be obtained by merging the available models and normalizing the weights according to the class conditional priors.

Using expectation-maximization (EM) to train GMMs is a process that is susceptible to getting stuck in local maxima and overfitting the training data, both of which we want to avoid. To deal with the former, we begin by randomly shuffling the syllable rate features for all of $train_{init}$ into as many bins as we had audio files. Each bin's data was then used to generate a separate GMM which could be trained relatively quickly. These models were initialized with means equal to randomly selected data points and a covariance matrix set to that of the bin's data normalized by the number of mixtures being trained. This process was repeated 25 times for each bin, and EM iterations were run until convergence. The model amongst all of the random starts for all the bins that gave the highest average log probability to its own bin's data was selected as the initial model.

The initial model was then trained further on all of $train_{init}$ as long as the log probability of $train_{heldout}$ increased. This was done to avoid overfitting the model to the training data to ensure it would generalize well to other data from the same domain.

3.4.1 Average Log Probability Baseline

One would expect that anomalous audio would have a low probability under the NWM representing the typical syllabic rate distribution of conversational speech. To test this hypothesis, we derived a baseline anomaly detector using the average log probability of an input sequence. Here, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is labeled as anomalous if

$$\frac{1}{n}\sum_{i=1}^{n} \log p(\boldsymbol{x}_i; \boldsymbol{\theta}) < \lambda \tag{3.3}$$

where the threshold λ is chosen so the false alarm rate is equal to the miss rate on the test set. This equal error rate (EER) allows us to summarize detection performance with a single number and automatically adapts to very low false alarm rates. In a real application, an *a priori* threshold [198] would be required. This threshold could be chosen using normal data percentiles as shown in Figure 3.3.

3.4.2 Distributional Anomaly Detection

The distribution of syllabic rate features for conversational speech represented by the NWM has a shape distinct from the distributions of other anomalous signals (Figure 3.2). In order to exploit these differences we explored several different approaches to distributional anomaly detection. Audio was considered anomalous if the divergence between the NWM and a test sequence exceeded some threshold λ.

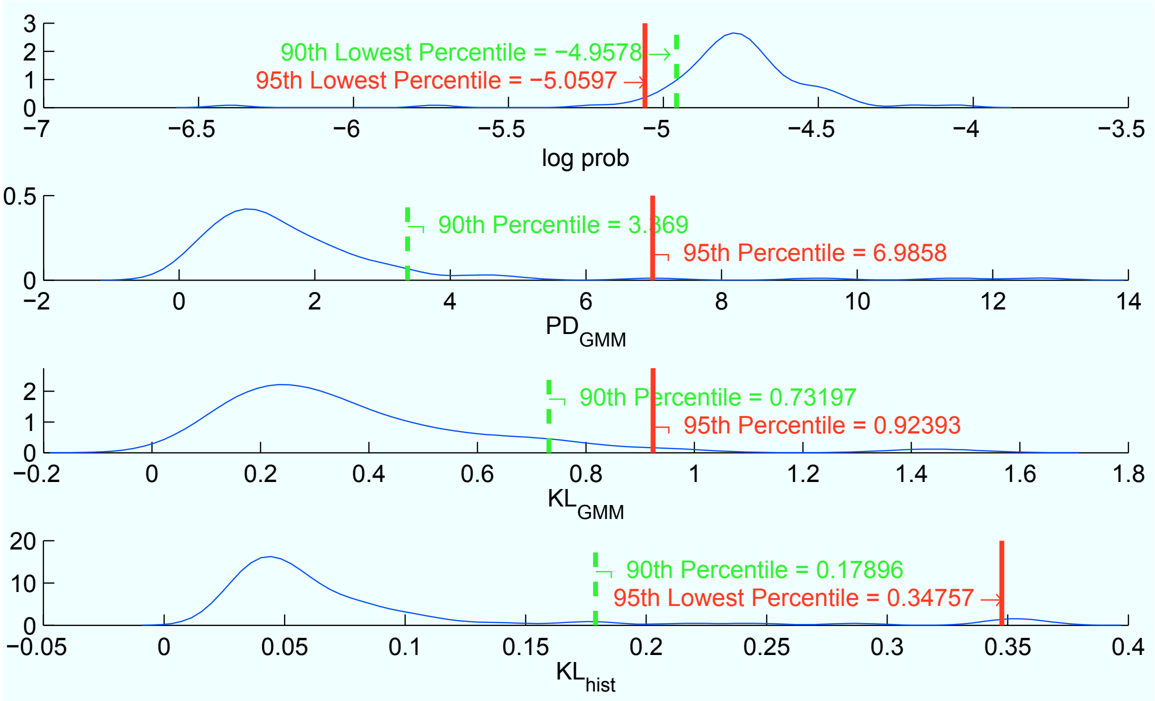

Figure 3.3: Kernel density estimates of various divergences of conversation sides in CallHome English *devel* and *eval* from the NWM trained on CallHome English *train*.

3.4.2.1 Parametric Divergence for GMMs

The first distributional approach we investigated was comparing the GMM parameters

$$\begin{aligned}\boldsymbol{\theta}_{\mathrm{p}} &= \left(w_{\mathrm{p},1}, \boldsymbol{\mu}_{\mathrm{p},1}, \boldsymbol{\Sigma}_{\mathrm{p},1}, \ldots, w_{\mathrm{p},m}, \boldsymbol{\mu}_{\mathrm{p},m}, \boldsymbol{\Sigma}_{\mathrm{p},m}\right) \text{ and} \\ \boldsymbol{\theta}_{\mathrm{q}} &= \left(w_{\mathrm{q},1}, \boldsymbol{\mu}_{\mathrm{q},1}, \boldsymbol{\Sigma}_{\mathrm{q},1}, \ldots, w_{\mathrm{q},m}, \boldsymbol{\mu}_{\mathrm{q},m}, \boldsymbol{\Sigma}_{\mathrm{q},m}\right),\end{aligned}$$

directly using a weighted mean square difference (MSD) heuristic between the mean and covariance parameters,

$$\begin{aligned}PD_{\mathrm{GMM}}\left(p \| q\right) = \sum_{i=1}^{m} &\left(\frac{w_{\mathrm{p},i} + w_{\mathrm{q},\pi(i)}}{2}\right) \times \\ &\Bigg(\alpha \cdot \mathrm{MSD}(\boldsymbol{\mu}_{\mathrm{p},i}, \boldsymbol{\mu}_{\mathrm{q},\pi(i)}) + \\ &\quad (1-\alpha) \cdot \mathrm{MSD}(\boldsymbol{\Sigma}_{\mathrm{p},i}, \boldsymbol{\Sigma}_{\mathrm{q},\pi(i)})\Bigg),\end{aligned}$$

which relies on a mapping function π between the components of each GMM. Rather than derive a correspondance between two independently trained GMMs, we adapt the NWM to each test segment. We use only the unique elements in the symmetric covariance matrix and set $\alpha = 0.5$. Though in principle one could set $\alpha = 1$ to only capture changes in the means, doing otherwise also captures the rare but important changes in the covariance matrices.

The primary change in the GMM parameters of syllabic rate features for conversational audio should only be due to changes in the prior probability of speech for different speakers. This would have the most impact on the mixture weights and we did not want this effect to contribute to the divergence. However, an individual Gaussian's contribution to the divergence should be proportional to its probability mass in the GMM, so we weighted each mixture's contribution to the whole using the average of its values for p and q. We experimented with keeping the mixture weights fixed during adaptation, but this caused the mixtures to move around to redistribute probability mass even when the *shape* of the distribution did not visibly change.

3.4.2.2 KL Divergence Approximation for GMMs

Comparing two densities $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ is often done using the Kullback-Leibler (KL) divergence [199],

$$KL\left(p\|q\right) = \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}. \tag{3.4}$$

This provides a measure of the information lost when using the test sequence $q(\boldsymbol{x})$ to approximate the NWM $p(\boldsymbol{x})$. For $\boldsymbol{x} \in \mathbb{R}^D$, the divergence between single Gaussians, $p(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ and $q(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$, can be computed directly [200],

$$\begin{aligned} KL_{\mathrm{G}}\left(p\|q\right) &= \frac{1}{2}\log\Bigg(\frac{|\boldsymbol{\Sigma}_p|}{|\boldsymbol{\Sigma}_q|} + \mathrm{Tr}|\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q| - D && (3.5) \\ &\quad + \left(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q\right)^T \boldsymbol{\Sigma}_q^{-1} \left(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q\right)\Bigg). && (3.6) \end{aligned}$$

However, there is no such closed form expression between two GMMs. We use an approximation for GMMs suggested by Goldberger et al. [201],

$$KL_{\mathrm{GMM}}\left(p\|q\right) = \sum_{i=1}^{m} w_{p,i} \left(KL_{\mathrm{G}}\left(p_i\|q_{\pi(i)}\right) + \log \frac{w_{p,i}}{w_{q,\pi(i)}} \right), \tag{3.7}$$

which relies on the mapping function π between the components of each GMM derived from adapting the NWM to each test segment. This provides a rigorous estimate of the divergence between the GMMs while remaining computationally efficient and able to scale well to higher dimensional distributions.

The NWM is a good starting point for modeling the test segments of conversational speech, especially if they are from a similar domain. Maximum *a posterior* (MAP) estimation [196] has been used successfully in a similar scenario to adapt GMM parameters of mel-frequency cepstral coefficients (MFCCs) [202] from a large set of unlabeled data to that of a particular speaker, channel, or language. MFCC-based features are typically 39 dimensions and are based on short-term spectral information derived from 20-30 ms segments of speech. MAP adaptation requires that each parameter have a prior distribution which makes an implicit assumption that this adaptation would only be done to something reasonable, like other speech. This is not necessarily the case for anomalies whose distribution could yield dramatically different parameters. Our compromise was to treat the NWM as a starting point for the parameters of a test sequence GMM and continue training using maximum likelihood estimation (MLE). This gave the parameters the necessary freedom to change dramatically for anomalies or adjust only slightly for other normal data. By taking this approach, we could use the trivial mapping function $\pi(i) = i$ for $i = 1, \ldots, m$ between corresponding mixtures of the GMMs.

3.4.2.3 KL Divergence using Histograms

Finally, since we are using features reasonably bounded in $\mathbb{R}^2$, we also estimate the KL divergence between the distributions using regularly spaced fixed-bin histograms,

$$KL_{\text{hist}}\left(p\|q\right) = \sum_{\Delta\boldsymbol{x}\in\mathbb{R}^2} p(\Delta\boldsymbol{x}) \log \frac{p(\Delta\boldsymbol{x})}{q(\Delta\boldsymbol{x})}. \tag{3.8}$$

With this approach we can test the validity of distributional anomaly detection even if GMMs turned out to be poor models for the syllabic rate features or our adaptation strategy is flawed. However, this represents an approximation that would likely suffer from undertraining if scaled to the high-dimensionality of features commonly used in other speech processing tasks.

3.5 Experimental Results

In order to evaluate the performance of these methods as an acoustic anomaly detector, we defined the task of detecting normal audio from the CallHome English *devel* and *eval* sets, and rejecting incorrect encodings (a-law excluded) and synthetic noises. A common method to display possible trade-offs between Type I and Type II errors using various thresholds is with a detection error trade-off (DET) curve (Figure 3.4) [203].

Anomaly detection results are shown in this form in Figure 3.5. The improvement of the distributional anomaly detection strategies over the log probability baseline

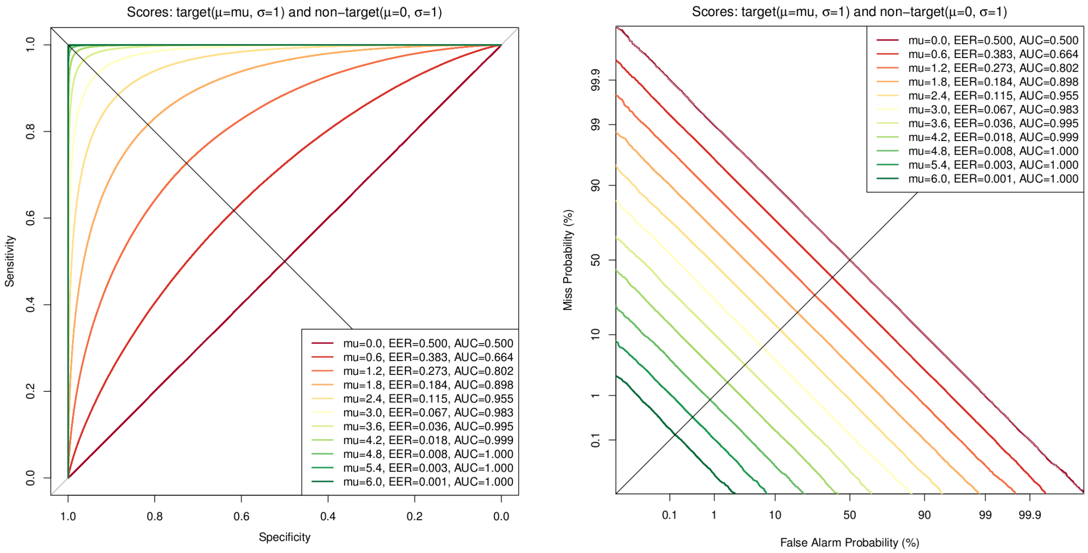

Figure 3.4: Receiver operating characteristic (ROC) curves (left) and detection error trade-off (DET) curves (right) for synthetic data. The DET curves display results in terms of error rates so results in the lower left-hand corner indicate better performance. Its axes are plotted on a normal deviate scale, so a system that outputs Gaussian-distributed scores for targets and non-targets will yield a straight line. The DET curves make for easier visual comparison of multi-system performance, especially at low error rates, unlike the ROC figure where some of the curves lie on top of each other. The black line indicates the the equal error rate (EER) where the miss rate is equal to the false alarm rate.

drove down the equal error rate from 36.3% to 5% using PD_{GMM}, 2.5% for KL_{GMM}, and 0.6% for KL_{hist}.

3.5.1 Anomalies Found in Normal Audio

The first area we explored was audio on either side of the divergence spectrum from sources we assumed would contain only normal audio. Since the NWM characterized CallHome English *train*, a natural place to look was in the *devel* set. As expected, the distributions of syllabic rate features for conversation sides with low divergences looked very similar to the NWM, but we were surprised at how different the distributions appeared for segments with high divergences (Figure 3.6).

Examining the spectrograms and energy envelopes of normal segments (Figure 3.7) revealed that most of the periods of non-speech had a low but steady amount of nearly white background noise. The spectrograms of the anomalous segments during periods of non-speech (Figure 3.9) were not as clean, and all looked very different from each other.

The first anomalous cut, en_4576_1 (Figure 3.8), had periods of background noise bursting out of almost no energy (Figure 3.9). The second anomaly, en_4686_2 (Figure 3.6), was extremely noisy with an SNR of 0.75 dB (Figure 3.9) compared to the average of 16.4 dB for the *devel* set estimated using NIST's *stnr* tool. It also had some signaling shown in the spectrogram that sounded similar to DTMF although it occasionally had three simultaneous tones. Other segments labeled as anomalous

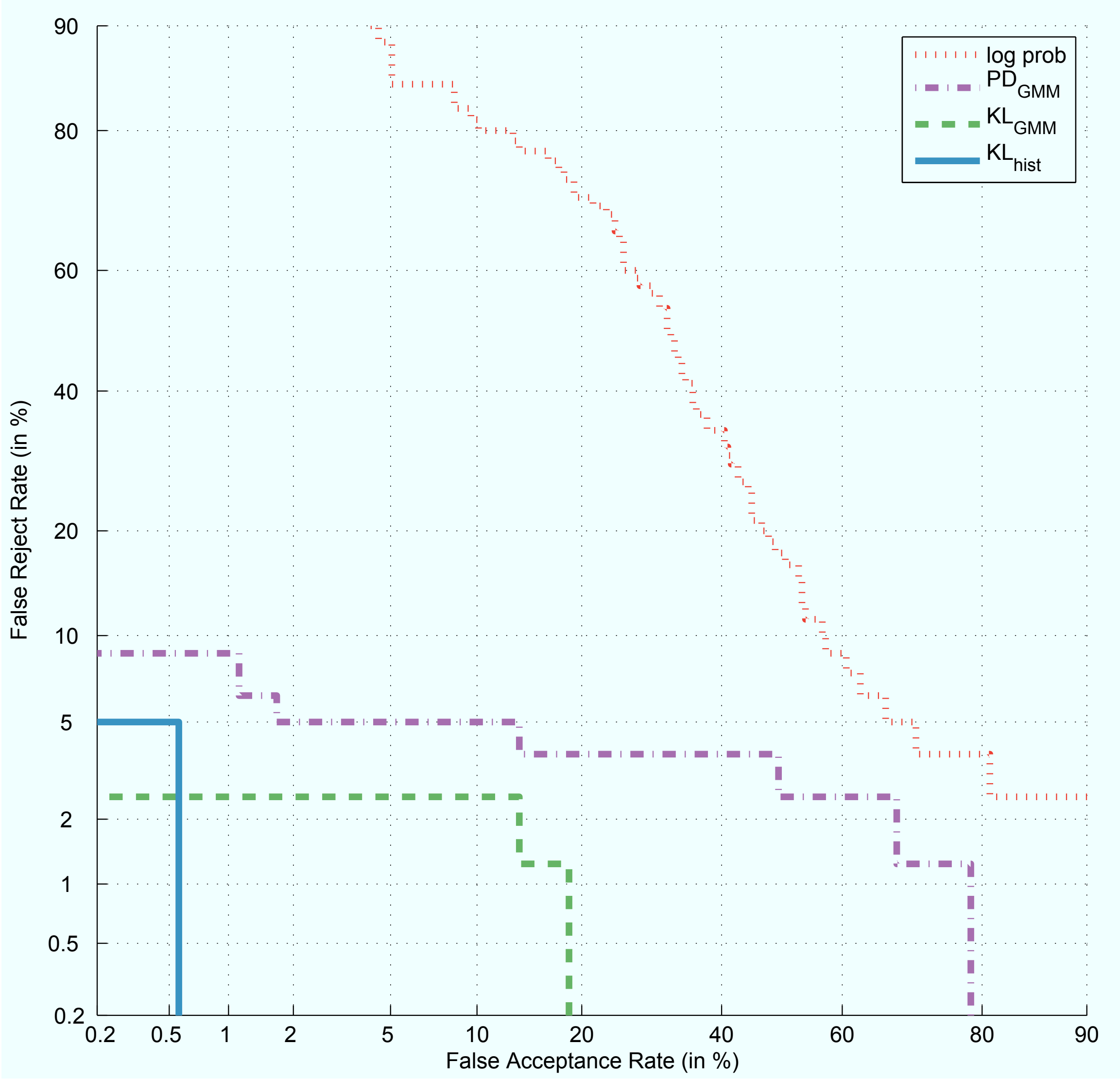

Figure 3.5: Normal audio detection error trade-off curves using various divergences from the NWM with CallHome English *devel* and *eval* conversation sides as targets and incorrectly decoded audio and noises as non-targets. The incorrectly decoded audio was performed by falsely assuming that μ-law data was reversed bit-ordered μ-law, reversed bit-ordered a-law, 8-bit linear, and 16-bit linear.

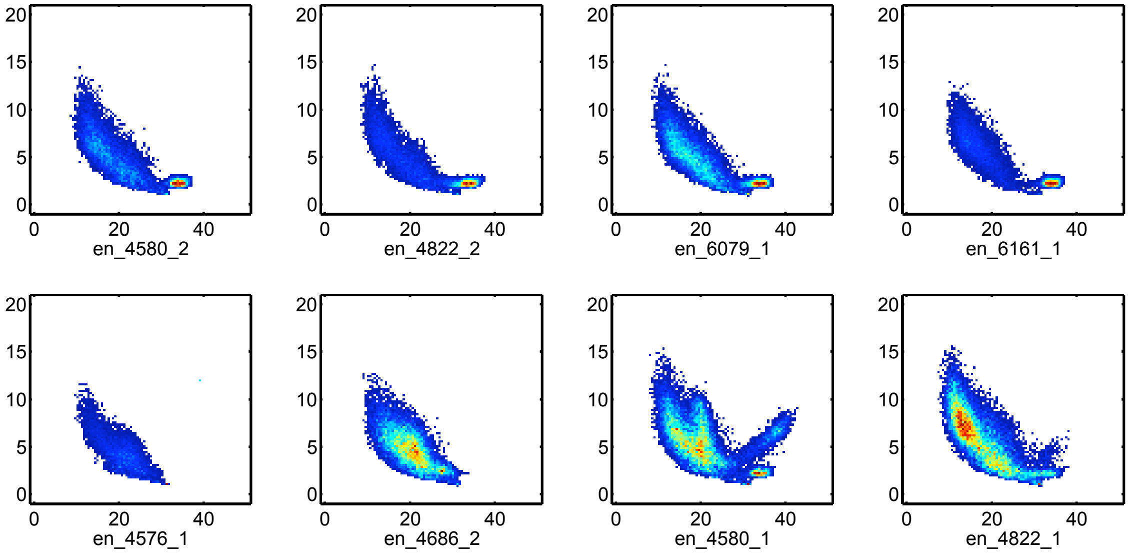

Figure 3.6: Histograms of syllabic rate features for conversation sides in CallHome English *devel* with low divergences (top four) and high divergences (bottom four) from the NWM.

were not quite as noisy, but often had bursts of energy during periods of non-speech (Figure 3.9). Those present in en_4580_1 were concentrated at 120 and 180 Hz which gave its distribution an entirely new region of mass, while those in en_4822_1 appeared to be bleed-over from the other side of the conversation. Both resulted in a modulating energy envelope during periods of non-speech that contributed to the increased divergence of their syllable rate feature distributions from the NWM.

It is important to note that segments with a low divergence from the NWM could have brief periods of anomalous audio that could go undetected. Detecting these short anomalies is certainly worthy of exploration, but that is outside the scope of this work. We focus on the long-term anomalies which are usually due to channel irregularities and can be extremely detrimental to the performance of speech processing systems.

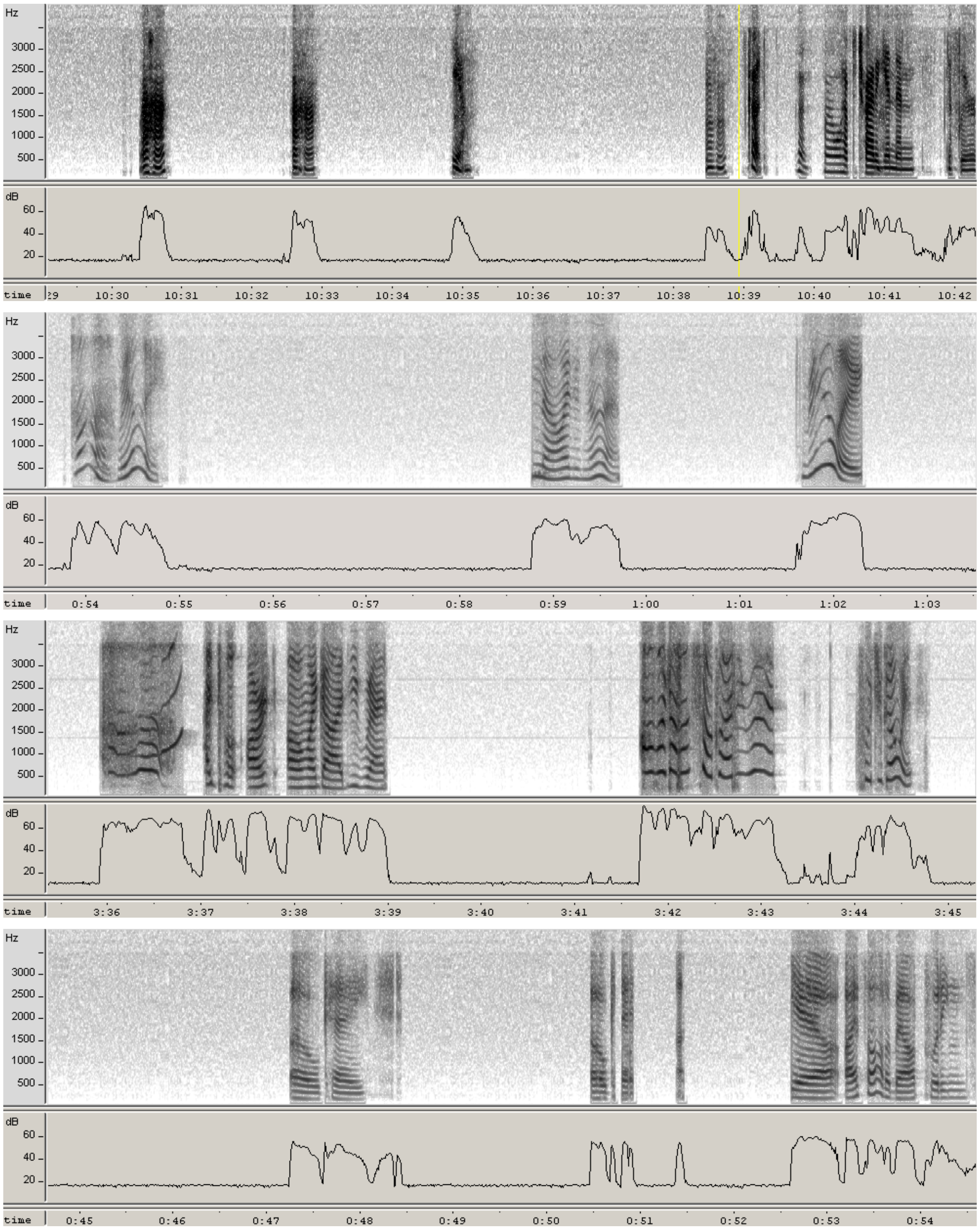

Figure 3.7: Spectrogram snippets of normal segments from en_4580_2, en_4822_2, en_6079_1, and en_6161_1 from top to bottom in CallHome English *devel* with low syllabic rate distribution divergences from the NWM. Note the low-level, stationary background noise indicative of normal conversational audio.

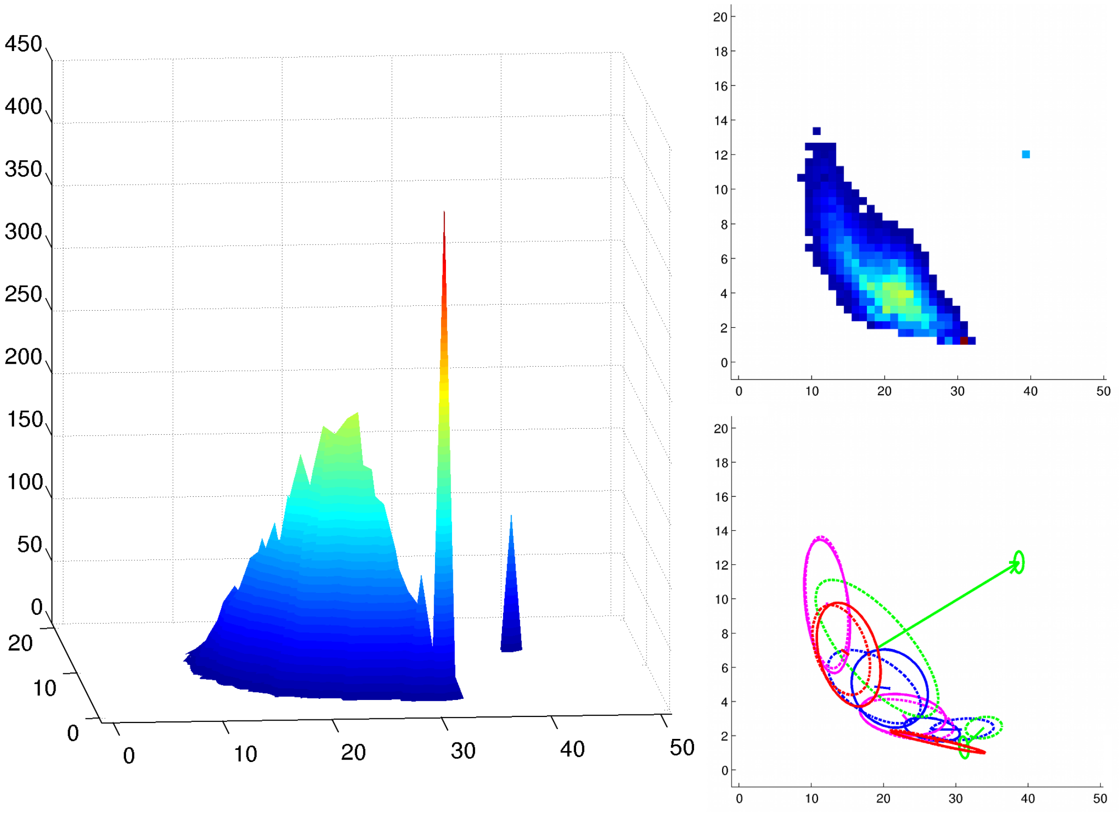

Figure 3.8: Syllable rate feature distributions (left and top-right) of en_4576_1 in CallHome English *devel* with a high divergence from the NWM, especially for the estimates based on GMM adaptation (bottom-right).

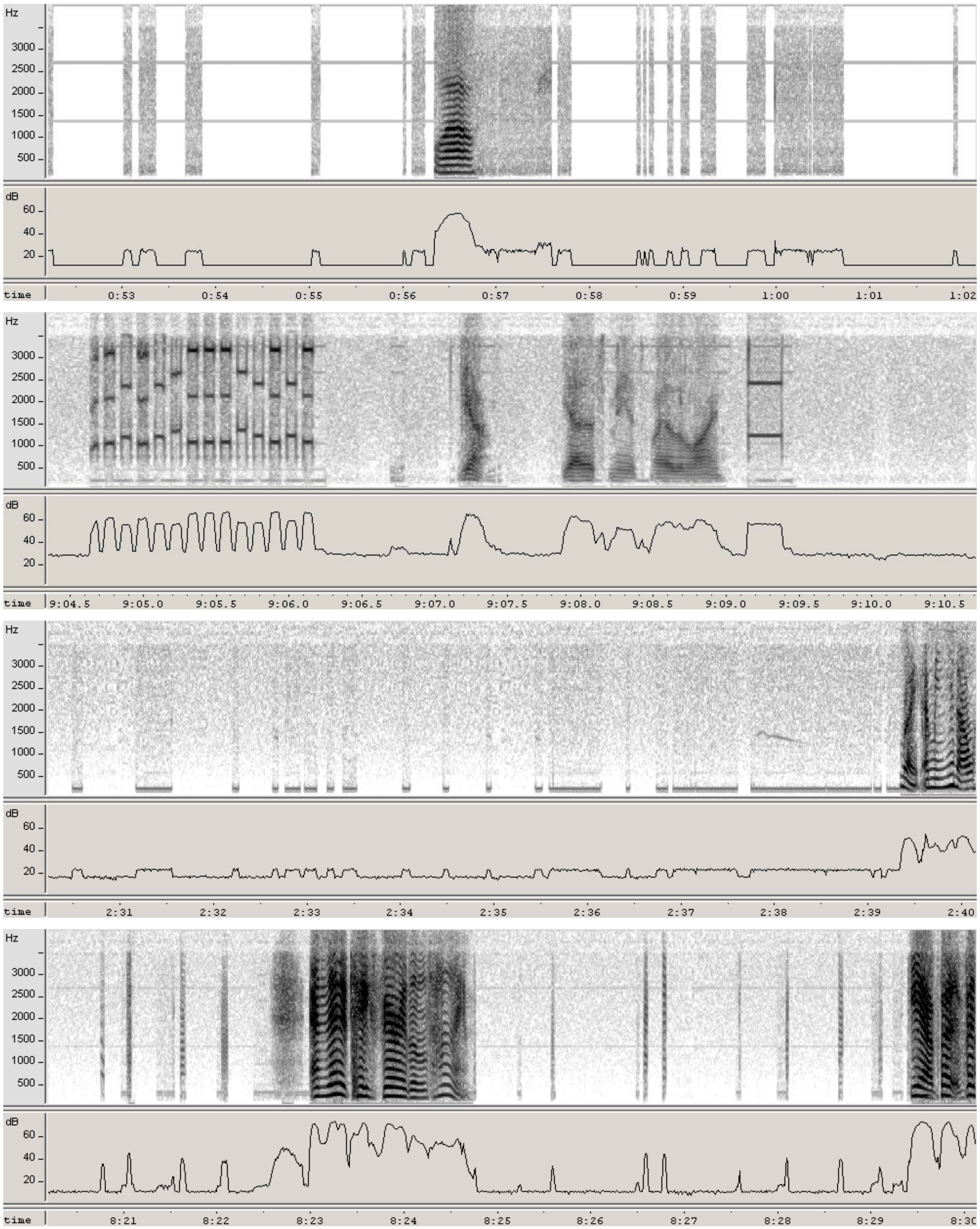

Figure 3.9: Spectrogram snippets of anomalous segments from en_4576_1, en_4686_2, en_4580_1, and en_4822_1 from top to bottom in CallHome English *devel* with high syllabic rate distribution divergences from the NWM.

3.5.2 Normal Audio Found Amongst Anomalies

When creating the incorrectly decoded audio files, we treated the μ-law data as if they were each of five other possible encodings commonly used for speech and examined the distribution of their KL_{GMM} divergences (Figure 3.10) expecting them all to be anomalous. While the distribution of a-law divergences is slightly shifted away from the origin, the bulk of the data fell within the 95th percentile of normal audio shown as a solid red line. After listening to some of these files, we discovered that most of the speech, while distorted, was still intelligible and could be easily distinguished from non-speech by looking at the energy envelope. So from the perspective of an acoustic anomaly detector based on syllable rate estimates, this type of incorrectly decoded audio was not that unusual after all.

Regardless of what decoding was performed there was always a difference in the energy envelope between the periods of speech and non-speech. The distortion was unique when assuming the data was 16-bit linear since the signal was also effectively downsampled by 2 without any regard for aliasing. Also visible in Figure 3.10 is the slow but steady migration of the divergences for each different incorrect decoding away from the origin. These distortion-consistent divergences suggest that our approach could also be used as a measure of how speech-like the energy envelope is for a set of audio on a continuous scale.

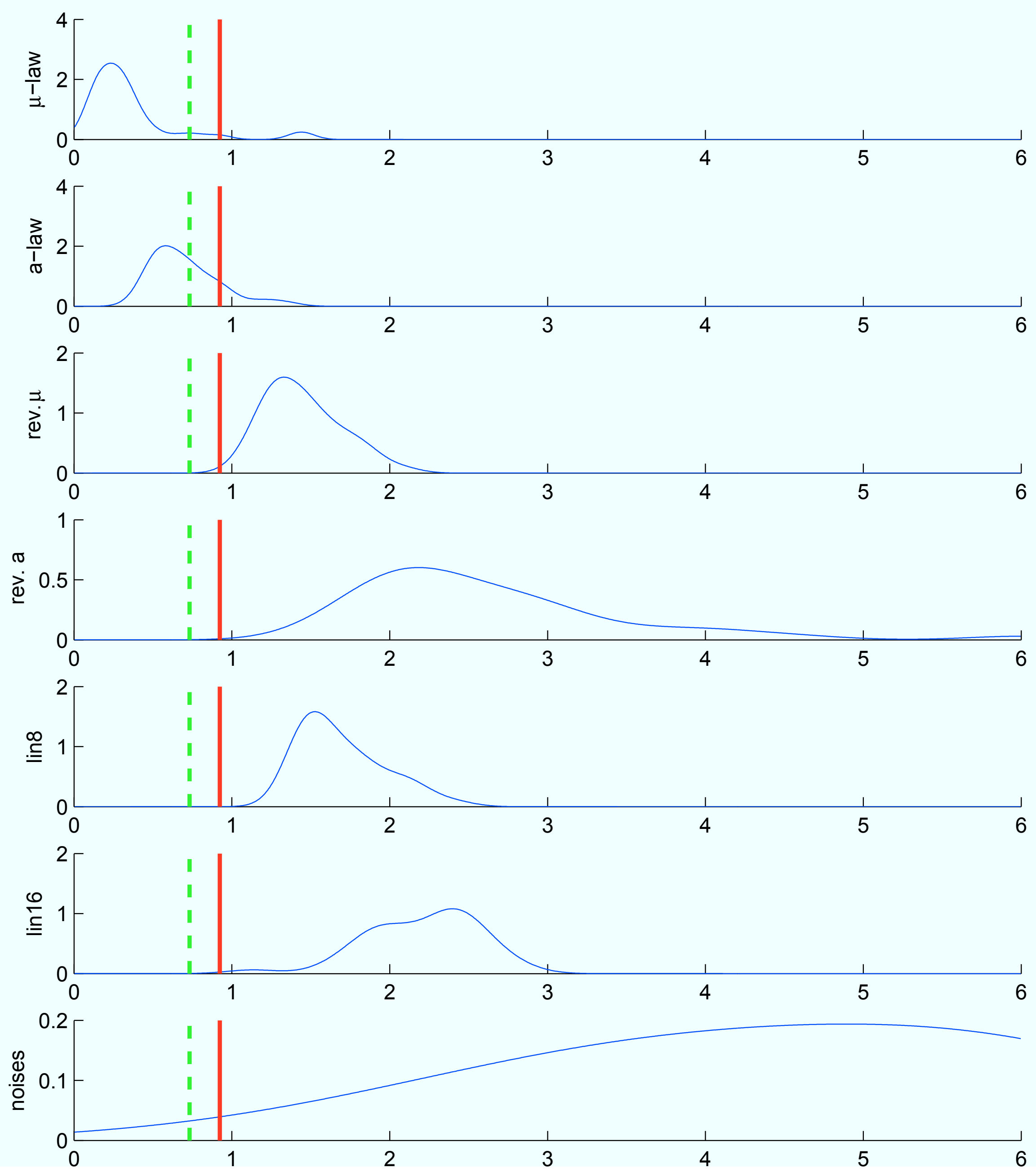

Figure 3.10: Kernel density estimates of 8 mixture KL_{GMM} divergences from the NWM when μ-law data is correctly decoded (top), incorrectly decoded as a-law, reversed-bit ordered μ-law, reversed-bit ordered a-law, 8-bit linear, and 16-bit linear. The divergences of the anomalous noises (bottom) are as large as 106, but the axis is kept short for ease of comparison. Normal percentile demarcations drawn from in-domain data seen in Figure 3.3.

3.5.3 Bimodal Distribution of Switchboard-2

While we did not have any transcripts or speech activity labels for Switchboard-2 Phase III, we were interested to see how the distribution of divergences for files would appear for a corpus similar to CallHome. Both are landline collections, although they differ in the amount of familiarity between the participants. We were surprised to find that the files clumped into two distinct groups with similar divergences for the methods employing GMM adaptation (Figure 3.11). To explore this further, we divided the set of files into two sets depending on whether or not their parametric divergence was less than 9.2. Even after pooling all of the features from all the files on either side of this split, we observed a dramatic difference between the feature distributions (Figure 3.12). Shown in the same figure are example GMM adaptations for files in each mode, depicting the component that moves to the origin for the group with a larger parametric divergence.

While this would not provide a challenge to SRSAD since the obvious heuristic of no energy being non-speech is employed before the features are even calculated, it is worth mentioning since this was discovered automatically. A similarly drastic change was observed in the most anomalous file, en_4576_1, of CallHome *devel* using the KL_{GMM} divergence. There is enough probability mass at $38.7, 12.2$ (Figure 3.8) to draw one of Gaussians to its location and make it anomalous even to the histogram based divergence.

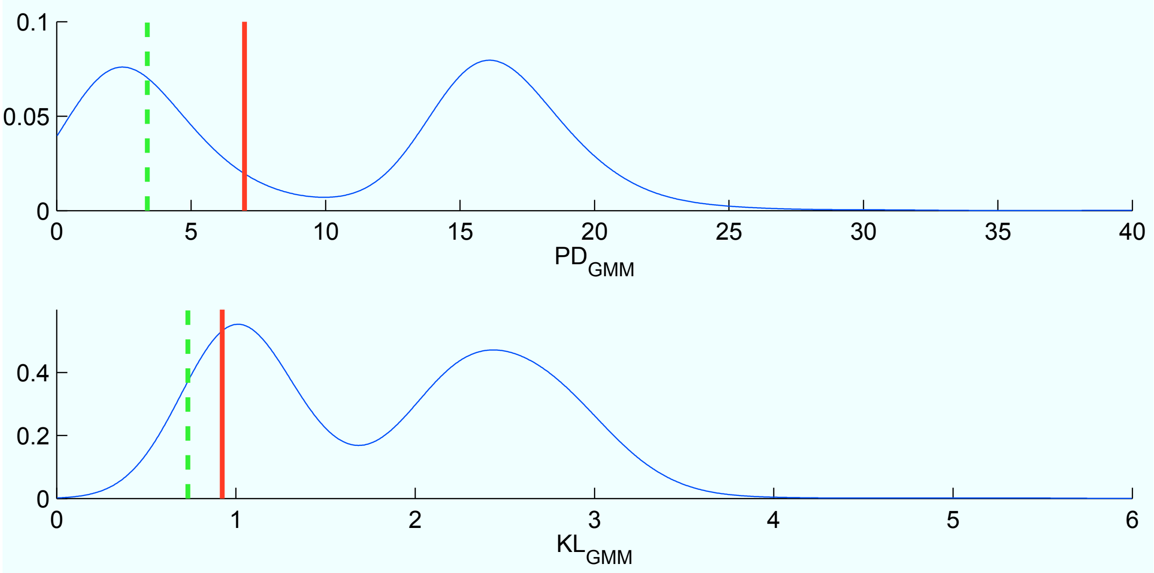

Figure 3.11: Kernel density estimates of 8 mixture PD_{GMM} and KL_{GMM} divergences of conversation sides in Switchboard 2 Phase III from the CallHome English NWM showing an unexpected bimodal distribution. Normal percentile demarkations drawn from in-domain data seen in Figure 3.3.

3.6 Conclusion

After carefully constructing a model of the normal world from the perspective of a speech activity detector, we presented results of explorations into files that were found to be somewhat anomalous in CallHome English and also reported findings of fairly normal audio discovered unexpectedly in a variety of incorrectly decoded files. We also showed that distributional anomaly detection significantly outperformed the log probability baseline. This is perhaps not surprising since several anomalies occupy the same regions of the feature space as conversational speech, which is why they cause false alarms and thus would have a high likelihood under a NWM. Similarly, incorrectly decoded audio is often noise-like with features similar to those during

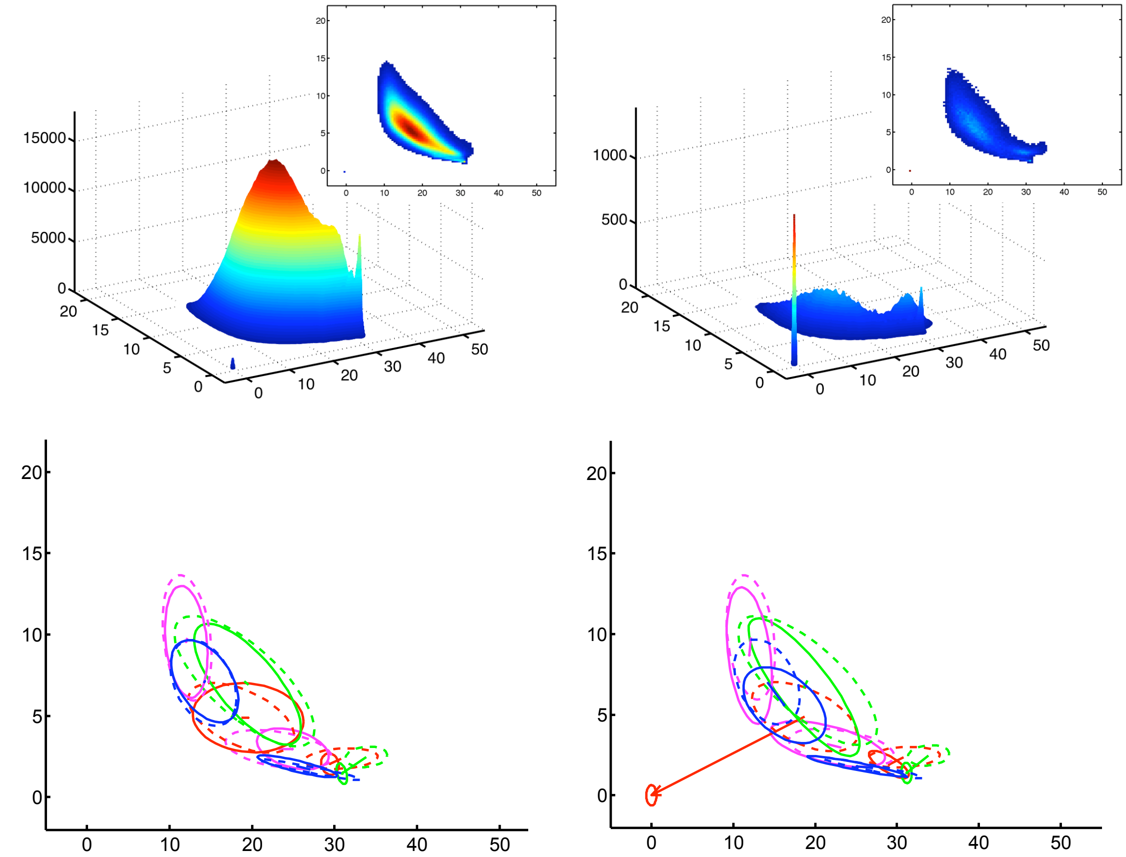

Figure 3.12: Syllable rate feature histograms (top two) drawn from each mode of the divergences seen in Figure 3.11. All the features from conversation sides with $PD\,(\mathrm{NWM}\|q) \leq 9.2$ are on the left and all the features from conversation sides with $PD\,(\mathrm{NWM}\|q) > 9.2$ are on the right. Below each are characteristic GMM adaptations from each mode for sw_30599_1 (left) and sw_32455_1 (right).

normal regions of non-speech. The obvious difference is the spread and shape of the syllable rate distributions for conversational audio compared to the anomalies; a difference which is better captured by estimating the divergence between statistical populations.

Our distributional approach provides self-reinforcement in the partially supervised detection paradigm by aggregating evidence over time, an approach which is somewhat analogous to visual perception where the brain receives a stream of information when observing a single object resulting from movements such as eye saccades and head shifts [204].

Chapter 4

Acoustic Anomaly Detection via Unsupervised Accommodation Learning from Contaminated Data using Histograms

In Chapter 3, we presented several distributional approaches to acoustic anomaly detection by training a model on normal data and estimating the divergence between it and other input. Now, we reformulate the problem into unsupervised accommodation learning and allow for anomalous contamination of the training data. In this chapter, we will explore this problem using histograms since they performed well in the partially supervised setting of Chapter 3. We first reported these results in [205].

4.1 Data

One challenge in speech processing is coping a large quantity of data. While many speech tasks use spectral features computed every 10 ms resulting in dozens of dimensions, we used two features from a syllable rate speech activity detector (SRSAD) [175] computed every 100 ms. Since speech has a syllable rate between 3.8 and 9.3 Hz [158], the frequency of its envelope modulation is different from that of white noise. Using a sliding half-second window of audio, SRSAD computes both the expected value of this modulation frequency and an estimate of its power (Figure 3.1). We modeled the distribution of this two-dimensional sequence as a set of independent observations.

We expanded on the set of synthetic anomalies from Chapter 3 that are known to be problematic to SRSAD, such as tones and noises of short duration and certain kinds of muzak [194]. The set of anomalies used in this chapter is comprised of 50 examples of each of the following: DTMF sequences, morse code, MIDI tones, MIDI songs, and various telephony noises. The MIDI songs were downloaded from the MIDI Database[1] and have a median length of 3.5 minutes. Telephony noises were obtained from FindSounds[2] using the following search terms: *busy signal*, *cell phone*, *dial tone*, *fax*, *keyboard*, *modem*, *off-hook*, *phone*, *printer*, *ringing*, and *typing*. Since some of these noises were of short duration, the audio was repeated until each was at least 5

[1] http://mididb.com

[2] http://findsounds.com

minutes long. Half of the examples of each anomaly type were randomly selected for testing and the other half were reserved for possible use as training contamination.

We again used the CallHome English corpus [191] to represent normal audio. Each conversation side in the *train* set was divided into 5 minute segments, and 250 of the total 918 were randomly selected for training. This enabled us to experiment with contamination percentages up to 33% when using all 125 anomalous segments, which we felt was adequate since "most estimators are known to fail when the fraction of outliers is greater than $\frac{1}{D+1}$, where D is the dimension of the data" [206]. The English *eval* set was similarly divided into 5 minute segments yielding 226 for testing. Since we will not use labeled training data in this chapter, we were not restricted to the subset of audio with associated transcripts. When testing on less than 5 minutes of audio, we will randomly select one continuous section of the desired length from each 5 minute segment. We did not investigate varying the amount of data used for training.

4.2 Histogram Methods

Since the syllable rate features were reasonably bounded in $\mathbb{R}^2$, we used histograms to model their distribution. This non-parametric density estimate (Section 2.1.3) makes fewer assumptions about the data than parametric or semi-parametric models. Histograms can use adaptive- or fixed-width bins. Adaptive binning can result in a lower error between the feature vectors and bin centroids, but it also eliminates many computationally efficient histogram dissimilarity measures [207]. We therefore chose not to use adaptive binning because we were willing to trade some accuracy for efficiency and increased robustness (Section 1.2.4).

Techniques using fixed bins come in two varieties: Either the feature space is divided up into regular intervals or the bin locations are derived from pre-clustering the data. The latter has been used extensively in image retrieval applications where a fixed database of images is being searched [208]. We used regular intervals to characterize the mostly normal training data since we could not guarantee that anomalies are present when the bin locations are derived. Our goal with unsupervised anomaly detection is recognizing unexpected anomalies long after training is completed, and we therefore did not consider it advantageous to use a set of bin locations derived from a fixed set of mostly normal training data.

In Section 2.1.3, we defined a non-parametric histogram as a set of parameters $\boldsymbol{\theta} = (\boldsymbol{b}_1, c_1, \ldots, \boldsymbol{b}_m, c_m)$ comprised of bin centroids $\boldsymbol{b}_j$ with corresponding counts

$$c_j = \sum_{i=1}^{n} \mathbb{1}_{\{j\}} \left(\arg\min_k d(\boldsymbol{b}_k, \boldsymbol{x}_i) \right) \tag{4.1}$$

using the Euclidean distance d for samples $\boldsymbol{x}_i$ for $i = 1, \ldots, n$. We convert this to a density estimate using add-one smoothing,

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{c_j + \frac{1}{m}}{\left(\sum_{k=1}^{m}(c_k)\right) + 1}, \tag{4.2}$$

where $j = \arg\min_k \; d(\boldsymbol{b}_k, \boldsymbol{x})$ for a histogram with m cells. Figure 4.1 shows examples

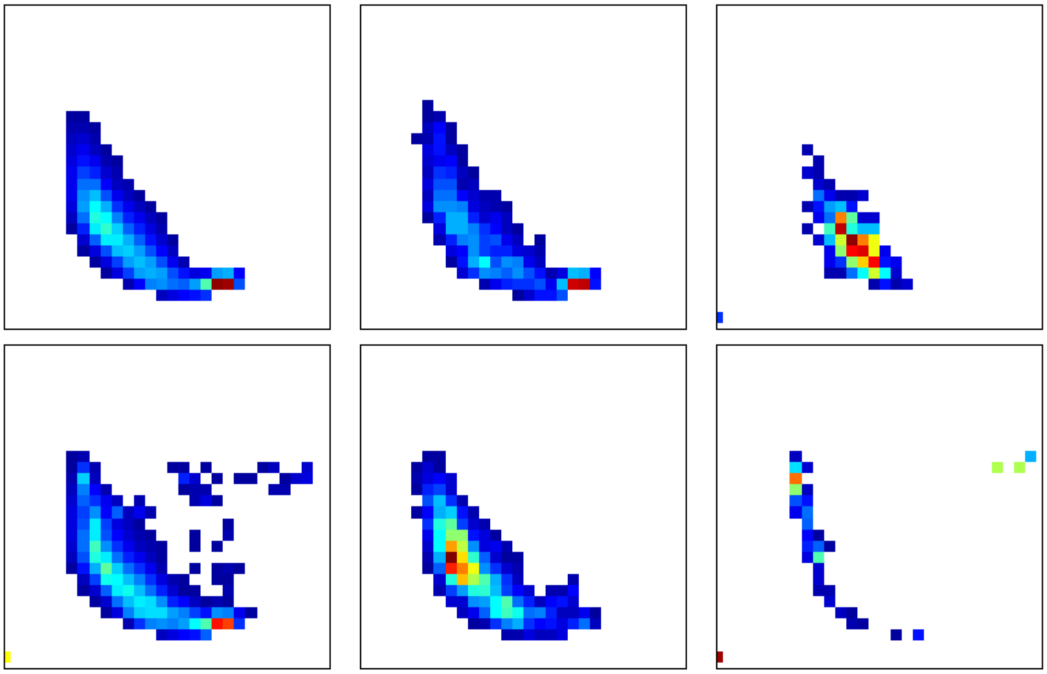

Figure 4.1: Histograms with 30 bins per axis for 0% contaminated training data (top left), 33% contaminated training data (bottom left), 5 minutes of CallHome English 5888 side 2 (top middle), 5 minutes of CallHome English 6825 side 2 (bottom middle), MIDI song (top right), and Morse code (bottom right). Those four test segments have histogram-based KL divergences of 0.13, 0.45, 5.83, and 9.89, respectively, to the uncontaminated training data.

of this for our mostly normal model (MNM) with 0% and 33% contamination along with histograms of normal and anomalous segments.

As in Chapter 3, if the dissimilarity between a histogram of a test sequence and the MNM exceeds a threshold λ, the test sequence is labeled as anomalous. The threshold λ is again chosen so that there is an equal error rate (EER) between misses and false alarms.

4.2.1 Log Likelihood Baseline

We first present a baseline anomaly detector using a histogram for the mostly normal model (MNM) to estimate the average log likelihood of a test sequence. In this case, a test segment $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is labeled as anomalous if

$$\frac{1}{n}\sum_{i=1}^{n} \log p_{\mathrm{MNM}}(\boldsymbol{x}_i) < \lambda. \tag{4.3}$$

While estimating the likelihood of individual feature vectors allows decisions to be made quickly with a single observation, the distribution of features over time provides additional evidence for deciding between normal and anomalous audio.

4.2.2 Distributional Anomaly Detection

As we saw in Chapter 3, the distribution of syllabic rate features for conversational speech has a shape that is distinct from the distributions of other anomalous signals. In this chapter, the MNM will be trained on conversational speech and some unknown amount of anomalous contamination. As seen in Figure 4.1, with as much as 33% contamination, the MNM still appears to be more similar to other segments of speech than it is to anomalous test segments. We will exploit these differences using a variety of techniques to compare histogram-based density estimates of the MNM, $p(\boldsymbol{x})$, and each test segment, $q(\boldsymbol{x})$.

4.2.2.1 Information Theory Divergences

Shannon formalized information theory as a study of communication and channel capacity in the presence of noise [209]. Kullback and Leibler followed by generalizing the concept of information in their study of the "statistical problem of discrimination" between distributions [199], the results of which we were interested in using for anomaly detection.

- **Jensen-Shannon (JS) Divergence**: The JS divergence,

$$JS\left(p\|q\right) = \frac{1}{2} KL\left(p\|g\right) + \frac{1}{2} KL\left(q\|g\right), \tag{4.4}$$

  is a symmetrization of the KL divergence (Equation 3.8) proposed by Lin [210], where $g(\boldsymbol{x}) = \frac{1}{2}p(\boldsymbol{x}) + \frac{1}{2}q(\boldsymbol{x})$, which adds numerical stability and bounds the divergence between 0 and 1.

4.2.2.2 Minkowski-form Metrics

- **$\boldsymbol{L_1}$, $\boldsymbol{L_2}$, and $\boldsymbol{L_\infty}$ Distances**: To compare two histogram density estimates, $p(\boldsymbol{x})$ as the MNM and $q(\boldsymbol{x})$ a test segment, we computed their L_1, L_2, and L_∞ distances

$$L_r\left(p\|q\right) = \left(\sum_{j=1}^{m} |p(\boldsymbol{b}_j) - q(\boldsymbol{b}_j)|^r\right)^{\frac{1}{r}}. \tag{4.5}$$

  The L_1 distance is the sum of the absolute cell differences and has been used to compute color dissimilarity between images [211]. L_2 is the sum of the squared differences and L_∞ is the max difference, which has been used to compute texture dissimilarities [212].

4.2.2.3 Test Statistics

We also used three statistics to test the null hypothesis that the test segment was generated from the same probability distribution as the MNM.

- $\boldsymbol{\chi^2}$ **Test Statistic**: For the Chi-squared statistic, we modified Puzicha et al.'s proposal [213] for histogram-based image retrieval,

$$\chi^2\left(p\|q\right) = \sum_{j=1}^{m} \frac{\left(q(\boldsymbol{b}_j) - p(\boldsymbol{b}_j)\right)^2}{p(\boldsymbol{b}_j)} \tag{4.6}$$

  to test if $q(\boldsymbol{x})$ differed from the mostly normal $p(\boldsymbol{x})$.

- **Kolmogorov-Smirnov (KS) Test Statistic**: The KS statistic is defined as the maximal difference between one-dimensional empirical cumulative distribution functions. A histogram-based approximation was proposed by Geman et al. [214] for grayscale boundary detection. Marginal distributions are often used in a multidimensional setting, but we take a different approach. The cells from the two dimensional histograms are ordered by descending probability mass in the MNM and cumulatively summed to obtain cumulative distribution functions, P and Q. We can then simply compute the test statistic

$$KS\left(p\|q\right) = \max_j \left|P(\boldsymbol{b}_j) - Q(\boldsymbol{b}_j)\right|. \tag{4.7}$$

- **Cramér-von Mises (CvM) Test Statistic**: To compute the CvM statistic, we use the same strategy of converting two-dimensional histograms into a one-dimensional cumulative density function. This test statistic

  $$CvM\,(p\|q) = \sum_{j=1}^{m} \left(P(\boldsymbol{b}_j) - Q(\boldsymbol{b}_j)\right)^2 \tag{4.8}$$

  is then straightforward to estimate with histograms.

4.3 Experimental Results

4.3.1 Model Selection

Model complexity has a substantial impact on anomaly detection performance, especially for short test segments. In Figure 4.2, we show the performance of each approach on 30 second segments as the number of histogram bins was varied when training with 33% anomalous contamination. The interquartile range of the EER was derived from random resamplings of the test set via statistical bootstrapping [215]. The log likelihood and L_∞ distance made their fewest errors when using only 5 bins per axis, the χ^2 statistic when using 10 bins per axis, and the L_2 distance when using 20 bins per axis. The L_1 distance and information theory divergences achieved their optimum performance with 30 bins per axis, and the Kolmogorov-Smirnov (KS) test and Cramér-von Mises (CvM) criterion did slightly better using 50 bins per axis. These histogram bin sizes were used for the remainder of the results as we varied other parameters.

This sensitivity to bin size is similar to the bias-variance trade-off in parameter estimation and optimizing for it is therefore important when engineering statistical systems. When the model complexity is too low we can accurately estimate the cell probabilities, but the restricted model space is biased away from the true distribution and performance suffers. As the bin sizes decreases the model space expands, yielding the ability to better model the true distribution. However, when the bins become too small the increased variance in probability estimates eclipses modeling power, again leading to a decrease in performance.

When we tested on the entirety of the 5 minute segments (Figure 4.3), the models performed well for a much wider range of complexities. With more data we were able to accurately estimate more parameters, which lead to better performance when using higher complexity models.

4.3.2 Test Segment Length

To further explore the effect of test segment length, we evaluated anomaly detection performance when training with 33% contamination as we varied the length of the test segments from 5 seconds to 5 minutes (Figure 4.5). Some general rankings of the various approaches began to emerge, especially for tests using longer segments. The L_1 distance, χ^2 statistic, and both information theory divergences significantly outperformed the other methods when testing on segments longer than 15 seconds. The

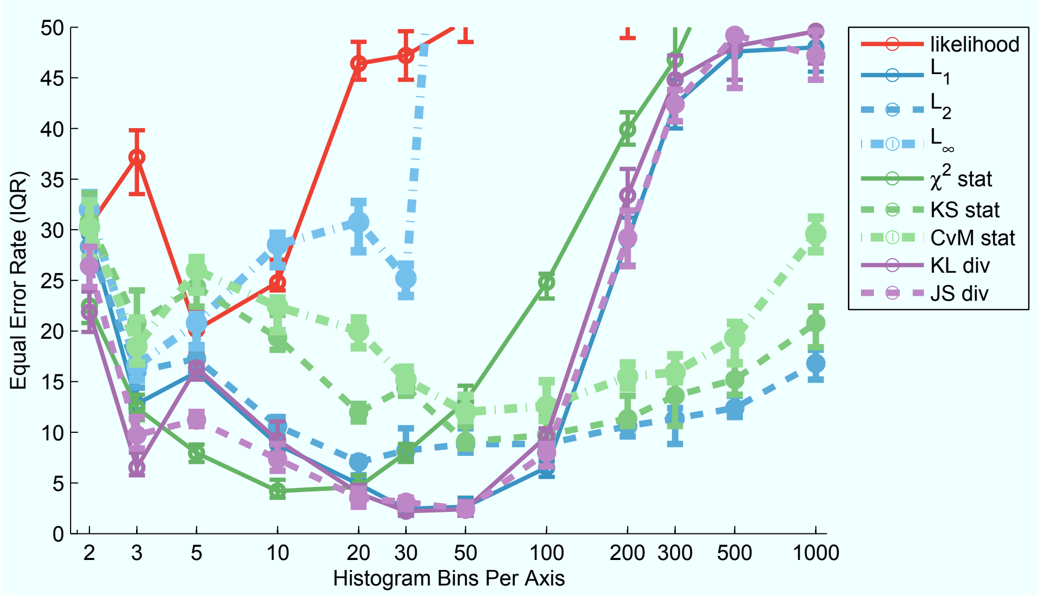

Figure 4.2: Anomaly detection equal error rate (showing interquartile range) as a function of model complexity with 33% anomalous contamination of the training data for 30 second test segments.

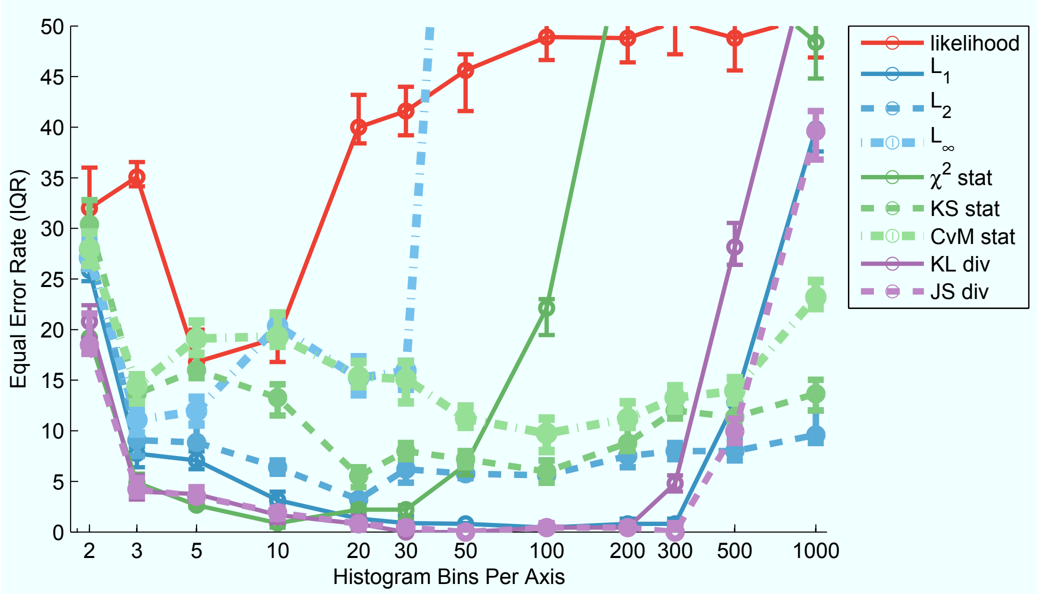

Figure 4.3: Anomaly detection equal error rate (showing interquartile range) as a function of model complexity with 33% anomalous contamination of the training data for the full 5 minute test segments.

KL divergence made *no errors* on the full 5 minute segments and the JS divergence, χ^2 statistic, and L_1 distance achieved EERs of 0.8%, 0.9%, and 1.3%, respectively. These four methods were also among the best performers on the shorter segments, but the differences were not always statistically significant.

The L_2 distance was the next best performer with an EER of 4.4% on the 5 minute segments. The KS test and L_∞ distance had comparable EERs of 7.2% and 8% while the CvM criterion and log likelihood baseline did not perform as well with EERs of 12.4% and 16.8%, respectively.

To explore the interplay between test segment length and model complexity, we show the EER for a range of both (Figure 4.4) when using the L_1 distance, χ^2 statistic, and KL divergence. The L_1 distance is especially attractive because its simplicity and numerical stability lead to a robustness unparalleled by the other two approaches. The χ^2 statistic performed well with 10 bins per axis while the KL divergence was harder to predict. We report the performance of all the approaches when testing on several segment lengths in Table 4.1.

4.3.3 Training Data Contamination

One aim of this chapter was to relax the requirement in Chapter 3 for large amounts of uncontaminated training data. Obtaining mostly normal, unlabeled data is often easier and far less expensive since labeled data requires human time and effort. Using unlabeled data also allows us to avoid the quagmire of defining what

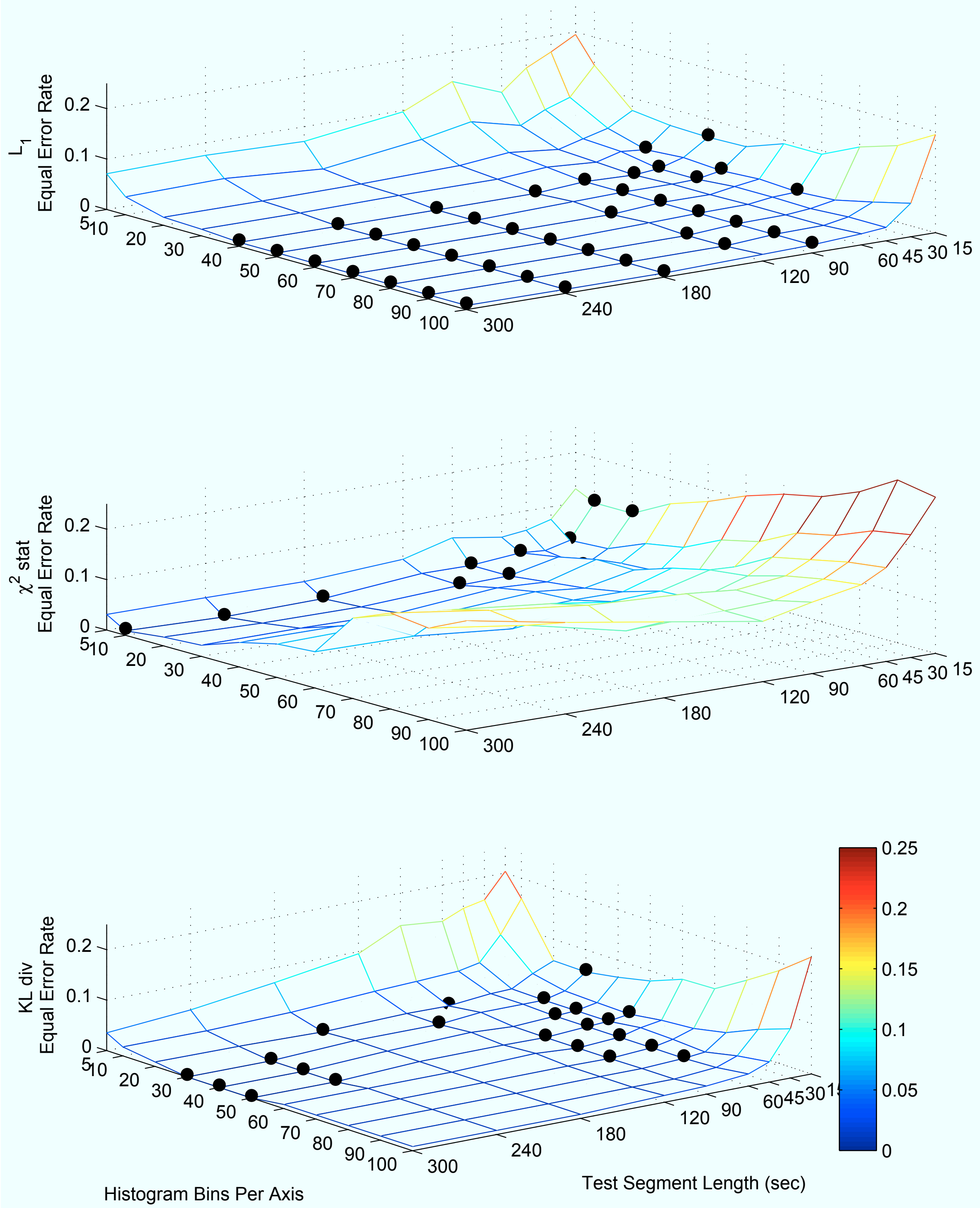

Figure 4.4: Anomaly detection equal error rate as the model complexity and test segment length are varied when training for 33% anomalous training contamination. A black dot is placed at the minimum(s) of each test segment length.

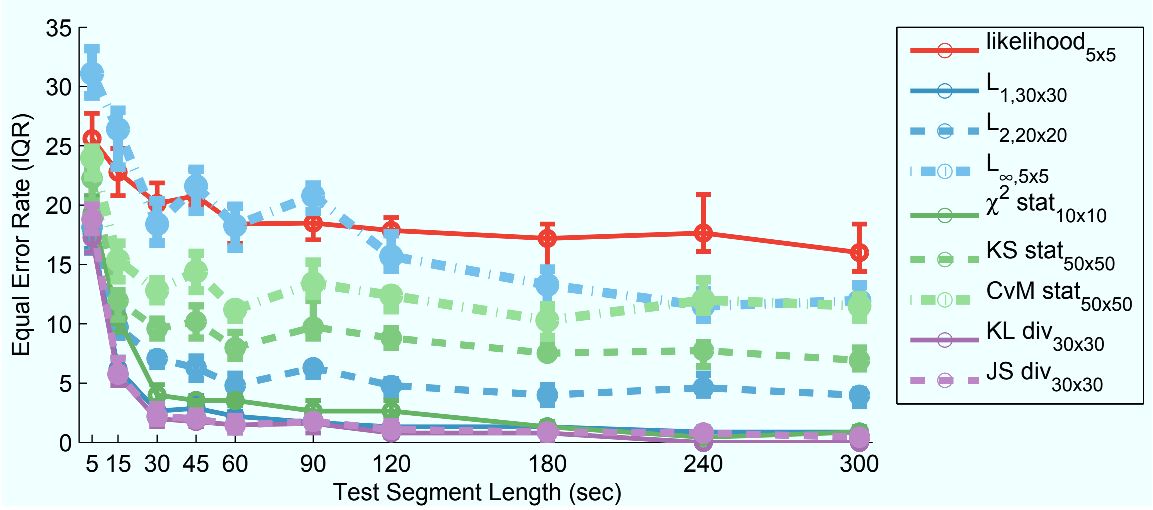

Figure 4.5: Interquartile range of anomaly detection equal error rate as a function of test segment length for 33% anomalous training contamination.

constitutes an anomalous sound. We are primarily interested in finding audio that is anomalous from the perspective of speech processing algorithms, not humans.

Given this desire, we investigated whether our methods could robustly model the normal data even if it was partially contaminated with anomalies. This was tested by incrementally adding anomalies into the data used for training the MNM and evaluating the anomaly detection performance. At each contamination level from 0% to 33% in approximate increments of $3\frac{1}{3}$%, we display the interquartile range of the EER via statistical bootstrapping (Figure 4.6).

We were encouraged by the robustness of all of the histogram-based approaches. The information theory divergences, χ^2 statistic, and L_1 and L_2 distances were minimally affected by the contamination. Our strategy of cell-ordering based on the MNM for the KS test and CvM criterion did not fare as well. With purely normal

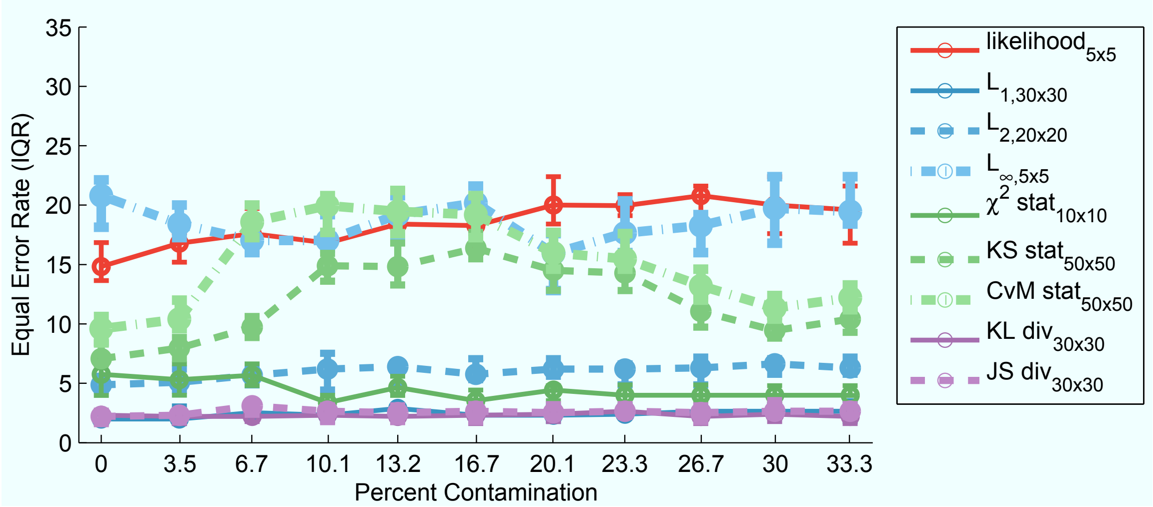

Figure 4.6: Interquartile range of anomaly detection equal error rate as a function of the amount of training data contamination for 30 second test segments.

data, the EER was cut in half as compared to an arbitrary ordering, but the effect quickly diminished as the contamination level reached 6.7%. For higher contamination levels performance started to improve. We attribute this unexpected behavior to our optimization of model complexity at 33% contamination. This also explains the performance of the L_∞ distance that was inversely related to the contamination level. The log likelihood baseline was the only method negatively affected by the amount of contamination throughout the entire range investigated.

We end our analysis of these methods by evaluating the detection error trade-off curve [216] when training at 33% contamination and testing on 30 second segments (Figure 4.7). The L_1 distance and information theory divergences were the best performers in this harsh test condition achieving EERs of 2.7% and 3.1%, respectively. The χ^2 and L_2 distance formed the next tier of performers with EERs of 4.9% and

6.4%, followed by KS, CvM, and L_∞ with EERs between 10.2% and 13.7%. The log likelihood baseline remained the worst performer with a 20.8% EER.

4.4 Conclusion

We were pleasantly surprised that the non-parametric histogram-based methods presented here were barely affected by up to 33% training contamination. The L_1 distance, χ^2 statistic, and information theory divergences were typically the best performers. While the KL divergence made no errors on the full 5 minute test segments regardless of the contamination level, we prefer the L_1 distance since its performance was comparable and its simplicity yielded the most robust performance to model complexity and test segment length. As we will see in the next chapter, coping with training data contamination can be a significant challenge for semi-parametric models and we will need to develop techniques to increase their robustness.

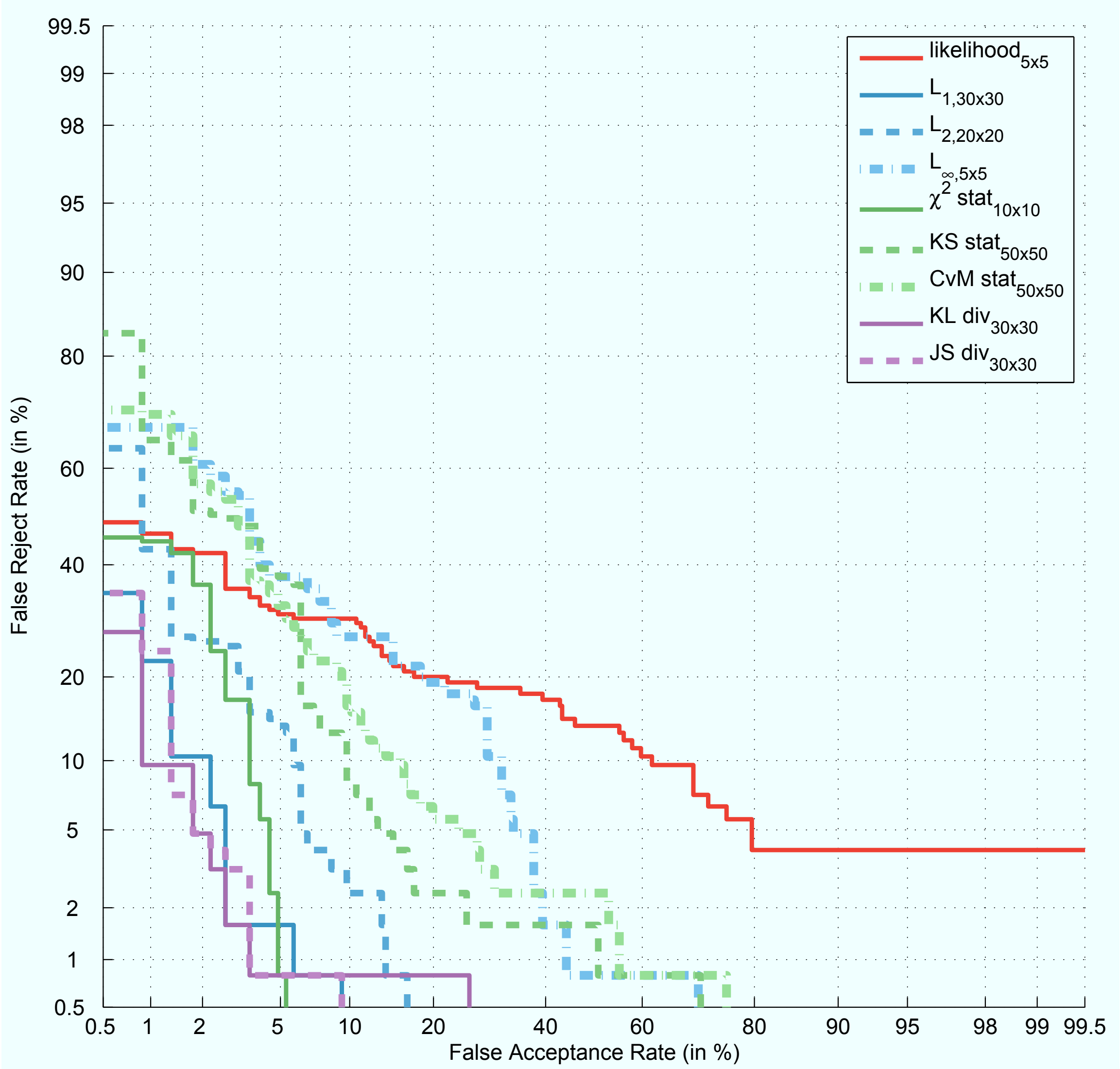

Figure 4.7: Anomaly detection error trade-off curve when training with 33% contaminated data and testing on 30 second segments.

Table 4.1: Percent equal error rate for histograms

Test Length (sec)		5	15	30	60	300
0% Contamination	likelihood$_{5x5}$	21.6	15.2	14.4	12.0	10.2
	L$_{1,30x30}$	13.3	4.9	2.4	1.3	0.4
	L$_{2,20x20}$	14.4	8.8	4.9	4.0	2.7
	L$_{\infty,5x5}$	36.0	30.1	19.5	16.8	12.0
	χ^2 stat$_{10x10}$	12.0	8.4	4.9	4.4	3.5
	KS stat$_{50x50}$	17.6	10.4	7.2	7.2	4.8
	CvM stat$_{50x50}$	18.1	12.8	9.6	8.0	5.8
	KL div$_{30x30}$	11.5	4.0	2.7	0.9	0.0
	JS div$_{30x30}$	12.4	4.9	2.7	1.3	0.4

Continued on next page

Table 4.1 – continued from previous page

Test Length (sec)		5	15	30	60	300
10% Contamination	likelihood$_{5x5}$	22.4	17.6	17.6	14.6	13.6
	L$_{1,30x30}$	15.2	4.9	2.7	1.6	0.4
	L$_{2,20x20}$	15.0	8.4	7.2	6.2	3.1
	L$_{\infty,5x5}$	30.1	29.6	18.6	16.4	13.6
	χ^2 stat$_{10x10}$	13.3	8.4	3.5	3.1	1.6
	KS stat$_{50x50}$	25.6	17.6	16.8	11.2	11.2
	CvM stat$_{50x50}$	28.0	21.2	20.8	17.3	14.2
	KL div$_{30x30}$	12.8	4.0	2.7	1.3	0.0
	JS div$_{30x30}$	13.3	4.9	2.7	1.3	0.4
33% Contamination	likelihood$_{5x5}$	27.2	23.5	22.1	20.8	16.8
	L$_{1,30x30}$	17.7	5.8	2.7	1.8	1.3
	L$_{2,20x20}$	18.1	10.6	6.4	5.8	4.4
	L$_{\infty,5x5}$	31.4	18.6	13.6	11.9	8.0
	χ^2 stat$_{10x10}$	20.8	12.0	4.9	4.0	0.9
	KS stat$_{50x50}$	21.7	13.3	10.2	8.8	7.2
	CvM stat$_{50x50}$	24.0	16.4	13.7	13.6	12.4
	KL div$_{30x30}$	18.1	5.3	3.1	1.6	0.0
	JS div$_{30x30}$	18.6	6.6	3.1	1.8	0.8

Chapter 5

Acoustic Anomaly Detection via Unsupervised Accommodation Learning from Contaminated Data using Gaussian Mixture Models

In Chapter 4, we began exploring unsupervised accommodation learning from contaminated data by computing the statistical divergence between histograms to detect acoustic anomalies. Here, we return to the methods from Chapter 3 employing Gaussian mixture models (GMMs) and introduce a promising approach to increase their robustness in the face of contamination. We first reported these results in [217] and we will use the same data as described in Section 4.1.

5.1 Gaussian Mixture Model Methods

Semi-parametric density estimation (Section 2.1.2) offers additional flexibility over parametric models, while maintaining a fixed number of parameters regardless of sample size. We began by training a GMM (Section 3.4) to characterize the data that we assumed was mostly normal. In this chapter, we experimented with GMMs using up to 16 components with full covariance matrices.

Unlabeled data often comes at minimal cost, so we were not concerned with using all of it. Whenever training the mostly normal model (MNM), we randomly assigned 66% of the data to an *initial* set, keeping the remainder in a *heldout* set. Training GMMs using the Expectation-Maximization (EM) algorithm can be a delicate process and we wanted to avoid local maxima and overfitting to the training data. To deal with the former we trained eight separate initial models. Each of these models was initialized using k-means clustering on 1000 samples randomly chosen without replacement from the *initial* set. After a maximum of 10 iterations of k-means clustering, we performed ML estimation using the EM algorithm until the parameters converged. The initial model with the highest log likelihood for the *heldout* set was then selected. Using this model and all of the data in the *initial* set, we continued to perform EM iterations while the log likelihood of the *heldout* set increased to ensure that the model would generalize. Our baseline anomaly detector is again the average log likelihood of an input sequence (Equation 3.2).

5.1.1 Trimming Gaussians for Robustness

The distribution of syllabic rate features is noticeably different between anomalies and normal audio (Figure 3.2). To obtain a model of a test sequence, we initialized the parameters to those of the MNM and performed ML estimation using the EM algorithm on the test data. While it is common to use MAP adaptation in such a scenario, we felt that deriving prior probabilities for the parameters using mostly normal data could not be justified when adapting to data that might be anomalous.

With sufficient contamination some Gaussians in the MNM would inevitably model anomalous regions of the feature space. When adapting to normal data, changes in these Gaussians could lead to false alarms. With labelled training data we could have estimated which Gaussians were modeling anomalous data and then discarded them before estimating the KL divergence. In our unsupervised setting we did not have such labels, so instead we exploit the mostly normal data by discarding a fraction of the most divergent Gaussians. Our approach began by treating the summands of Equation 3.7,

$$s_i = w_{p,i}\left(KL_{\mathrm{G}}\left(p_i \| q_{\pi(i)}\right) + \log\frac{w_{p,i}}{w_{q,\pi(i)}}\right) \tag{5.1}$$

as observations of a random variable whose location we want to estimate robustly. We do so by discarding α of the largest s_i's using the one-sided trimmed mean,

$$KL_{\alpha\text{-trimmed}}\left(p \| q\right) = m\left(\frac{1}{m-k}\sum_{j=1}^{m-k} s_{(j)}\right) \tag{5.2}$$

with $s_{(j)}$ denoting the order statistics and $k = \lfloor \alpha m \rfloor$. We also tried the more tra-

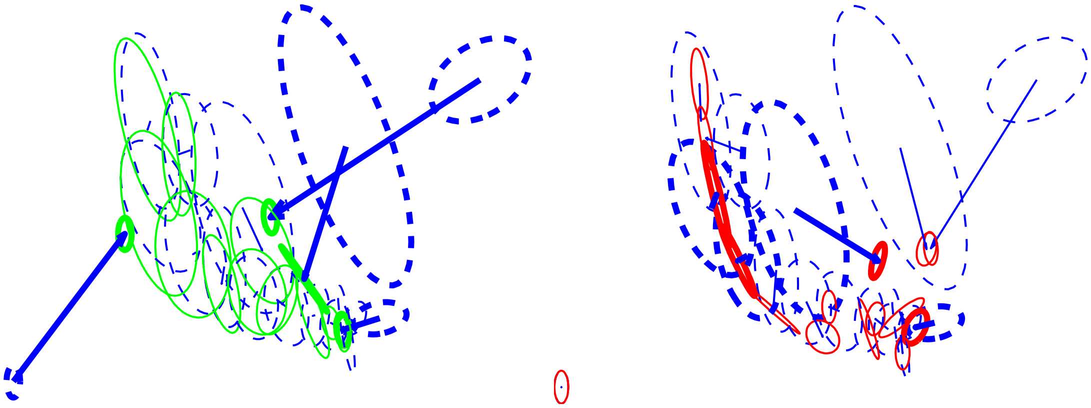

Figure 5.1: GMM ML adaptation from the MNM (dashed blue) with 33% training contamination to CallHome English cut 4829 side A (solid green on left) and a random sequence of DTMF tones keyed on and off every 700 ms (solid red on right). The four most divergent Gaussians that would be trimmed are shown with thicker lines.

ditional two-sided trimmed mean [218], but found it did not perform as well. We attributed this to the lack of negative outliers in the right-skewed distribution of the weighted Gaussian divergences.

An example of the adaptation from the MNM to a normal segment is shown on the left in Figure 5.1. A change in a few of the Gaussians lead to a divergence of 20.9 using Equation 3.7. Adaptation from the same MNM to an anomaly is shown on the right in Figure 5.1. The resulting divergence was only 2.9 despite more of the Gaussians being affected by the adaptation. After discarding the four most divergent Gaussians (shown with thicker lines) in an unsupervised manner, the normal and anomalous segments had trimmed KL divergences of 0.2 and 1.3, respectively.

5.2 Experimental Results

5.2.1 Training Data Contamination

Obtaining data that is mostly normal is relatively inexpensive since it does not require any annotation. We investigated if any methods could robustly model the normal data, even when it was partially contaminated with anomalies. Figure 5.2 shows the EER for each of the investigated methods using mixtures of 16 Gaussians as contamination levels were varied from 0% to 33% in approximate increments of $3\frac{1}{3}$%. To examine the relationship between performance and contamination, we first performed linear regression and then fit natural cubic splines to assess the linearity. Model selection for the splines was performed using the Bayes information criterion (BIC) [219].

Linear regression for the log likelihood method suggested that error rate increased with the amount of contamination (EER% = 22 + 0.87 per contamination percent, $r^2 = 0.87, P < 0.001$). However, spline fitting suggested that the relationship was slightly nonlinear ($df = 2, P < 0.001$). For purely normal training data, the KL divergence achieved the lowest EER (5.8%) with a median absolute deviation (MAD) of 1.3%. The difference between its performance and the $\frac{1}{4}$-trimmed KL divergence (6.3% EER, 0.9% MAD) was not significant ($P = 0.44$) using a Wilcoxon paired-sample signed rank test.

For contaminated training data, trimming one quarter of the Gaussians resulted

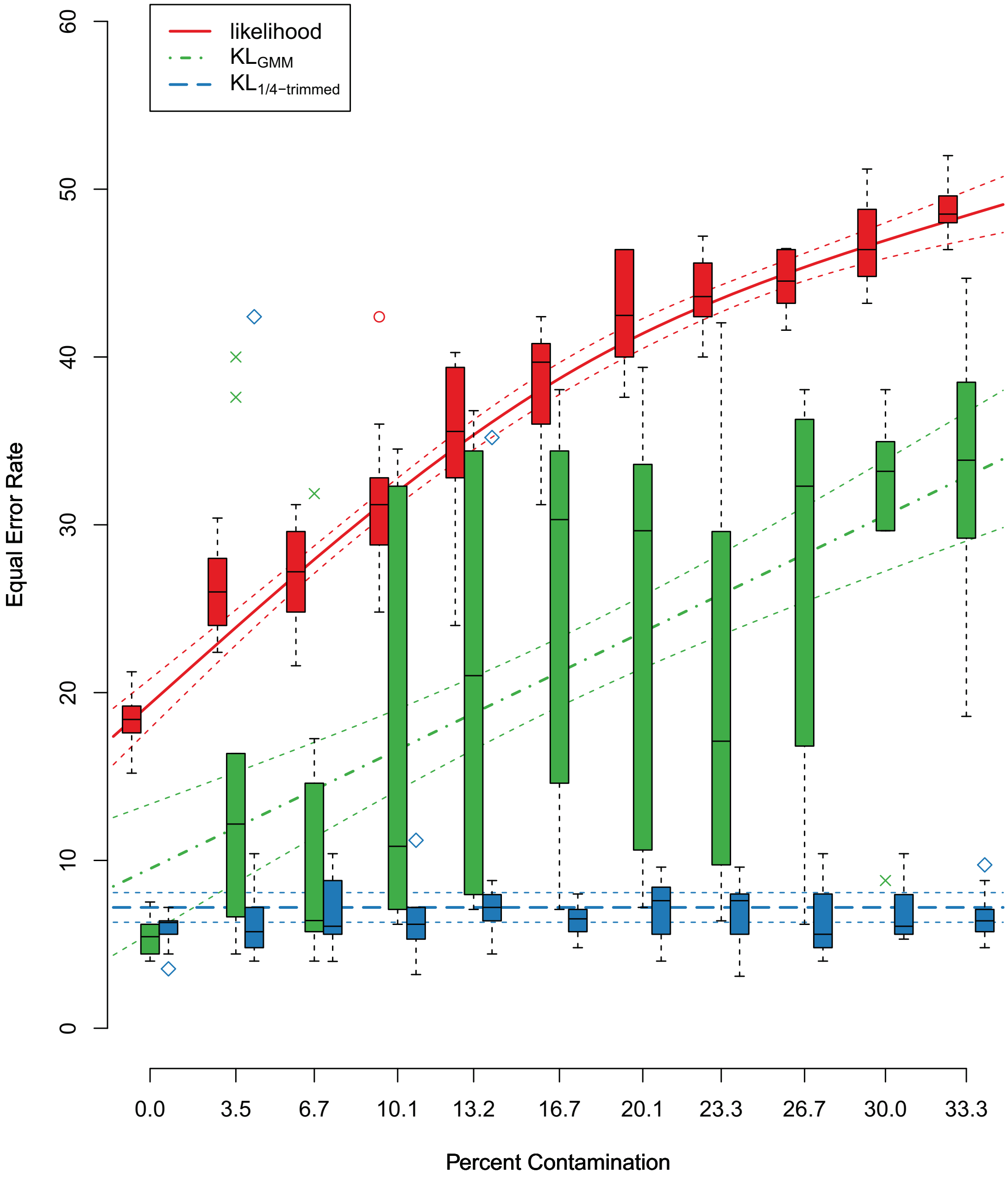

Figure 5.2: Box and spline plots of anomaly detection EER at various contamination levels using 16 Gaussians. The dotted lines around the splines indicate 95% confidence levels.

in significantly better performance, with one exception at 6.7% contamination ($P = 0.19$). The variability of performance for the untrimmed KL divergence was not well accounted for with a linear model (EER% = 9.5 + 0.70 per contamination percent, $r^2 = 0.32, P < 0.001$), but the BIC suggested that higher order splines offered no better fit. The performance of the $\frac{1}{4}$-trimmed KL divergence did not show a significant dependence on the amount of contamination ($P = 0.53$) and a constant model resulted in a better fit (EER% = 7.6, $P < 0.001$). The median EER for all methods including $\frac{1}{2}$-trimming KL divergence are shown in Table 5.1 for select contamination levels.

5.2.2 Model Selection

We also evaluated the performance of each method as we varied the number of Gaussians from 1 to 16 when training on 33% contaminated data (Figure 5.3). Spline fitting suggested that all relationships were nonlinear, with three degrees of freedom for both the log likelihood method and KL divergence and four degrees of freedom for the $\frac{1}{4}$-trimmed KL divergence. The log likelihood method achieved its lowest EER of 28.8% (1.2% MAD) using a single Gaussian, although this was not consistent for other contamination levels. The KL divergence achieved its lowest EER of 7.7% (1.5% MAD) using 6 Gaussians. Both methods showed a dependence on model complexity that would require careful optimization for any new data set. In contrast, trimming one quarter of the Gaussians resulted in robust performance over a wide range of model complexities (6.4% EER, 1.0% MAD for 16 Gaussians). This performance was significantly better than the untrimmed KL divergence with a comparable model complexity ($P < 0.001$), but not when compared to the untrimmed divergence using 6 Gaussians ($P = 0.053$).

5.3 Conclusion

Using only unlabeled data, our goal was to develop a robust acoustic anomaly detector using two syllable rate features from a speech activity detector. When trained on purely normal data, we found that the KL divergence achieved the lowest EER of the three methods employing GMMs. When subjected to training contamination, the performance of the KL divergence suffered dramatically and optimization of its model complexity became extremely important. Seeing the merit in this approach, we wanted to improve its robustness to contamination.

By trimming one quarter of the most divergent Gaussians we were able to statistically remove the effect of contamination up to 33%. We experimented with other trimming ratios, but one quarter had the most consistent performance regardless of contamination level and model complexity. Such a detector could work in tandem with other speech processors, enabling the overall system to have a means of detecting anomalous audio at little additional cost.

Table 5.1: Percent equal error rate (median with n=10) for GMMs

Contamination		0%	3.5%	10%	20%	33%
4 Gauss.	likelihood	19.2	22.4	22.8	24.4	44.8
	KL_{GMM}	4.8	8.8	10.1	12.0	22.3
	$KL_{1/4\text{-trim}}$	4.6	7.2	8.4	15.8	16.4
	$KL_{1/2\text{-trim}}$	5.9	11.8	15.1	27.1	24.8
8 Gauss.	likelihood	20.0	24.0	29.6	38.8	45.6
	KL_{GMM}	4.4	5.2	5.6	6.9	7.2
	$KL_{1/4\text{-trim}}$	4.4	4.2	5.9	6.8	8.6
	$KL_{1/2\text{-trim}}$	6.8	8.4	9.2	10.8	13.0
16 Gauss.	likelihood	18.4	26.0	31.2	42.5	48.5
	KL_{GMM}	5.5	12.2	10.8	29.6	33.8
	$KL_{1/4\text{-trim}}$	6.3	5.8	6.2	7.6	6.4
	$KL_{1/2\text{-trim}}$	11.6	12.8	12.2	12.4	11.2

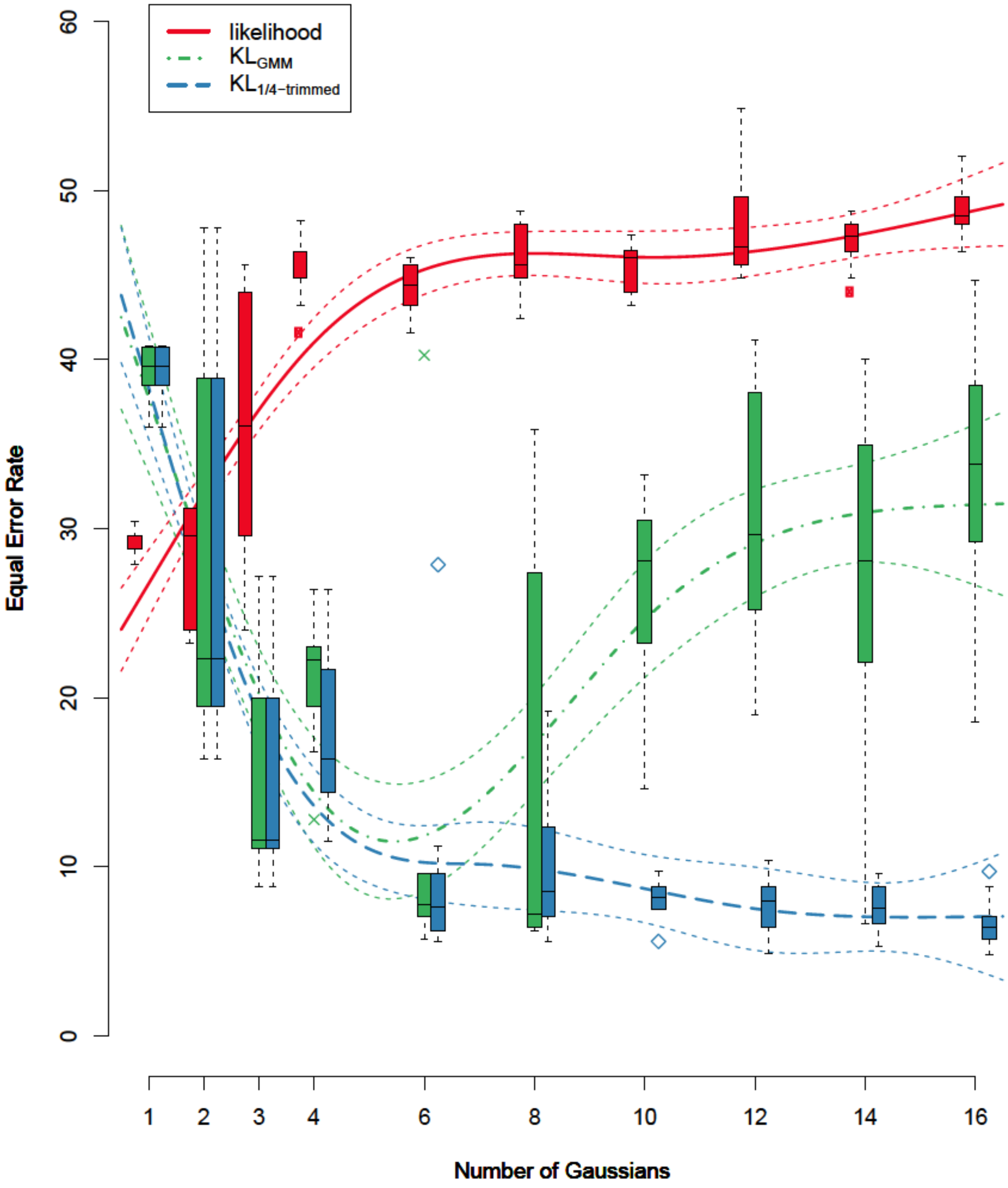

Figure 5.3: Box and spline plots of anomaly detection EER for various model complexities when training with 33% anomalous contamination. The dotted lines around the splines indicate 95% confidence intervals.

Chapter 6

Anomaly Detection for Graphs via Partially Supervised Learning from Normal Data

In Chapters 3-5, we developed various techniques for distributional anomaly detection and evaluated their performance on acoustics. In order to demonstrate the generality of our approach, we will now investigate the same techniques in a completely different application area: graph matching. The only difference in our methodology will be the feature extraction that must be tailored to the new domain.

A graph $G = (V, E)$ is comprised of a set V of objects called *vertices* and a set E of connections between them called *edges*. The theory of graphs can trace its origin back to 1735 when Euler used it to solve a vexing problem involving the bridges of

Kőnigsberg [220]. The question was if the seven bridges connecting four landmasses could each be traversed without crossing any bridge twice. By simplifying the problem and representing the landmasses as vertices and the bridges as edges in a graph, he offered an elegant proof that a non-backtracking traversal is an impossible task since all four vertices have an odd number of incident edges.

Since then, graphs have been used to characterize a multitude of real world social, informational, technological, and biological networks [221]. They have been used to study everything from the synaptic connections between neurons of a nemotode [222] to the macro-economic and political relationships between nation states [223]. Graphs have enabled a better understanding of the decentralized insurgent networks in Afghanistan and Iraq [224] and the preferential attachment between scientific collaborators [225]. They have offered insights into the structural richness and omnivory of ecological food-webs [226] and aided in the design and analysis of protocols given the Internet's topological properties [227]. Graphs can represent the co-starring roles of Hollywood actors [228] and the hundreds of trillions of connections in the human brain [229]. They can be used to infer gender, age, and education levels from human interactions and turn-taking behavior [230] and model everything from protein folding [231] and metabolic networks [232] to the world wide web [233] and telephone calls [234]. We can use them to study how a new product penetrates the consumer market [235], physicians adopt a new drug [236], and epidemics spread through computer networks [237] and society [238]. Graph theory is a burgeoning field of study with numerous survey papers [221, 239–241] and several popular books [242, 243] describing the significant achievements that have been made in the field over the last few decades. In the 21st century we live in a data-rich, connected world [244] and graphs provide a concise mathematical representation in which to study it.

The goal of this chapter is to detect anomalies in such graphs, which could serve as a useful tool for determining when changes have occurred in these wide ranging phenomena. Anomaly detection is a longstanding problem with many applications in statistics and signal processing [117]. Here, we consider anomaly detection on graphs, a subject which has not previously had treatment in such depth.

6.1 Preliminaries

Consider graph $G = (V, E)$ from the space of simple graphs $\mathcal{G}$. The order of G is the number $n = |V|$ of vertices and the size of the graph $s = |E|$ is the number of edges. We denote an edge between u and v in V as (u, v) in E. Such vertices u and v are said to be adjacent, and each are incident to the edge (u, v). We only consider simple graphs with undirected edges, so $(u, v) = (v, u)$.

The adjacency matrix A of a simple graph G is the symmetric binary matrix in which entry $a_{ij} = 1$ if $(v_i, v_j) \in E$, otherwise $a_{ij} = 0$. If there exists a nonzero vector $\boldsymbol{x} \in \mathbb{R}^D$ such that $A\boldsymbol{x} = \lambda\boldsymbol{x}$, then $\boldsymbol{x}$ is referred to as an eigenvector of A associated with eigenvalue λ. Since A is symmetric for undirected graphs, we know that all of its eigenvalues are real and A is orthogonally diagonalizable [245]. This implies that there exists an orthogonal matrix Q such that $\Lambda = Q^{-1}AQ$ is diagonal. Since $Q^{-1} = Q^T$, we also know that $Q\Lambda Q^T = A$. If we choose $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$, then Q can be constructed with the corresponding orthonormal eigenvectors using Gram-Schmidt orthonormalization, such that $Q = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$, where $A\boldsymbol{x}_i = \lambda_i \boldsymbol{x}_i$ and $||\boldsymbol{x}_i|| = 1$ for $i = 1, \ldots, n$.

6.2 Previous Work

Graph matching is the process of determining whether or not two graphs have the same structure. It has been the subject of research for over four decades [246, 247] and is related to anomaly detection with the antithetical goal of determining when two graphs are the same. Two graphs, $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, are said to be isomorphic if there exists a one-to-one mapping, f, between the vertex sets, V_1 and V_2, that preserves the edge structure. That is, if there exists an edge $(u, v) \in E_1$ then there also exists an edge $(f(u), f(v)) \in E_2$. In terms of computational complexity, the graph isomorphism problem is the subject of a great deal of research since it is probably neither in P nor NP-complete [248].

Graph matching has many applications dating all the way back to 1965 where it was used for chemical information retrieval [249] and it continues to be used in a variety of bioinformatic [250] and chemoinformatics tasks today [251]. It is a powerful tool often used by the computer vision community for tasks such as 3D object recognition [252], character recognition [253, 254], and shape analysis [255–257]. In addition to many other applications, it has also been used to monitor computer networks [258], mine software design patterns [259], and cluster concepts [260].

Techniques for graph matching can be broadly divided into two categories depending on whether they are testing for an *exact* isomorphic match as described above or if some structural differences are allowed in the search for an approximate or *inexact* match [247].

6.2.1 Exact Graph Matching

The most widely known algorithm for exact graph matching is Ullman's [261] enumeration algorithm and tree search with a look-ahead function to reduce the search space. Modifications to this approach have included branch-and-bound techniques [262] along with more recent approaches to reformulate it as an optimization problem [263] that can be addressed using techniques such as constraint propagation [264]. The complexity of these tree-search techniques is exponential in the worst case [265] although heuristics can help achieve polynomial complexity in certain situations, such as using distance metrics between vertices and edges when working with attributed relational graphs [266].

One of the fastest methods for exact graph matching is McKay's group-theoretic canonical labeling algorithm [267], available in the highly regarded *nauty* (no automorphisms, yes?) software package[1]. After determining which automorphism group each vertex-colored graph belongs to, McKay provides information about the group in the form of a set of generators as well as its size and orbits. This is then used to produce a canonical labeling for each graph to assist in isomorphism testing [268]. It enables equality verification in $O(n^2)$, making it suitable for fast database lookup after deriving the canonical labeling, which can require exponential time in the worst case [269]. However, on average it performs quite well, although it cannot easily exploit vertex and edge attributes [247].

Others [270, 271] have used decision trees [272] to similarly enable fast retrieval from a graph library. However, the pre-processing time and storage space is often exponential in the size of the graphs in the library [247].

When an exact graph match cannot be found between G_1 and G_2, we can still search for a subgraph isomorphism [261] where G_2 contains a subgraph G_2' that is isomorphic to G_1 with vertices $V_2' \subseteq V_2$ and edges $E_2' \subseteq E_2$. From the set of all possible subgraphs, we are often interested in finding the maximum common subgraph (MCS) [273]. However, subgraph matching is known to be NP-complete [274] and the added computational complexity of exact MCS methods restricts their use to relatively small graphs.

[1] http://cs.anu.edu.au/~bdm/nauty

6.2.2 Inexact Graph Matching

Many methods of inexact graph matching try to overcome these computational drawbacks by producing an approximate answer in a reasonable amount of time. Their error tolerance can also provide more useful results when dealing with graphs representing real word data, which often suffer from noise or other distortions. For example, when dealing with incomplete sampling, an observed graph G' might have some differences in its vertex or edge sets compared to the true graph G. This has been formalized in the notion of a graph edit distance, which is the minimal number of vertex and edge insertions, deletions, or substitutions necessary to transform G' into G. Bunke showed that the graph edit distance

$$d(G, G') = |V| + |V'| - 2|V''| \tag{6.1}$$

is equivalent to the MCS problem for a particular class of cost functions [275] where $G'' = (V'', E'')$ is the MCS of G and G'. This naturally leads to a graph similarity measure,

$$\delta(G, G') = 1 - \frac{|V''|}{\max(|V|, |V'|)} \tag{6.2}$$

which is a metric that can be used where reflexivity or the triangle inequality are desired [276].

Inexact graph matching techniques can generally be divided up into methods using tree search, continuous optimization, and spectral graph theory [247]. Optimal methods of error-tolerant graph matching [266, 277–279] typically use a tree search guided by a cost function of a partial match to find the optimal subgraph isomorphism, but this comes at the cost of exponential time since that problem is NP-complete. A heuristic estimate of the future matching cost can speed up the search if the estimate is accurate, but these are often application dependent.

Although suboptimal methods are not guaranteed to find the best solution, the approximate answer they can provide quickly is useful for a wide variety of applications, especially when working with large graphs. One such method casts the discrete matching problem into one of continuous optimization, where many existing algorithms can be brought to bear on the problem. A prime example of this is relaxation labeling [280–284] where the probability of each vertex in one graph mapping to each vertex in the other is initialized based on their attributes or connectivity. These are then modified in successive iterations until the process converges to a fixed point where the best solution is represented by the mapping with the highest probability [247].

The only problem with relaxation labeling is that there is no guarantee that the mapping is one-to-one. Weighted graph matching allows for two way constraints on the correspondence by means of a matching matrix between vertex sets whose elements are constrained to be either 0 or 1 with each row and column summing to unity [247]. This is often transformed into a continuous optimization problem by allowing the elements to have continuous values between 0 and 1, which can be solved in polynomial time using linear programming and then converted back into discrete form using the Hungarian method [285].

The expectation-maximization (EM) algorithm has also been used for optimization rather than search. Luo and Hancock [286] develop a likelihood function for the graph matching problem and use EM and singular value decomposition to iteratively estimate a set of assignment variables. This approach is only guaranteed to find local optima and is highly dependent on the initial conditions. Other recent methods of continuous optimization techniques for inexact graph matching include fuzzy graph matching [287] and reproducing kernel Hilbert spaces [288].

Spectral methods for inexact graph matching seek to represent and distinguish structural properties of graphs using eigendecompositions of graph adjacency matrices which are invariant to vertex/edge reorderings [289]. Rather than use optimization techniques that are only guaranteed to find a local optimum, Umeyama [290] pioneered an analytic approach based on the observation that isomorphic graphs will have the same spectra regardless of the vertex labelings. Umeyama's approach is guaranteed to find the optimal match if two graphs are isomorphic and a suboptimal solution if the graphs are nearly isomorphic. Umeyama assumes that each graph has a set of distinct eigenvalues, but he states that they can be perturbed if multiple roots exist without significantly affecting the result. Recently, Xu and King [291] offered speed-ups to Umeyama's approach by approximating the permutation matrix with a generic orthogonal matrix and using principle components analysis to derive an objective function that they claim is faster and more accurate.

In order to use spectral approaches to match graphs of different sizes, Carcassoni and Hancock [292] adopted a hierarchical approach using the modal structure of the graph adjacency matrices. After assigning vertices to clusters using an eigendecomposition, they use the within-cluster and between-cluster adjacency matrices to compute cluster-conditional correspondence probabilities. Kosinov and Caelli take a different approach and use the eigenvalues to project the vertices and their structure onto the graph's most important eigenvectors using principle components analysis to reduce dimensionality [289]. All these methods of eigendecompositions work on any directed or undirected graphs with non-negative weights on the edges, but they cannot make use of other vertex attributes.

Allowing for error-tolerant isomorphisms becomes even more important when dealing with an attributed graph $G = (V, E, \mu^V, \mu^E)$ where $\mu^V : \mathcal{V} \rightarrow \mathcal{A}^{\mathcal{V}}$ and $\mu^E : \mathcal{E} \rightarrow \mathcal{A}^{\mathcal{E}}$ represent attributes on the vertices and edges, respectively. These attributes often come from measurements or other inference algorithms relevant to a particular application, and they can provide useful information when assessing graph similarity.

6.2.3 Anomaly Detection using Graph Invariants

Graph invariants are properties or parameters of a graph that must be preserved under isomorphisms [293]. Pao, et al. [294] analyzed the inferential capability of scalar graph invariants as test statistics for differentiating homogeneous graphs from heterogeneous "chatter" alternatives where a subset of the vertices are overly-connected. Their results indicate that there is no uniformly most powerful summary statistic across the space of "chatter" alternatives, although the maximum locality statistic has significantly more power than the others over large regions of the alternative parameter space. The methodology of using graph invariants as test statistics to differentiate homogeneous graphs from "chatter" alternatives is important when deciding how to approach anomaly detection in this setting, but it does not speak to the robustness of using a scalar graph invariant for the detection of a wide variety of anomalous graphs.

6.3 Vertex Invariants

Vertex invariants are functions $i : V_G \to \alpha$ from the vertices V in a graph G to a value in α, such that if an isomorphism maps v onto v' then $i(v) = i(v')$ [295]. Such vertex properties have been used in heuristic techniques to speed up the testing for graph isomorphisms [267]. Many vertex invariants are also referred to as centrality measures [296], since they can be used to estimate "power and influence" in social networks [297]. One famous example involves mapping the covert network of terrorist cells around the 9/11 hijackers [298]. Krebs states that "after one month of investigation it was 'common knowledge' that Mohamed Atta was the ring leader" and his analysis showed that he had the highest degree and closeness centrality measures.

- **Degree**: The simplest local vertex invariant is its degree or number of incident edges. When used as a measure of connectivity independent of graph order n, it is often normalized [296],

$$D(v) = \frac{\deg(v)}{n-1}. \tag{6.3}$$

The **max degree** $\max_{v \in V} D(v)$ is also investigated here as a graph invariant.

- **Betweenness**: The betweenness of a vertex v,

$$B(v) = \frac{\displaystyle\sum_{v \neq t \neq u} \frac{\sigma_{tu}(v)}{\sigma_{tu}}}{(n-1)(n-2)}, \tag{6.4}$$

  is a measure of global centrality where σ_{tu} is the number of shortest paths between vertices t and u and $\sigma_{tu}(v)$ is the number of shortest paths between vertices t and u that pass through v.

- **Closeness**: The closeness of vertex v is the reciprocal of the average distance to other reachable vertices,

$$C(v) = \frac{n-1}{\displaystyle\sum_{u \in V \backslash v} d(u,v)}, \tag{6.5}$$

  where $d(u,v)$ is the shortest path distance between vertices u and v.

- **Eigenvector Centrality**: The eigenvector centrality [299] of vertex v is proportional to the sum of scores of all adjacent vertices,

$$EV(v) = \frac{1}{\lambda} \sum_{u \in V, (u,v) \in E} EV(u). \tag{6.6}$$

  It is so named because it is the solution to the eigenvector equation of the adjacency matrix A. In general, many solutions exist, but the additional constraint that $EV(v) > 0$ for all $v \in V$ restricts us to the eigenvector corresponding to the largest eigenvalue λ.

- **PageRank**: Originally defined on directed graphs representing the World Wide Web [44], PageRank is a modified version of eigenvector centrality which can be reformulated for undirected graphs as the solution to the recursive equation,

$$PR(v) = (1 - df) + df \sum_{u \in V, (u,v) \in E} \frac{PR(u)}{\deg(u)}, \tag{6.7}$$

  where $df = 0.85$ is a commonly used damping factor.

- **Triangles**: We denote the number of triangles (cycles of length 3) involving vertex v as $\tau(v)$.

- **Locality Statistic**: The first order locality statistic [300] of vertex v is

$$L(v) = \text{size}(\Omega(N[v])), \tag{6.8}$$

  where N is the first order neighborhood of v and Ω is the induced subgraph. The first order **scan statistic** of graph G is $S(G) = \max_{v \in V} L(v)$, which is also investigated here as a graph invariant.

- **Clustering Coefficient**: The local clustering coefficient of vertex v is defined as

$$CC(v) = \frac{2\tau(v)}{\deg(v)\,(\deg(v) - 1)}, \tag{6.9}$$

  which measures how close its neighbors are to being fully-connected [228].

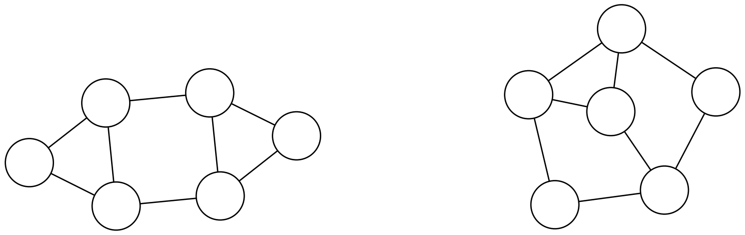

Figure 6.1: Two non-isomorphic graphs with the same degree sequence {3, 3, 3, 3, 2, 2}, but different spectra.

6.3.1 Anomaly Detection using Vertex Invariants

In this Chapter, we propose a methodology of detecting anomalous graphs that follows naturally from Chapters 3-5, where we performed acoustic anomaly detection by estimating the divergence between statistical populations of syllable rate estimates. When comparing graphs, we will substitute local measurements in time with local measurements about each vertex (i.e. vertex invariants) and then estimate the divergences between their statistical populations. In some ways this is a generalization to Kosinov and Caelli's projection of vertices into the eigenspaces of graphs [289]. However, our approach can make use of any vertex invariant or attribute, which may be readily available in many real-world graph databases. We first reported these results in [301].

The simplest corollary to our approach that is often used as a fast heuristic for inexact graph matching is comparing degree sequences. The problem is that there are many non-isomorphic graphs with the same degree sequences, such as those in Figure 6.1. A similar approach, which compares the spectral decompositions of each adjacency matrix, suffers from the same problem. While all isomorphic graphs are

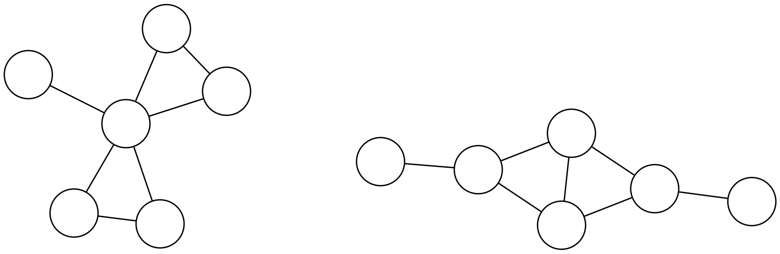

Figure 6.2: Two non-isomorphic graphs with the same spectra of approximately {2.7, 1.0, 0.19, -1.0, -1.0, -1.9}, but different degree sequences.

isospectral, not all isospectral graphs are isomorphic. Several examples of the latter were first discovered in 1957 by Collatz and Sinogowitz [302]. Since then, many other isospectral graphs have been discovered with the smallest connected examples having 6 vertices (Figure 6.2). Mowshowitz [303] later demonstrated a way of constructing connected, regular, isospectral, non-isomorphic digraphs and Schwenk showed that as the number of vertices grow, the probability of occurrence of two isospectral, non-isomorphic trees approaches unity [304].

Any such one-dimensional approach is bound to be problematic. If we compared both degree sequences and spectra, we could have determined that the graphs in Figures 6.1 and 6.2 were not isomorphic. However, comparing two attributes is not necessarily enough. For example, Figure 6.3 shows two graphs with the same degree sequences and spectra, but different clustering coefficients. Even though these graphs are not isomorphic, similar substructures can be seen in both.

If we ignore all numerical and computational considerations and use all n eigenvectors and eigenvalues, then the isomorphism problem can be "solved" for most practical purposes [305]. The caveat is that it is difficult to estimate the complex-

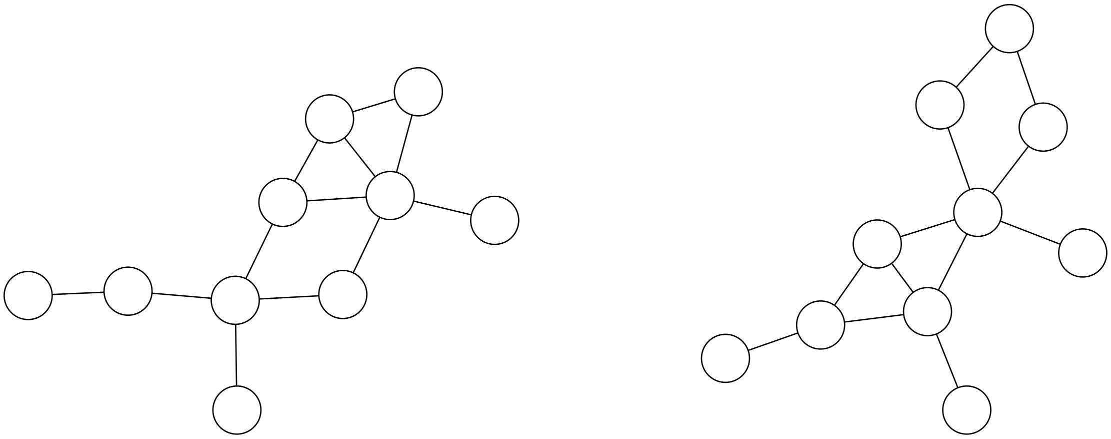

Figure 6.3: Two non-isomorphic graphs with the same degree sequence of {5, 4, 3, 3, 2, 2, 2, 1, 1, 1} and spectra of approximately {3.01, 1.79, 0.79, 0.50, 0.00, 0.00, -0.74, -1.31, -1.56, -2.48}, but different clustering coefficients.

ity of the backtracking algorithm required when coping with eigenspaces of multiple dimensions.

The unanswered question is if there is a “decent” *complete set of invariants* that determines a graph up to isomorphism [305]. We have chosen to investigate eight invariants which are useful for characterizing and visualizing large scale graphs (Figure 6.4). Another question, which is application dependent, is how important is distinguishing non-isomorphic graphs that are so similar that it is hard to find any differences among sets of multiple vertex invariants. As that level of importance rises, more vertex invariants can be used in the comparison. It is just a matter of engineering the proper trade-off (Section 1.2.4) between accuracy and other considerations such as speed and cost.

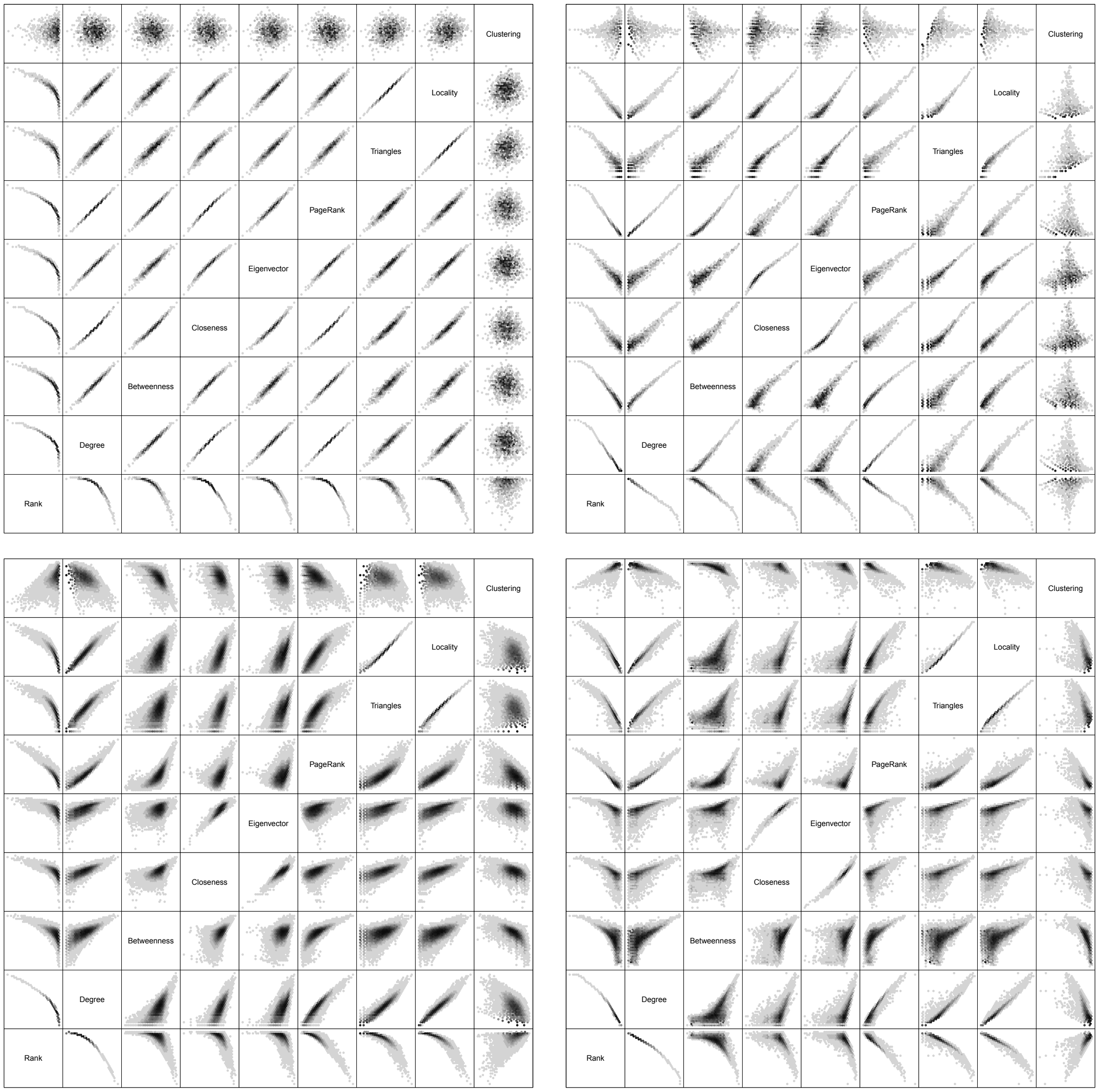

Figure 6.4: Histograms of vertex invariants for an Erdös-Rényi random graph (n=1000, p=0.1; top left), Barabási-Albert preferential attachment graph (n=1000, m=10; top right) [3], a citation network of 34,546 arXiv articles on High Energy Physics Phenomenology (bottom left) [4], and an Enron-based communication network with 36,692 email addresses (bottom right) [5]. The invariants for the largest connected component are plotted on logarithmic axes after removing the zeros, along with the rank of each vertex according to decreasing degree.

6.4 Experimental Design

We are interested in demonstrating the generality of our approach to anomaly detection by showing that our techniques can discover anomalies present in Erdős-Rényi random graphs. Erdős and Rényi [306] were the first to analyze large graphs with equiprobable edges [307]. Before their work, graph theory research was largely focused on proofs associated with small and regular graphs that could be analyzed manually. Over the years, increasing computational power has lead to models that have become progressively more realistic. Random graphs can now be created which are *highly-clustered*, *small-world* networks [228] with *scale-free* degree distributions [3]. The properties of such complex networks have been rigorously studied by van der Hofstad [308], but our goal is more modest. We want to compare our approach of distributional anomaly detection to existing baselines on Erdős-Rényi graphs [294].

6.4.1 Null Hypothesis

Our null hypothesis (H_0) is that the observed graph is drawn from the class of Erdös-Rényi random graphs, $ER(n, p)$, with n vertices where each of the $\binom{n}{2}$ possible edges exist independently with probability p.

6.4.2 Alternative Hypotheses

The alternative hypothesis (H_A) is that the observed graph is not drawn from the class of Erdös-Rényi random graphs. We test this by generating graphs where a subset of vertices are connected according to a different process. Let the set of anomalous vertices be Γ (of order γ) and the set of non-anomalous vertices be $V \setminus \Gamma$ (of order $n-\gamma$). In all cases, the $\binom{n-\gamma}{2}$ possible edges connecting vertices in $V \setminus \Gamma$, exist independently with probability p, as in H_0. Moreover, the $\gamma(n-\gamma)$ possible edges between vertices in Γ and $V \setminus \Gamma$ also exist independently with probability p, as in H_0. The four types of anomalous graphs treat the $\binom{\gamma}{2}$ possible edges between vertices in Γ differently (shown in Figure 6.5).

- **Increased Connectivity** H_1: A *kidney-egg* graph, $\kappa(n, p, \gamma, q)$, where connections between the γ anomalous vertices exist independently with probability q where $q > p$. This condition is directly comparable to [294].

- **Decreased Connectivity** $H_{\{2,3,4\}}$: A *kidney-void* graph, $\kappa_c(n, p, \gamma, \rho)$, where the γ vertices are either not connected to each other or form a tree with branching factor ρ (each vertex connected to at most ρ children in Γ and one parent[2]). We investigated H_2: $\rho = 0$, where there are no connections between the vertices in Γ ; H_3: $\rho = 1$, where Γ form a path; and H_4: $\rho = 2$, where Γ form a binary tree.

[2] For $\rho > 0$, this means there are exactly $\gamma - 1$ edges present between vertices in Γ ; equivalent values of q are extremely small.

We also evaluate detection performance when distinguishing Erdös-Rényi graphs from a pooled test set of these anomalous graphs, where half have a local region of increased connectivity (H_1) and half have a local region of decreased connectivity ($H_{\{2,3,4\}}$).

6.4.3 Divergence of Vertex Invariant Distributions

For each vertex v in a graph G, we can fuse the information from D vertex invariants into $\boldsymbol{\psi}(v, G) \in \mathbb{R}^D$. For notational convenience, we will occasionally drop the operands and refer only to $\boldsymbol{\psi}$. Given a set of graphs $\mathbb{G} = \{G_1, \ldots, G_s\}$, each of order n, we estimate the joint probability density function $p_{\mathbb{G}}(\boldsymbol{\psi})$ of vertex invariants using $\{\boldsymbol{\psi}(v_1, G_1), \ldots, \boldsymbol{\psi}(v_n, G_s)\}$. Our general approach is to measure the divergence,

$$d(p_{\{G\}}(\boldsymbol{\psi}) || p_{\mathcal{N}}(\boldsymbol{\psi})), \tag{6.10}$$

between density estimates of vertex invariants for graph G and the training set $\mathcal{N}$ of normal graphs.

6.4.3.1 Histograms

The oldest, simplest, and most popular form of nonparametric density estimation is the histogram (Section 2.1.3), which dates back as far as 1662 to mortality tables in the age of the plague [309]. Histograms provide a consistent estimate of the true underlying probability density function [310] while not making parametric assumptions

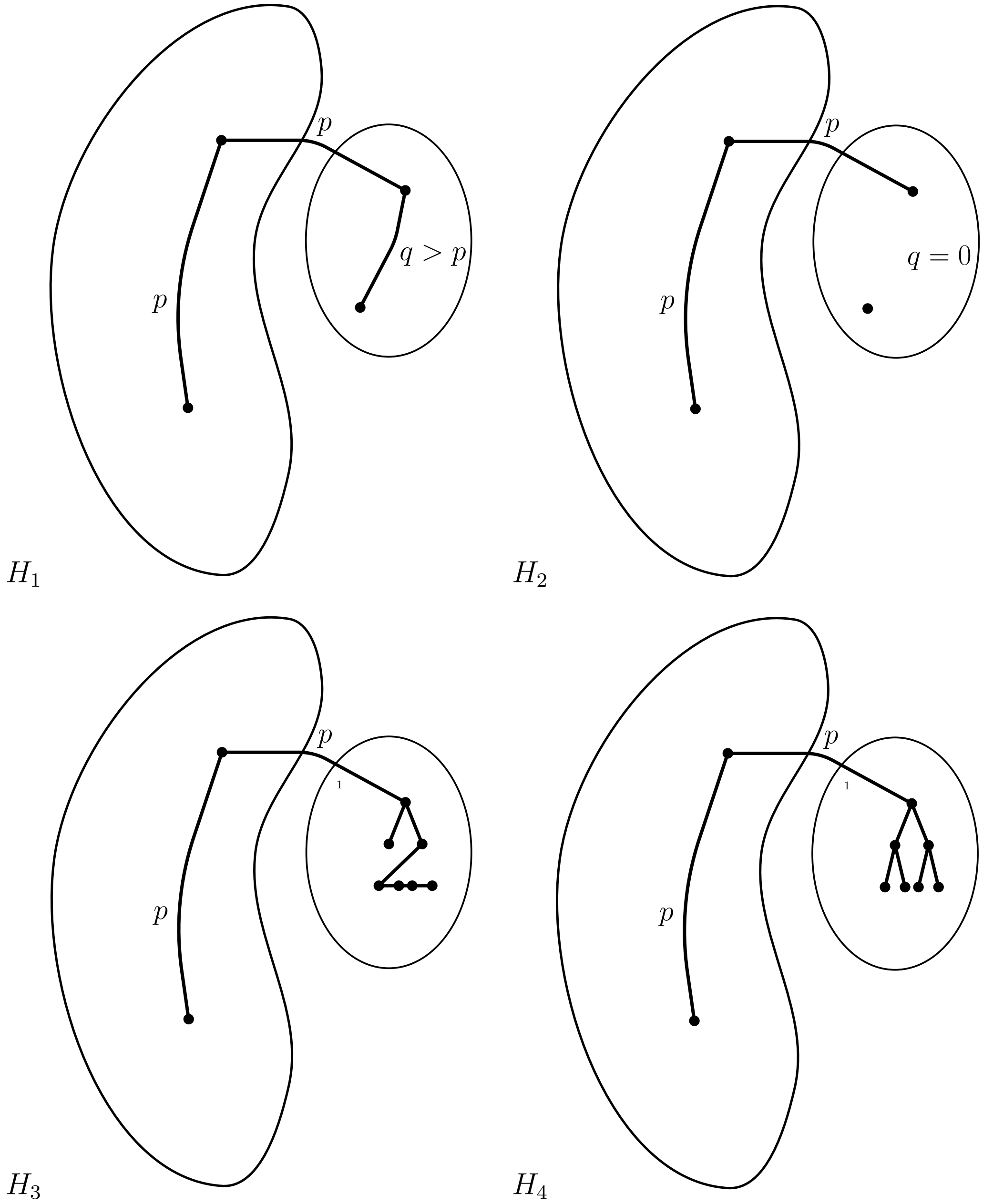

Figure 6.5: Depictions of the anomalous graphs investigated. $H_1 : \kappa(n, p, \gamma, q)$ where $q > p$, the *kidney-egg* graph. $H_2 : \kappa_c(n, p, \gamma, \rho = 0)$, the *disappearing-egg* graph. $H_3 : \kappa_c(n, p, \gamma, \rho = 1)$, the *kidney-line* graph. $H_4 : \kappa_c(n, p, \gamma, \rho = 2)$, the *kidney-tree* graph.

1 1

about its form. Kernel density estimates converge to the true distribution faster than histograms, but this can come at considerable computational and storage costs [310] especially for large sample sizes in a multivariate setting. Adaptive histograms with variable bin widths offer an intriguing compromise, but finding an *optimal* adaptive grid is difficult in practice, and *ad hoc* methods that are easier to implement "need not be better and in fact can be much worse" [80].

Given a set of graphs, $\mathbb{G} = \{G_1 = (V_1, E_1), \ldots, G_s = (V_s, E_s)\}$, each of order n, let a frequency histogram of D vertex invariants be defined by parameters $\boldsymbol{\theta} = \{\boldsymbol{b}_1, c_1, \ldots, \boldsymbol{b}_m, c_m\}$ of bin centroids $\boldsymbol{b}_j \in \mathbb{R}^D$ with corresponding counts,

$$c_j = \sum_{i=1}^{s} \sum_{v \in V_i} \mathbb{1}\left(j = \arg\min_k d(\boldsymbol{b}_k, \boldsymbol{\psi}(v, G_i))\right), \tag{6.11}$$

using the indicator function $\mathbb{1}$. We convert this to a density estimate using add-one smoothing,

$$p_{\mathbb{G}}(\boldsymbol{\psi}; \boldsymbol{\theta}) = \frac{c_j + \frac{1}{m}}{ns + 1}, \tag{6.12}$$

where $j = \arg\min_k \; d(b_k, \boldsymbol{\psi})$. This allows for an arbitrary number of vertex invariants, although the sparsity of their joint distribution is known to be problematic if $D > 5$ [311]. In this work, we chose fixed-bin histograms to model the joint distribution of each pair of vertex invariants.

6.4.3.2 Cross Entropy

To measure the abnormality of an observed graph $G^\star = (V^\star, E^\star)$ of order n, we can efficiently compute the negative average log likelihood of its vertex invariants

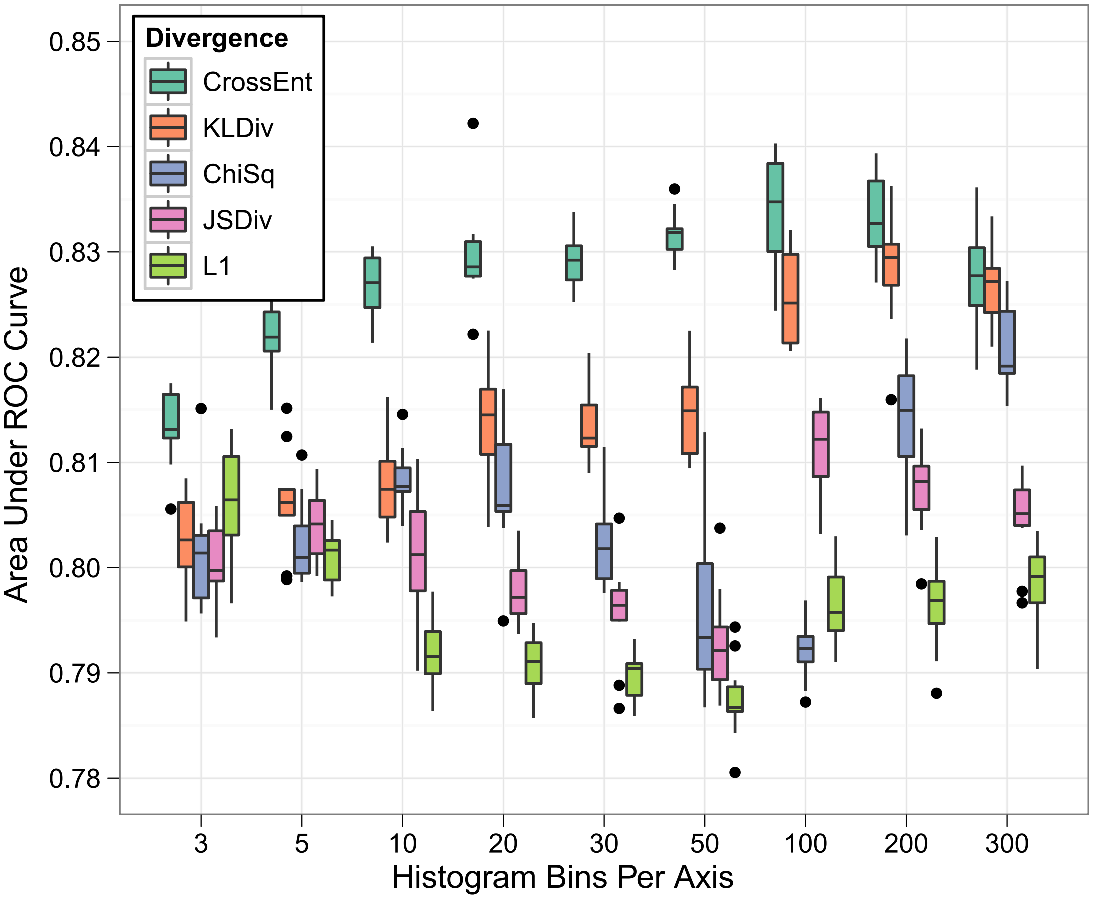

Figure 6.6: Performance of several methods of histogram comparison while varying the number of bins per axis when using the clustering coefficient and PageRank vertex invariants. Boxplots depict the interquartile range of the area under ROC curves for 10 resampling experiments of the pooled test condition where half of the anomalous graphs are sampled from those with a region of increased connectivity (H_1) and half are sampled from those with a region of decreased connectivity ($H_{\{2,3,4\}}$).

under the model of $\mathcal{N}$ normal graphs,

$$\begin{aligned} & -\frac{1}{n}\sum_{v\in V^\star}\log p_{\mathcal{N}}\left(\boldsymbol{\psi}(v,G^\star)\right) && (6.13)\\ \approx\ & -\sum_{i=1}^{m} p_{\{G^\star\}}(\boldsymbol{b}_i)\log p_{\mathcal{N}}(\boldsymbol{b}_i) && (6.14)\\ =\ & H(p_{\{G^\star\}},p_{\mathcal{N}}) && (6.15)\end{aligned}$$

using the equivalent cross entropy between histograms.

6.4.3.3 Other Methods

We tested several other methods of histogram comparison (Chapter 4) including the Kullback-Liebler divergence, χ^2 test statistic, Jensen-Shannon divergence, and L_1 norm, but none were able to outperform cross entropy (Figure 6.6).

6.5 Monte Carlo Experiments

Since we only consider random graphs in this work, the asymptotic distributions of some vertex invariants can be found analytically, especially for H_0 [294]. However, invariant distributions of finite graphs are typically known only for an extremely small number of vertices and those for $H_{\{1,2,3,4\}}$ would be even more complex. Thus, we estimate performance via Monte Carlo simulation.

We explore anomaly detection against five compound alternatives, namely each of $H_{\{1,2,3,4\}}$, along with their pooled combination. We first generate the set $\mathcal{N}$ of

$R = 1000$ graphs according to $H_0 : ER(n = 1000, p = 0.1)$. Another set of graphs are generated according to the same process for testing. We also generate a large set of anomalous graphs according to H_A, which depends on the test condition described below. We conduct 10 trials, each time randomly sampling 1,000 graphs from H_0 and 10,000 graphs from H_A. For each method of anomaly detection we compute the test statistic $T_{\mathcal{N}} : \mathcal{G} \to \mathbb{R}$ of each graph and reject the null for large values. We compute the area under the receiver operating characteristic (ROC) curve (AUC) to assess performance across the range of possible thresholds.

In order to assess the performance of detecting *kidney-egg* anomalies, we randomly sample graphs for H_A from $H_1 : \kappa(n = 1000, p = 0.1, \gamma, q)$ with γ drawn uniformly from $\{5, 10, \ldots, 100\}$ and q from $\{0.15, 0.20, \ldots, 1.00\}$. When assessing the detection of each type of *kidney-void* graph, we randomly sample graphs for H_A from $\kappa_c(n = 1000, p = 0.1, \gamma, \rho)$ with γ drawn uniformly from $\{5, 10, \ldots, 200\}$ and $\rho = 0$ for H_2, $\rho = 1$ for H_3, and $\rho = 2$ for H_4.

Our primary goal is to find a test statistic that can robustly reject the null hypothesis when presented with a graph from any of the four types of anomalies. To assess this, we compute the AUC when graphs are sampled with equal probability from H_1 and the set $H_{\{2,3,4\}}$. This is referred to as the pooled test condition.

As a baseline, we evaluate the performance of eight graph invariants when detecting anomalies from $H_{\{1,2,3,4\}}$ along with their pooled combination. We compare this to our family of methods comprising the Cartesian product of (a) all $\binom{8}{2}$ pairs of

Table 6.1: Median of the area under 10 ROC curves

Alternative Hypothesis		H_1	H_2	H_3	H_4	Pooled
Graph Invariant	MaxDegree	0.770	0.515	0.519	0.516	0.644
	Size	0.782	0.809	0.783	0.783	0.786
	AvgPathLength	0.788	0.813	0.788	0.784	0.792
	MADg	0.788	0.809	0.783	0.784	0.789
	MADe	0.802	0.801	0.776	0.774	0.792
	ScanStat	0.807	0.562	0.553	0.555	0.681
	Triangles	0.812	0.805	0.779	0.779	0.801
	AvgCC	0.816	0.787	0.761	0.756	0.792
Distrib.	Triangles,CC$_{100\text{x}100}$	**0.869**	0.798	0.763	0.764	**0.822**
	Locality,Degree$_{3\text{x}3}$	0.772	**0.849**	**0.818**	**0.820**	0.802
	Locality,Closeness$_{3\text{x}3}$	0.766	**0.846**	**0.821**	**0.820**	0.797
	PageRank,CC$_{100\text{x}100}$	**0.862**	**0.831**	**0.797**	**0.795**	**0.835**

vertex invariants using (b) $\{3, 5, 10, 20, 30, 50, 100, 200, 300\}$ bins per axis when comparing histograms via (c) cross entropy, Kullback-Leibler divergence, χ^2 test statistic, Jensen-Shannon divergence, and L_1 norm. We jointly optimized these three parameters and show the interquartile range of AUC for the top performing systems along with that of the graph invariants in Figure 6.7. We also report the median AUC for these systems across all test conditions in Table 6.1, emphasizing in bold the performance of distributional systems that perform significantly better ($p < 0.05$) than each graph invariant using a two-sided Wilcoxon paired-sample signed rank test for the 10 trials.

Among the vertex invariants investigated here, clustering coefficient is the least correlated with measures of vertex centrality in Erdös-Rényi graphs. When optimizing for the detection of increased activity in H_1, clustering coefficient therefore provides complementary information to the number of triangles involving each vertex. However, when optimizing for the detection of decreased activity in $H_{\{2,3,4\}}$, locality is chosen along with another measure of vertex centrality (degree or closeness) even though they are highly correlated.

While clustering coefficient is not used by the systems optimized for detecting decreased activity, it is used in the top performing system for the pooled test condition along with PageRank. These vertex invariants provide complimentary information about local neighborhood connectivity and global centrality, both of which are useful when detecting anomalies that can have regions of increased or decreased activity

(Figure 6.8). When computing the cross entropy between distributions of these invariants using 100 histogram bins per axis, this system achieves an overall median AUC of 0.835 for the pooled test condition, which is significantly greater ($p = 0.002$) than each graph invariant. It also significantly outperforms the graph invariants when testing against each anomaly type separately. When detecting H_1 graphs, the median AUC of 0.862 was significantly greater ($p = 0.002$) than the top performing graph invariant, average clustering coefficient (median AUC = 0.816). Performance improvements were also significant for H_2 (median AUC = 0.831, $p = 0.002$), H_3 (median AUC = 0.797, $p = 0.014$), and H_4 (median AUC = 0.795, $p = 0.006$) when compared to average path length, which was the top performing graph invariant with median AUCs of 0.813, 0.788, and 0.784, respectively.

Considerable research has been done to optimize histogram bins for a given amount of data, especially for a single dimension [310, 312]. In the multivariate setting, optimal bin size also depends on the correlation coefficient between variables [313] along with the measure of divergence between histograms when performing anomaly detection [205]. We thus left it as another free variable in the joint optimization and found that two highly correlated centrality measures, like locality and degree, are best utilized with as few as 3 histogram bins when detecting anomalies with decreased activity ($H_{\{2,3,4\}}$). When using uncorrelated vertex invariants like clustering coefficient and PageRank, 100 bins per axis yields the top performing overall system. While this is contrary to accepted theory where more bins are required to track the diagonal distributions of highly correlated variables, as we note that our goal is to optimize anomaly detection performance, not distributional accuracy.

For each test condition, we find that cross entropy is the best method of histogram comparison of those investigated as demonstrated in Figure 6.6. We attribute this to the tendency of anomalous vertex invariants to be in low density regions of the space making the likelihood-equivalent, cross entropy, a natural choice for measuring anomalousness. Also, the disparity in the amount of training and test data (10^6 and 10^3 measurements, respectively) is well accommodated by the cross entropy computation where the logarithmic emphasis is not performed on the poorly estimated $p_{\{G^\star\}}$.

While most of our investigation explores compound alternatives to assess overall performance, we show the median AUC for each simple alternate hypothesis in Figures 6.9 and 6.10. For the "chatter" alternative, the tradeoff between the γ vertices involved in the anomaly and their increased connectivity q is well covered in [294]. We display the AUC results here (Figure 6.9) to confirm that this performance metric is highly correlated with statistical power and to demonstrate that our new methods experience similar trends across this space of parameterized alternate hypotheses.

For the anomalous graphs with regions of decreased activity (Figure 6.10), performance tends to be better for $\rho = 0$ compared to $\rho = \{1, 2\}$. This is further demonstrated for the other methods in Figure 6.7 when comparing H_2 to $H_{\{3,4\}}$. This is likely due to the greater degree of connectivity change when $q = 0$ for $\rho = 0$

compared to $q \approx \frac{\gamma-1}{\binom{\gamma}{2}}$ for $\rho = \{1, 2\}$.

6.6 Conclusion

In this chapter, we demonstrate the generality of our approach by achieving state-of-the-art performance when detecting anomalies in graphs. Our approach to the antithetical goal of graph matching is in some ways a generalization to Kosinov and Caelli's projection of vertices into the eigenspaces of the adjacency matrix [289]. However, we are not limited to eigenvectors and can use any vertex invariant or attribute. Given a graph database with existing vertex attributes, our approach scales well since fixed-width histograms can be compared using "embarrassingly parallel"[3] algorithms unlike the agglomerative clustering used by Kosinov and Caelli.

By estimating the joint distribution of two vertex invariants using histograms and assessing its divergence from normality, we significantly outperform all available graph invariants when detecting anomalies with a local region of increased or decreased connectivity. We demonstrate that clustering coefficient and PageRank provide complementary information about vertices, and modeling their distribution makes for a robust system capable of detecting multiple types of anomalous graphs. While we chose these features for their performance in the pooled test condition, they also outperform all available graph invariants when constraining the problem to the detection of each of four anomalous graph types.

[3] This phrase was first found in [314] although its etymology remains unknown.

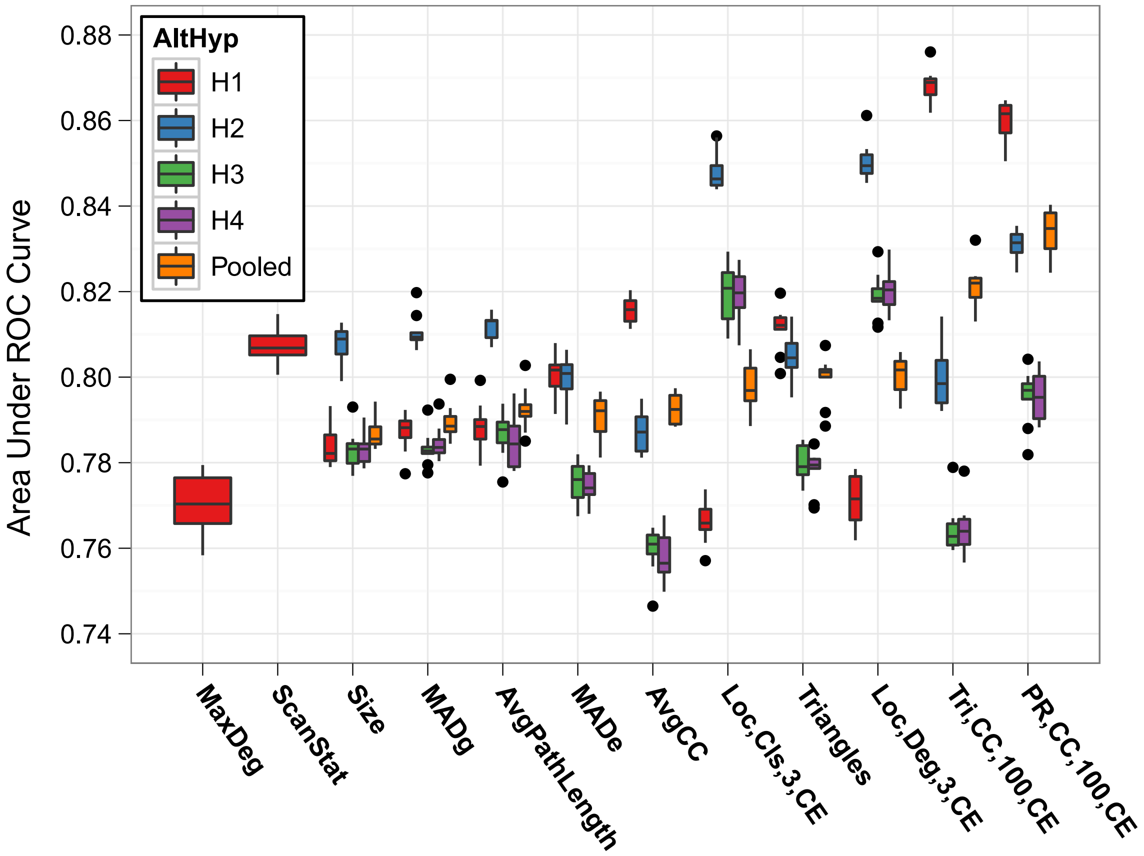

Figure 6.7: Performance of graph invariants and the best systems using distributions of vertex invariants. Boxplots depict the interquartile range of the area under ROC curves for 10 resampling experiments with each type of anomaly. The anomalous graphs present in the H_1 test condition are $\kappa(n = 1000, p = 0.1, \gamma, q)$ graphs with γ drawn uniformly from $\{5, 10, \ldots, 100\}$ and q from $\{0.15, 0.20, \ldots, 1.0\}$. The anomalous graphs in the $H_{\{2,3,4\}}$ test conditions are $\kappa_c(n, p, \gamma, \rho)$ graphs with γ drawn uniformly from $\{5, 10, \ldots, 200\}$ and $\rho = 0$, $\rho = 1$, and $\rho = 2$ for H_2, H_3, and H_4, respectively. We do not show results for max degree and scan statistic for $H_{\{2,3,4\}}$ since their construction gives them no power for these type of anomalies.

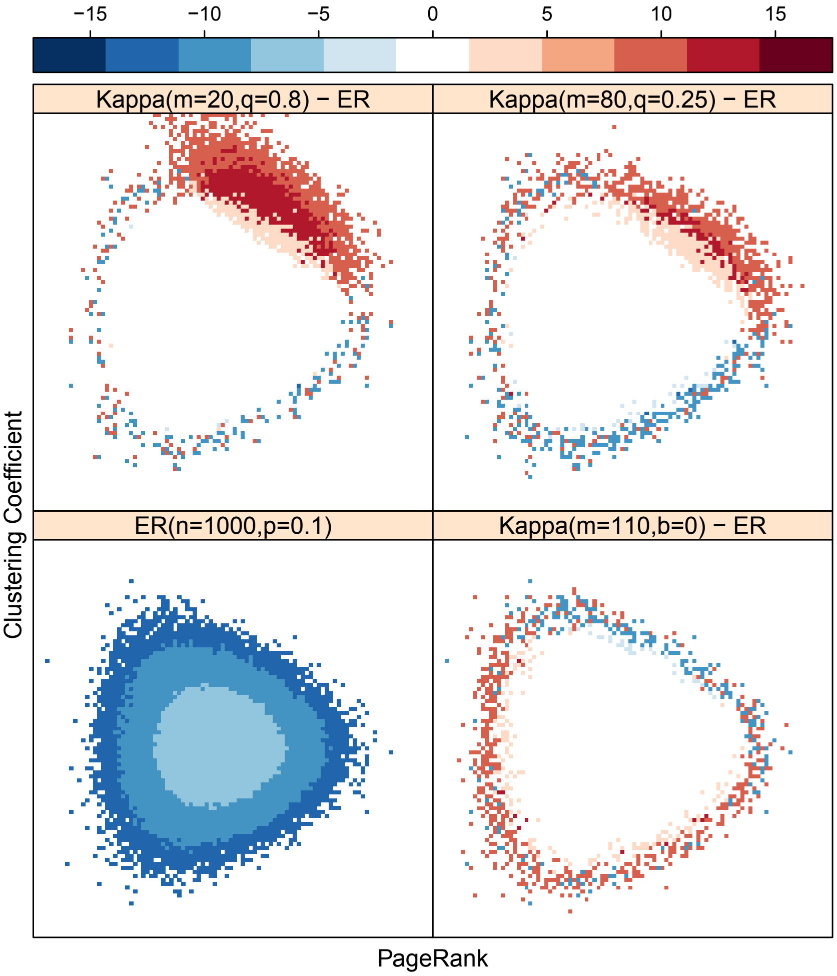

Figure 6.8: Log probability histogram of clustering coefficient and PageRank vertex invariants. Lower left panel shows their distribution for H_0: Erdös-Rényi graphs with $n = 1000$ vertices where each edge exists independently with probability $p = 0.1$. Upper panels show log differences between histograms of invariants from H_0 and H_1 (*kidney-egg*) where $H_1 = \kappa(n = 1000, p = 0.1, \gamma = 20, q = 0.8)$ on the left and $H_1 = \kappa(n = 1000, p = 0.1, \gamma = 80, q = 0.25)$ on the right. Lower right panel shows log differences between histograms of invariants from H_0 and H_2 (*disappearing-egg*) where $H_2 = \kappa_c(n = 1000, p = 0.1, \gamma = 110, \rho = 0)$. The histograms shown here have 100 bins per axis resulting in the best performance of the distributional methods for the pooled test condition. The κ parameters were chosen to show different conditions leading to approximately the same AUC of 0.99 when using cross entropy for histogram comparison.

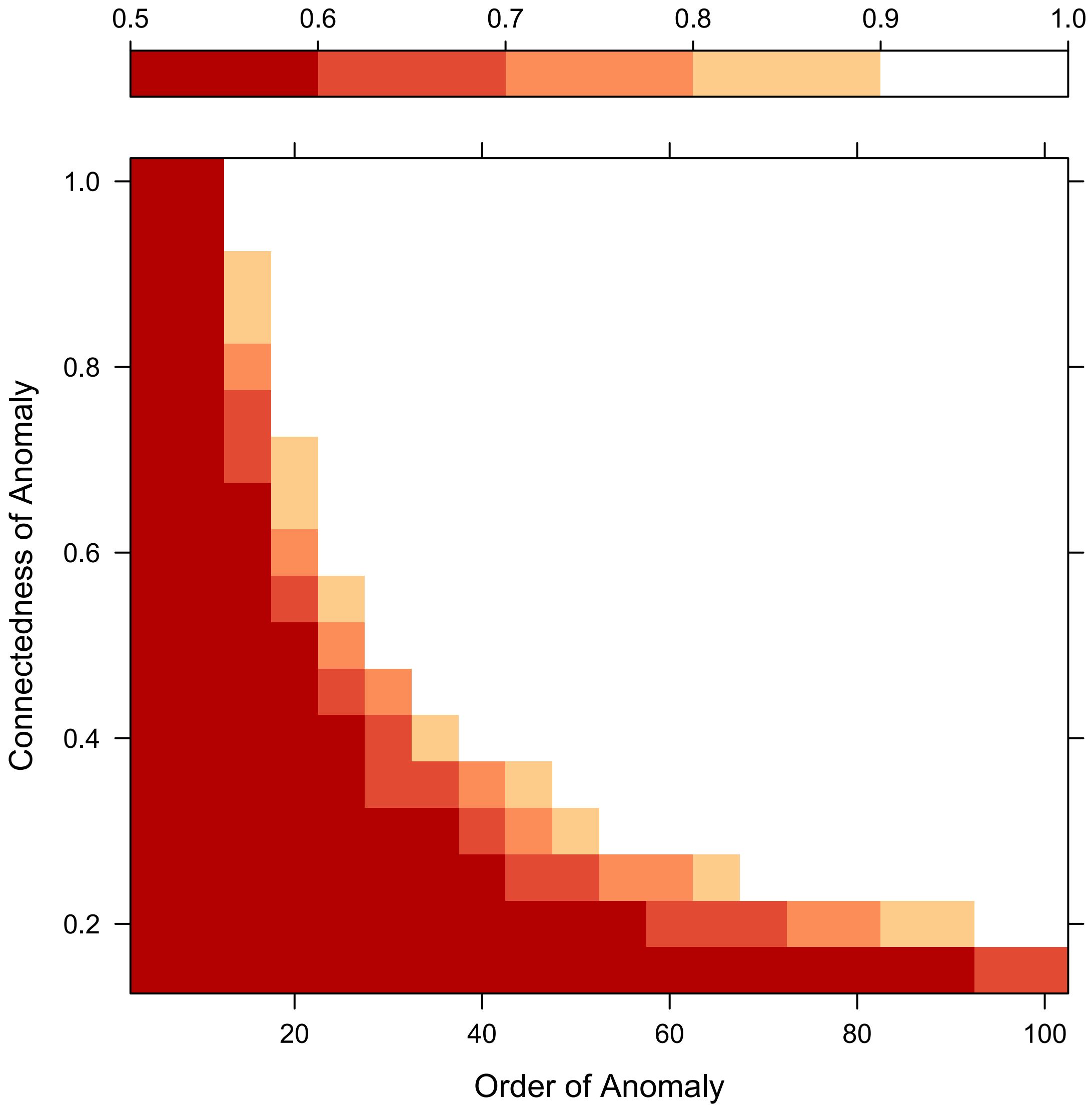

Figure 6.9: Area under the ROC curve as a function of the anomalous subgraph's order m and connectivity q when using histograms with 100 bins per axis to compare the joint distribution of clustering coefficient and PageRank vertex invariants between $ER(n = 1000, p = 0.1)$ and $\kappa(n = 1000, p = 0.1, \gamma, q)$ random graphs.

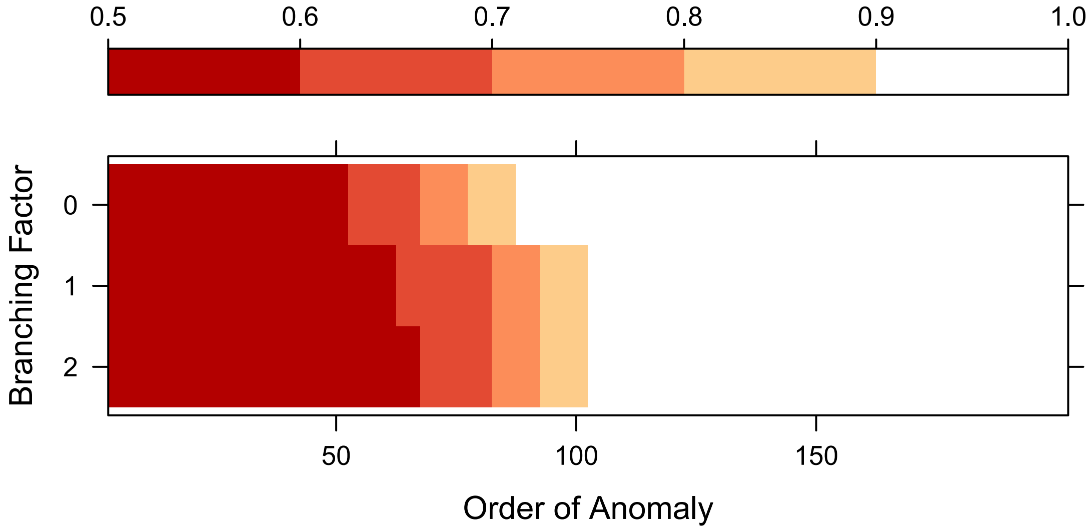

Figure 6.10: Area under the ROC curve when varying the anomalous subgraph's order γ and branching factor ρ when using histograms with 100 bins per axis to compare the joint distribution of clustering coefficient and PageRank vertex invariants between $ER(n = 1000, p = 0.1)$ and $\kappa_c(n = 1000, p = 0.1, \gamma, \rho)$ random graphs.

Chapter 7

Conclusion

The human ability to *know what we do not know* is a concept often overlooked in the development of pattern recognition systems that discriminate between subpopulations of normal data in controlled experiments. Robustness to anomalies often comes at some cost to classification performance on normal data, so it is not in a researcher's best interest to worry about something that will not present itself in their experiments. It is only when using pattern recognizers in real-world applications that extra care must be taken to avoid making egregious errors, which can quickly lead to the distrust of automated systems.

Since the capability of recognizing the unknown is straightforward for humans, we have taken some cues from biological systems. When performing tasks such as reading and scene perception [315], humans extract *local* information during eye fixations using the fovea centralis, the part of the retina responsible for visual acuity. Most visual perception occurs during fixational eye movements [316] in between short and rapid movements called saccades. These saccades relocate the point of fixation elsewhere so additional information can be extracted. The visual system then aggregates all of this information to build up a *global* representation of the entire scene.

Our approach to distributional anomaly detection is analogous. We perform feature extraction to gather *local* information and then aggregate all such evidence to form a *global* density estimate which can be compared to a model of normal data. When determining whether a 30 second segment of audio contains conversational speech, we extract syllable rate estimates from each 500 ms frame of audio. We then use those local features to estimate the global distribution of syllable rate for the entire segment and compare it to normal conversational speech. When determining whether a random graph contains an anomalous region of excessive or negligible connectivity we perform a similar process. We focus our attention on each vertex, extract local features that are invariant up to isomorphism, estimate the global distribution of vertex invariants for the graph in question, and compare it to a non-anomalous distribution.

7.1 Summary

Throughout our investigations, we demonstrated that our approach of comparing distributions outperformed log likelihood baselines. In retrospect, this is not surprising since the anomalies we were investigating often occupied the same regions of the feature space as normal data. The difference was typically found in the spread and shape of the feature distributions between normal and anomalous data; a difference that is better captured by estimating the statistical divergence. And while our goal was to detect anomalies, we did not learn from them by studying any specific examples, both because they are often rare and difficult to find, and because they can result from a multitude of poorly understood causes. We therefore developed methods for acoustics and graphs that could detect anything that markedly deviates from normal data using robust estimates of statistical divergence.

Anomaly detection is a longstanding problem with many applications in signal processing. In the interest of developing a robust speech activity detector, we reformulated the problem into acoustic anomaly detection (AAD) in Chapter 3 using a partially supervised strategy learning from normal acoustics. We presented results of explorations into files that were found to be anomalous in a normal speech corpus, and conversely reported findings of fairly normal audio discovered unexpectedly in a variety of incorrectly decoded files. We also showed that partially-supervised distributional anomaly detection significantly outperformed the log probability baseline.

In Chapter 4, we reformulated AAD into unsupervised accommodation learning and allowed for anomalous contamination of the training data. Nonparametric histograms were found to be naturally robust to contamination and we demonstrated state-of-the-art performance when using the L_1 distance on 30 seconds of audio when subjected to 33% training contamination to achieve an equal error rate of 2.7%.

In the interest of being able to robustly compare two distributions in a high dimensional space we returned to Gaussian mixture models (GMMs) in Chapter 5. We analyzed what caused the Kullback-Leibler (KL) divergence between GMMs to break down in the face of training contamination and came up with a promising solution: By trimming one quarter of the most divergent Gaussians from the mixture model, we were able to statistically remove the effect of training data contamination levels as high as 33%. We significantly outperformed the untrimmed KL approximation for contamination levels of 10% and above, reducing the equal error rate from 33.8% to 6.4% when subjected to 33% training contamination.

In Chapter 6, we demonstrated the generality of our approach to anomaly detection by considering the application of inexact graph matching. To do this, we compared distributions of vertex invariants to those obtained from non-anomalous graphs. We considered homogeneous Erdös-Rényi random graphs to be non-anomalous, and compared them to four classes of heterogeneous alternatives, where a subset of the vertices were connected according to a different process with excessive or negligible connectivity. In this context, we demonstrated superior performance to available state-of-the-art approaches against a specific type of anomaly and further demonstrated superior generalization to entire classes of graph anomalies.

7.2 Future Work

Our work in acoustic anomaly detection focused on the perspective of a speech activity detector that used envelope modulation features known to be indicative of syllabic rate. Recently, long-term acoustic frames have been used to model the amplitude variations in critical bands as a way of extending the modulation spectrum to the higher dimensions necessary for many other speech processing tasks. Frequency domain linear prediction (FDLP) [317] is one such technique that has proven to be robust to noise and reverberation for phoneme recognition, speaker verification, and audio compression [318–320]. Leveraging acoustic anomaly detection in these contexts could allow for even more accurate confidence estimation. It could also foster the development of a more robust and integrated speech processing system, rather than the traditional pipelined approach where speech activity labels are passed to other subsystems without confidence estimation or feedback loops.

Another promising area of research could be combining temporal and spectral features while using *missing feature theory* to concentrate on the salient pieces of information and de-emphasize those corrupted by noise. Missing feature theory was originally developed to deal with occluded objects in computer vision, but Cooke [321] adapted the techniques to acoustic scene analysis [322–324]. It has since been used to improve speaker identification [325, 326] and speech recognition [327–329]. One of the largely unexplored areas of missing feature theory is how to accurately estimate the noise mask without *a priori* knowledge. We believe that our approach to acoustic anomaly detection could be extended to improve noise mask estimation by temporally localizing abnormalities that deviate markedly from normal audio.

The task of localizing anomalies is outside the scope of this work, but it would be useful for both acoustics and graphs. While many acoustic anomalies are transient, it is still important to identify and exclude them from downstream processing without having to discard an entire segment. Anomaly localization is even more important in graphs because the anomaly is often what we are interested in finding and studying, rather than just discarding as a nuisance to other systems.

Local anomaly detection could benefit from recent developments in distribution-free methods of comparison that require less data than density estimation. One approach uses k-nearest neighborhoods for multi-dimensional divergence estimation [330], while another uses locality sensitive hashing [331] to compare data in higher dimensions without constructing a density estimate [332].

Regardless of the approach to estimating divergence, it is important to note that we do not require a good match to a recognizable distribution. By design, our process can decide that an anomaly is unlike anything we have seen before. We believe that such a system can and should be used as part of a more robust pattern recognition system that can detect the “unknown unknowns” [1] and properly react to any “black swan events” [2].

Bibliography

[1] D. Rumsfeld. (2002) Transcript: DoD news briefing - Secretary Rumsfeld and Gen. Myers. [Online]. Available: http://www.defense.gov/transcripts/transcript.aspx?transcriptid=2636

[2] N. N. Taleb, *The Black Swan: The Impact of the Highly Improbable.* Random House, 2007.

[3] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[4] J. Gehrke, P. Ginsparg, and J. Kleinberg, "Overview of the 2003 KDD Cup," *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 2, pp. 149–151, 2003.

[5] B. Klimt and Y. Yang, "Introducing the Enron corpus," in *CEAS*, 2004.

[6] J. K. Francis, "Noni, Morinda citrifolia," U.S. Department of Agriculture, Forest Service, 2002.

[7] A. Hirazumi and E. Furusawa, "An immunomodulatory polysaccharide-rich

substance from the fruit juice of Morinda citrifolia (noni) with antitumour activity," *Phytotherapy Research*, vol. 13, no. 5, pp. 380–387, 1999.

[8] B. S. Nayak, G. N. Isitor, A. Maxwell, V. Bhogadi, and D. D. Ramdath, "Wound-healing activity of Morinda citrifolia fruit juice on diabetes-induced rats." *Journal of wound care*, vol. 16, no. 2, pp. 83–86, Feb. 2007.

[9] M.-Y. Y. Wang, B. J. West, C. J. Jensen, D. Nowicki, C. Su, A. K. Palu, and G. Anderson, "Morinda citrifolia (Noni): a literature review and recent advances in Noni research." *Acta pharmacologica Sinica*, vol. 23, no. 12, pp. 1127–1141, Dec. 2002.

[10] C. Younos, A. Rolland, J. Fleurentin, M.-C. Lanhers, R. Misslin, and F. Mortier, "Analgesic and behavioural effects of morinda citrifolia," *Planta Med*, vol. 56, no. 05, 1990.

[11] F. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, pp. 1–21, 1969.

[12] H. Bulthoff and S. Edelman, "Psychophysical support for a two-dimensional view interpolation theory of object recognition," *Proceedings of the National Academy of Sciences*, vol. 89, no. 1, p. 60, 1992.

[13] A. Wong and W. Hayward, "Constraints on view combination: effects of self-occlusion and differences among familiar and novel views," *Journal of Experi-*

mental Psychology: Human Perception and Performance, vol. 31, no. 1, p. 110, 2005.

[14] A. Friedman, D. Waller, T. Thrash, N. Greenauer, and E. Hodgson, “View combination: A generalization mechanism for visual recognition,” *Cognition*, 2011.

[15] D. Simons and R. Wang, “Perceiving real-world viewpoint changes,” *Psychological Science*, vol. 9, no. 4, pp. 315–320, 1998.

[16] M. Tarr and S. Pinker, “Mental rotation and orientation-dependence in shape recognition,” *Cognitive psychology*, vol. 21, no. 2, pp. 233–282, 1989.

[17] M. Tarr and H. Bulthoff, “Image-based object recognition in man, monkey and machine,” *Cognition*, vol. 67, no. 1-2, pp. 1–20, 1998.

[18] T. Poggio and S. Edelman, “A network that learns to recognize 3D objects,” *Nature*, vol. 343, no. 6255, pp. 263–266, 1990.

[19] J. Stone, “Object recognition: View-specificity and motion-specificity,” *Vision research*, vol. 39, no. 24, pp. 4032–4044, 1999.

[20] T. M. Kamm, “Active learning for acoustic speech recognition modeling,” Ph.D. dissertation, Johns Hopkins University, 2004.

[21] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[22] J. Dean, G. S. Corrado, R. Monga, K. Chen, M. Devin, Q. Le, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Ng, "Large scale distributed deep networks," *NIPS*, 2012.

[23] L. Deng, D. Yu, and J. Platt, "Scalable stacking and learning for building deep architectures," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on.* IEEE, 2012, pp. 2133–2136.

[24] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[25] Q. V. Le, R. Monga, M. Devin, K. Chen, G. S. Corrado, J. Dean, and A. Y. Ng, "Building high-level features using large scale unsupervised learning," in *Proceedings of the ICML 2012 workshop*, 2012.

[26] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[27] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. Interspeech*, 2011, pp. 437–440.

[28] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Large vocabulary continuous speech recognition with context-dependent dbn-hmms," in *Acoustics, Speech and Sig-*

nal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011, pp. 4688–4691.

[29] T. N. Sainath, B. Kingsbury, B. Ramabhadran, P. Fousek, P. Novak, and A.-r. Mohamed, "Making deep belief networks effective for large vocabulary continuous speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on.* IEEE, 2011, pp. 30–35.

[30] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *ICASSP*, 2013.

[31] G. Anthes, "Deep learning comes of age," *Communications of the ACM*, vol. 56, no. 6, pp. 13–15, 2013.

[32] V. N. Vapnik, *Statistical learning theory.* John Wiley and Sons, Inc., 1998.

[33] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *Computational learning theory.* Springer, 1995, pp. 23–37.

[34] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[35] ——, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[36] D. J. Hand, “Supervised classification and tunnel vision,” *Applied Stochastic Models in Business and Industry*, vol. 21, no. 2, pp. 97–109, 2005.

[37] ——, “Classifier technology and the illusion of progress,” *Statistical Science*, vol. 21, no. 1, pp. 1–14, 2006.

[38] G. E. P. Box and N. R. Draper, *Empirical model-building and response surface.* New York, NY, USA: John Wiley and Sons, Inc., 1986.

[39] G. Box, “Non-normality and tests on variances,” *Biometrika*, vol. 40, no. 3-4, pp. 318–335, 1953.

[40] J. Tukey, “A survey of sampling from contaminated distributions,” *Contributions to probability and statistics*, pp. 448–485, 1960.

[41] P. Huber, “Robust estimation of a location parameter,” *The Annals of Mathematical Statistics*, pp. 73–101, 1964.

[42] F. Hampel, “A general qualitative definition of robustness,” *The Annals of Mathematical Statistics*, pp. 1887–1896, 1971.

[43] E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel, “Ontonotes: the 90% solution,” in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers.* Association for Computational Linguistics, 2006, pp. 57–60.

[44] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” in *International Conference on the World Wide Web*, 1998, pp. 107–117.

[45] L. Page, S. Brin, R. Motwani, and T. Winograd, “The pagerank citation ranking: Bringing order to the web.” Stanford InfoLab, Technical Report 1999-66, November 1999.

[46] B. Martins and M. Silva, “Language identification in web pages,” in *Proceedings of the 2005 ACM symposium on Applied computing*. ACM, 2005, pp. 764–768.

[47] F. Xia, W. Lewis, and H. Poon, “Language id in the context of harvesting language data off the web,” in *Proceedings of The 12th Conference of the European Chapter of the Association of Computational Linguistics (EACL 2009)*, 2009.

[48] L. Von Ahn and L. Dabbish, “Labeling images with a computer game,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2004, pp. 319–326.

[49] L. Von Ahn, M. Kedia, and M. Blum, “Verbosity: a game for collecting common-sense facts,” in *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 2006, pp. 75–78.

[50] L. von Ahn, “Duolingo: learn a language for free while helping to translate the

web," in *Proceedings of the 2013 international conference on Intelligent user interfaces.* ACM, 2013, pp. 1–2.

[51] C. Callison-Burch and M. Dredze, "Creating speech and language data with amazon's mechanical turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk.* Association for Computational Linguistics, 2010, pp. 1–12.

[52] V. Raykar, S. Yu, L. Zhao, A. Jerebko, C. Florin, G. Valadez, L. Bogoni, and L. Moy, "Supervised learning from multiple experts: Whom to trust when everyone lies a bit," in *Proceedings of the 26th Annual International Conference on Machine Learning.* ACM, 2009, pp. 889–896.

[53] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng, "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, 2008, pp. 254–263.

[54] A. Sorokin and D. Forsyth, "Utility data annotation with amazon mechanical turk," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on.* IEEE, 2008, pp. 1–8.

[55] M. Marge, S. Banerjee, and A. Rudnicky, "Using the amazon mechanical turk for transcription of spoken language," in *Acoustics Speech and Signal Processing*

(ICASSP), 2010 IEEE International Conference on. IEEE, 2010, pp. 5270–5273.

[56] C. Vondrick, D. Ramanan, and D. Patterson, “Efficiently scaling up video annotation with crowdsourced marketplaces,” *Computer Vision–ECCV 2010*, pp. 610–623, 2010.

[57] S. Novotney and C. Callison-Burch, “Cheap, fast and good enough: Automatic speech recognition with non-expert transcription,” in *Association for Computational Linguistics*, 2010, pp. 207–215.

[58] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel, “The automatic content extraction (ACE) program–tasks, data, and evaluation,” in *Proceedings of LREC*, vol. 4, 2004, pp. 837–840.

[59] C. Yong and S. K. Foo, “A case study on inter-annotator agreement for word sense disambiguation,” in *SIGLEX Workshop On Standardizing Lexical Resources*, 1999.

[60] B. W. Silverman, *Density estimation for statistics and data analysis.* Chapman & Hall/CRC, 1986, vol. 26.

[61] “Nix v. Hedden,” Supreme Court of the United States, p. 149, 1893.

[62] J. Le, A. Edmonds, V. Hester, and L. Biewald, “Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribu-

tion," in *SIGIR 2010 workshop on crowdsourcing for search evaluation*, 2010, pp. 21–26.

[63] C. Callison-Burch, "Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1.* Association for Computational Linguistics, 2009, pp. 286–295.

[64] L. Atlas, D. Cohn, and R. Ladner, "Training connectionist networks with queriers and selective sampling," in *NIPS.* Morgan Kaufman, 1990, pp. 566–573.

[65] K. Tumer and J. Ghosh, "Estimating the Bayes error rate through classifier combining," in *Pattern Recognition, Proceedings of the 13th International Conference on*, vol. 2. IEEE, 1996, pp. 695–699.

[66] D. J. Hand, *Construction and assessment of classification rules.* John Wiley and Sons, Inc., 1997.

[67] A. Jamain and D. J. Hand, "Mining supervised classification performance studies: A meta-analytic investigation," *Journal of Classification*, vol. 25, no. 1, pp. 87–112, 2008.

[68] I. J. Good, "Some statistical applications of Poisson's work," *Statistical Science*, vol. 1, no. 2, 1986.

[69] M. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary Physics*, vol. 46, no. 5, pp. 323–351, 2005.

[70] V. Pareto, *Cours d'Economie Politique.* F. Rouge and Cie., Lausanne, Switzerland, 1897, vol. 2.

[71] G. K. Zipf, *Human Behaviour and the Principle of Least-Effort.* Cambridge MA: Addison-Wesley, 1949.

[72] A. J. Lotka, "The frequency distribution of scientific productivity," *J Wash Acad Sci*, vol. 16, no. 12, pp. 317–323, 1926.

[73] B. Mandelbrot, "The variation of certain speculative prices," *The Journal of Business*, vol. 36, no. 4, pp. pp. 394–419, 1963.

[74] B. Mandelbrot and H. M. Taylor, "On the distribution of stock price differences," *Operations Research*, vol. 15, no. 6, pp. pp. 1057–1062, 1967.

[75] J. Phillips, *Robert Wadlow: The Unique Life of the Boy Who Became the World's Tallest Man.* CreateSpace, 2010.

[76] A. Kornai, *Mathematical Linguistics.* Springer, 2008.

[77] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification.* Wiley-Interscience, 2001.

[78] S. Wasserman and G. Robins, "An introduction to random graphs, dependence graphs, and p*," in *Models and Methods in Social Network Analysis*, P. J.

Carrington, J. Scott, and S. Wasserman, Eds. Cambridge University Press, 2005.

[79] P. J. Bickel and K. A. Doksum, *Mathematical Statistics, Volume I.* Prentice Hall Englewood Cliffs, NJ, 2001.

[80] D. W. Scott, *Multivariate Density Estimation.* John Wiley and Sons, Inc., 1992.

[81] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[82] G. R. Terrell and D. W. Scott, "Variable kernel density estimation," *The Annals of Statistics*, pp. 1236–1265, 1992.

[83] J. R. Thompson and R. A. Tapia, *Nonparametric function estimation, modeling, and simulation.* Society for industrial and applied mathematics, 1987, no. 21.

[84] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *The Annals of Mathematical Statistics*, pp. 832–837, 1956.

[85] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[86] V. A. Epanechnikov, "Non-parametric estimation of a multivariate probability

density," *Theory of Probability & Its Applications*, vol. 14, no. 1, pp. 153–158, 1969.

[87] K. Pearson, *Maps and Chartograms.* KP:UCL, 1891, vol. 21, no. 49.

[88] R. A. Fisher, *Statistical Methods, Experimental Design, and Scientific Inference: A Re-issue of Statistical Methods for Research Workers, The Design of Experiments, and Statistical Methods and Scientific Inference.* Oxford University Press, 1990.

[89] D. E. Stokes, *Pasteur's quadrant: Basic science and technological innovation.* Brookings Inst Press, 1997.

[90] V. N. Vapnik, *The Nature of Statistical Learning Theory.* Springer, 2000.

[91] V. N. Vapnik and A. J. Chervonenkis, *Theory of pattern recognition.* Nauka, 1974.

[92] X. Zhu, "Semi-supervised learning literature survey," Computer Sciences, University of Wisconsin-Madison, Tech. Rep. 1530, 2007.

[93] K. Nigam, A. McCallumzy, S. Thruny, and T. Mitchelly, "Learning to classify text from labeled and unlabeled documents," *AAAI/IAAI*, 1998.

[94] O. Chapelle and A. Zien, "Semi-supervised classification by low density separation," *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 2004.

[95] B. Liu, W. Lee, P. Yu, and X. Li, "Partially supervised classification of text documents," in *ICML*, 2002, pp. 387–394.

[96] H. Hu, C. Sha, X. Wang, and A. Zhou, "A unified framework for semi-supervised pu learning," *World Wide Web*, pp. 1–18, 2013.

[97] C. Kılıç and M. Tan, "Positive unlabeled learning for deriving protein interaction networks," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 1, no. 3, pp. 87–102, 2012.

[98] C. Elkan and K. Noto, "Learning classifiers from only positive and unlabeled data," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 2008, pp. 213–220.

[99] F. Denis, R. Gilleron, and F. Letouzey, "Learning from positive and unlabeled examples," *Theoretical Computer Science*, vol. 348, no. 1, pp. 70–83, 2005.

[100] H. Yu, J. Han, and K.-C. Chang, "Pebl: Web page classification without negative examples," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 16, no. 1, pp. 70–81, 2004.

[101] W. S. Lee and B. Liu, "Learning with positive and unlabeled examples using weighted logistic regression," in *ICML*, 2003.

[102] X. Li and B. Liu, "Learning to classify text using positive and unlabeled data," in *IJCAI*, 2003.

[103] B. Liu, Y. Dai, X. Li, W. Lee, and P. Yu, "Building text classifiers using positive and unlabeled examples," *ICDM*, pp. 179–186, Nov. 2003.

[104] F. Denis, "PAC learning from positive statistical queries," in *Algorithmic Learning Theory*, 1998, pp. 112–126.

[105] F. Denis, R. Gilleron, A. Laurent, and M. Tommasi, "Text classification and co-training from positive and unlabeled examples," in *Proceedings of the ICML 2003 workshop: the continuum from labeled to unlabeled data*, 2003, pp. 80–87.

[106] F. Denis, R. Gilleron, M. Tommasi *et al.*, "Text classification from positive and unlabeled examples," in *Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU'02*, 2002, pp. 1927–1934.

[107] D. M. J. Tax, "One-class classification," Ph.D. dissertation, Delft University of Technology, 2001.

[108] Y. Chen, X. S. Zhou, and T. S. Huang, "One-class svm for learning in image retrieval," in *Image Processing, 2001. Proceedings. 2001 International Conference on*, vol. 1. IEEE, 2001, pp. 34–37.

[109] L. M. Manevitz and M. Yousef, "One-class svms for document classification," *The Journal of Machine Learning Research*, vol. 2, pp. 139–154, 2002.

[110] A. Ng and M. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes," in *NIPS*. Cambridge, MA: MIT Press, 2002.

[111] A. Dawid, "Properties of diagnostic data distributions," *Biometrics*, pp. 647–658, 1976.

[112] D. J. Hand, H. Mannila, and P. Smyth, *Principles of data mining.* MIT press, 2001.

[113] C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh, "Sample selection bias correction theory," in *Algorithmic Learning Theory.* Springer, 2008, pp. 38–53.

[114] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *ICASSP*, vol. 2, April 2003.

[115] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, April 1995.

[116] T. Jebara, *Machine Learning: Discriminative and Generative.* Springer, 2003.

[117] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, pp. 85–126, 2004.

[118] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, and J. Srivastava, "A comparative

study of anomaly detection schemes in network intrusion detection," in *SIAM International Conference on Data Mining*, 2003.

[119] S. Hickinbotham and J. Austin, "Novelty detection in airframe strain data," in *Proceedings of the 15th International Conference on Pattern Recognition*, vol. 2, 2000.

[120] R. J. Bolton and D. J. Hand, "Unsupervised profiling methods for fraud detection," in *Credit Scoring Conference*, 2001.

[121] J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi, "Topic detection and tracking with spatio-temporal evidence," in *Proceedings of the 25th European Conference on IR Research*, 2003.

[122] E. Keogh, S. Lonardi, and C. A. Ratanamahatana, "Towards parameter-free data mining," in *Proceedings of the tenth international conference on knowledge discovery and data mining*, 2004.

[123] D. Guthrie, L. Guthrie, B. Allison, and Y. Wilks, "Unsupervised anomaly detection," in *IJCAI*, 2007.

[124] P. Rousseeuw and A. Leroy, *Robust Regression and Outlier Detection.* John Wiley and Sons, Inc., 1987.

[125] T. Rossing, *Springer handbook of acoustics.* Springer, 2007.

[126] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32–80, 2001.

[127] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Muller, R. Huber, B. Andrassy, J. Bauer *et al.*, "Comparison of four approaches to age and gender recognition for telephone applications," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–1089.

[128] L. Cerrato, M. Falcone, and A. Paoloni, "Subjective age estimation of telephonic voices," *Speech Communication*, vol. 31, no. 2, pp. 107–112, 2000.

[129] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *ICASSP*, 1992.

[130] B. Bernstein, "Language and social class," *British journal of sociology*, pp. 271–276, 1960.

[131] D. Kolossa and R. Haeb-Umbach, *Robust Speech Recognition of Uncertain or Missing Data: Theory and Applications*. Springer, 2011.

[132] P. Dermody, "Human capabilities for speech processing in noise," in *Speech Processing in Adverse Conditions*, 1992.

[133] N. Jayant, B. McDermott, S. Christensen, and A. Quinn, "A comparison of

four methods for analog speech privacy," *Communications, IEEE Transactions on*, vol. 29, no. 1, pp. 18–23, 1981.

[134] N. Jayant, "Analog scramblers for speech privacy," *Computers & Security*, vol. 1, no. 3, pp. 275–289, 1982.

[135] J. Licklider and I. Pollack, "Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech," *The Journal of the Acoustical Society of America*, vol. 20, no. 1, pp. 42–51, 1948.

[136] R. Shannon, F. Zeng, V. Kamath, J. Wygonski, M. Ekelid *et al.*, "Speech recognition with primarily temporal cues," *Science*, pp. 303–303, 1995.

[137] H. Fletcher, "Speechand hearing in communication," 1953.

[138] G. Miller and J. Licklider, "The intelligibility of interrupted speech," *The Journal of the Acoustical Society of America*, vol. 22, no. 2, pp. 167–173, 1950.

[139] A. Sangwan, M. Chiranth, H. Jamadagni, R. Sah, R. Venkatesha Prasad, and V. Gaurav, "Vad techniques for real-time speech transmission on the internet," in *High Speed Networks and Multimedia Communications 5th IEEE International Conference on.* IEEE, 2002, pp. 46–50.

[140] K. Itoh and M. Mizushima, "Environmental noise reduction based on speech nonspeech identification for hearing aids," in *Acoustics, Speech, and Signal Pro-*

cessing, 1997. ICASSP-97., 1997 IEEE International Conference on, vol. 1. IEEE, 1997, pp. 419–422.

[141] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass, "Signal processing in high-end hearing aids: state of the art, challenges, and future trends," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 2915–2929, 2005.

[142] F. Basbug, K. Swaminathan, and S. Nandkumar, "Noise reduction and echo cancellation front-end for speech codecs," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 1, pp. 1–13, 2003.

[143] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 5, pp. 245–256, 2002.

[144] P. K. Gupta, S. Jangi, A. B. Lamkin, W. R. K. III, and A. J. Morris, "Voice activity detector for speech signals in variable background noise," US Patent No. 5649055.

[145] J. Ramirez, J. Górriz, and J. Segura, "Voice activity detection. fundamentals and speech recognition system robustness," *Robust Speech Recognition and Understanding*, pp. 1–22, 2007.

[146] C. Alberti, M. Bacchiani, A. Bezman, C. Chelba, A. Drofa, H. Liao, P. Moreno,

T. Power, A. Sahuguet, M. Shugrina *et al.*, "An audio indexing system for election video material," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on.* IEEE, 2009, pp. 4873–4876.

[147] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on.* IEEE, 2012, pp. 4117–4120.

[148] D. A. Eddins and D. M. Green, "Temporal integration and temporal resolution," *Hearing*, pp. 207–242, 1995.

[149] M. Florentine, H. Fastl, and S. Buus, "Temporal integration in normal hearing, cochlear impairment, and impairment simulated by masking," *The Journal of the Acoustical Society of America*, vol. 84, p. 195, 1988.

[150] W. O. Olsen and R. Carhart, "Integration of acoustic power at threshold by normal hearers," *The Journal of the Acoustical Society of America*, vol. 40, no. 3, pp. 591–599, 1966.

[151] J. Hughes, "The threshold of audition for short periods of stimulation," *Proceedings of the Royal Society of London. Series B-Biological Sciences*, vol. 133, no. 873, pp. 486–490, 1946.

[152] C. S. Watson and R. W. Gengel, "Signal duration and signal frequency in rela-

tion to auditory sensitivity," *The Journal of the Acoustical Society of America*, vol. 46, no. 4B, pp. 989–997, 1969.

[153] R. Plomp, "Rate of decay of auditory sensation," *The Journal of the Acoustical Society of America*, vol. 36, no. 2, pp. 277–282, 1964.

[154] M. J. Shailer and B. C. Moore, "Gap detection and the auditory filter: Phase effects using sinusoidal stimuli," *The Journal of the Acoustical Society of America*, vol. 81, p. 1110, 1987.

[155] P. J. Fitzgibbons and F. L. Wightman, "Gap detection in normal and hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 72, p. 761, 1982.

[156] J. J. Zwislocki, "Temporal summation of loudness: An analysis," *The Journal of the Acoustical Society of America*, vol. 46, no. 2B, pp. 431–441, 1969.

[157] I. J. Hirsh and C. E. Sherrick Jr, "Perceived order in different sense modalities." *Journal of experimental psychology*, vol. 62, no. 5, p. 423, 1961.

[158] S. Greenberg, J. Hollenback, and D. Ellis, "Insights into spoken language gleaned from phonetic transcription of the switchboard corpus," in *International Conference on Spoken Language Processing.* Citeseer, 1996, pp. S32–35.

[159] S. Greenberg, "Understanding speech understanding: Towards a unified theory of speech perception," in *Proceedings of the ESCA Tutorial and Advanced Re-*

search Workshop on the Auditory Basis of Speech Perception. Keele, England, 1996, pp. 1–8.

[160] ——, “On the origins of speech intelligibility in the real world,” in *Robust Speech Recognition for Unknown Communication Channels*, 1997.

[161] ——, “Speaking in shorthand–a syllable-centric perspective for understanding pronunciation variation,” *Speech Communication*, vol. 29, no. 2, pp. 159–176, 1999.

[162] S. Greenberg and T. Arai, “The relation between speech intelligibility and the complex modulation spectrum,” in *7th International Conference on Speech Communication and Technology*, 2001.

[163] S. Greenberg, “Pronunciation variation is key to understanding spoken language,” in *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, 2003, pp. 219–222.

[164] S. Greenberg, W. A. Ainsworth, and R. R. Fay, *Speech Processing in the Auditory System.* Springer, 2004.

[165] R. Drullman, J. Festen, and R. Plomp, “Effect of temporal envelope smearing on speech reception,” *The Journal of the Acoustical Society of America*, vol. 95, p. 1053, 1994.

[166] D. Van Tasell, S. Soli, V. Kirby, and G. Widin, “Speech waveform envelope cues

for consonant recognition," *The Journal of the Acoustical Society of America*, vol. 82, p. 1152, 1987.

[167] R. Drullman, *Listening to speech: an auditory perspective.* Lawrence Erlbaum, 2006, ch. The significance of temporal modulation for speech intelligibility.

[168] P. Price and H. Simon, "Perception of temporal differences in speech by "normal-hearing"adults: Effects of age and intensity," *The Journal of the Acoustical Society of America*, vol. 76, p. 405, 1984.

[169] S. Rosen, J. Walliker, J. Brimacombe, and B. Edgerton, "Prosodic and segmental aspects of speech perception with the house/3m single-channel implant," *Journal of Speech, Language and Hearing Research*, vol. 32, no. 1, p. 93, 1989.

[170] A. W. Huggins, "Temporally segmented speech," *Attention, Perception, & Psychophysics*, vol. 18, no. 2, pp. 149–157, 1975.

[171] H. Bourlard and S. Dupont, "A mew asr approach based on independent processing and recombination of partial frequency bands," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 1. IEEE, 1996, pp. 426–429.

[172] R. Drullman, J. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, p. 2670, 1994.

[173] N. F. Viemeister, "Temporal modulation transfer functions based upon modulation thresholds," *The Journal of the Acoustical Society of America*, vol. 66, p. 1364, 1979.

[174] T. Forrest and D. M. Green, "Detection of partially filled gaps in noise and the temporal modulation transfer function," *The Journal of the Acoustical Society of America*, vol. 82, p. 1933, 1987.

[175] D. J. Nelson, D. C. Smith, and J. L. Townsend, "Voice activity detector," US Patent No. 6556967, April 2003.

[176] J. Pencak and D. Nelson, "The NP speech activity detection algorithm," in *ICASSP*, vol. 1, 1995, pp. 381–384.

[177] J. D. Hoyt and H. Wechsler, "Detection of human speech in structured noise," in *ICASSP*, vol. 2, April 1994, pp. 237–240.

[178] J. Stegmann and G. Schroder, "Robust voice-activity detection based on the wavelet transform," in *IEEE Workshop on Speech Coding for Telecommunications*, 1997.

[179] T. Crystal, A. Schmidt-Nielsen, and E. Marsh, "Speech in noisy environments (SPINE) adds new dimension to speech recognition R&D," in *Proceedings of the second international conference on Human Language Technology Research.* Morgan Kaufmann Publishers Inc., 2002, pp. 212–216.

[180] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, K. Vesely, P. Matejka, X. Zhu, and N. Mesgarani, "Developing a speech activity detection system for the DARPA RATS program," *Interspeech*, 2012.

[181] P. Matejka, O. Plchot, M. Soufifar, O. Glembek, L. D'Haro, K. Veselỳ, F. Grézl, J. Ma, S. Matsoukas, and N. Dehak, "Patrol team language identification system for DARPA RATS P1 evaluation," in *Interspeech*, 2012.

[182] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, Jan 1999.

[183] Y. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *Signal Processing Letters, IEEE*, vol. 8, no. 10, pp. 276–278, 2001.

[184] S. Gazor and W. Zhang, "Speech probability distribution," *Signal Processing Letters, IEEE*, vol. 10, no. 7, pp. 204–207, 2003.

[185] P. Bell, M. Gales, P. Lanchantin, X. Liu, Y. Long, S. Renals, P. Swietojanski, and P. Woodland, "Transcription of multi-genre media archives using out-of-domain data," in *Proc. IEEE Workshop on Spoken Language Technology, Miama, Florida, USA*, 2012.

[186] Y. Raimond, C. Lowis, R. Hodgson, and J. Tweed, "Automated semantic tag-

ging of speech audio," in *Proceedings of the 21st international conference companion on World Wide Web.* ACM, 2012, pp. 405–408.

[187] M. Goto, J. Ogata, K. Yoshii, H. Fujihara, M. Mauch, and T. Nakano, "Podcastle and songle: Crowdsourcing-based web services for retrieval and browsing of speech and music content," *Proc. of CrowdSearch 2012*, 2012.

[188] K. Thadani, F. Biadsy, and D. Bikel, "On-the-fly topic adaptation for youtube video transcription," *Proceedings of Interspeech*, 2012.

[189] R. van Dalen, J. Yang, M. Gales, A. Ragni, and S. Zhang, "Generative kernels and score-spaces for classification of speech: Progress report," Tech. Rep. cued/f-infeng/tr. 676, Cambridge University Engineering Department, Tech. Rep., 2012.

[190] L. Nguyen, S. Abdou, M. Afify, J. Makhoul, S. Matsoukas, R. Schwartz, B. Xiang, L. Lamel, J. Gauvain, G. Adda *et al.*, "The 2004 bbn/limsi 10xrt english broadcast news transcription system," in *2004 Rich Transcriptions Workshop, Pallisades, NY*, 2004.

[191] A. Canavan, D. Graff, and G. Zipperlen, "CALLHOME American English Speech," Linguistic Data Consortium, Philadelpha, 1997.

[192] D. Graff, D. Miller, and K. Walker, "Switchboard-2 phase iii audio," Linguistic Data Consortium, Philadelpha, 2002.

[193] D. Graff, K. Walker, and D. Miller, "Switchboard cellular part 1 audio," Linguistic Data Consortium, Philadelpha, 2001.

[194] D. C. Smith, J. Townsend, D. J. Nelson, and D. Richman, "A multivariate speech actitivity detector based on the syllable rate," in *ICASSP*, 1999.

[195] N. Borges and G. G. L. Meyer, "Unsupervised distributional anomaly detection for a self-diagnostic speech activity detector," in *Conference on Information Sciences and Systems*, 2008.

[196] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[197] P. A. Lachenbruch and M. Goldstein, "Discriminant analysis," *Biometrics*, vol. 35, no. 1, pp. 69–85, 1979.

[198] K. P. Li and J. E. Porter, "Normalizations and selection of speech segments for speaker recognition scoring," in *ICASSP*, 1988.

[199] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[200] J. R. Hershey and P. A. Olsen, "Approximating the Kullback Leibler divergence between Gaussian mixture models," in *ICASSP*, April 2007.

[201] J. Goldberger, S. Gordon, and H. Greenspan, “An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures,” in *ICCV*, vol. 1, 2003, pp. 487–493.

[202] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, 1980.

[203] J. A. Swets, “The relative operating characteristic in psychology,” *Science*, vol. 182, no. 4116, pp. 990–1000, 1973.

[204] D. George and J. Hawkins, “A hierarchical Bayesian model of invariant pattern recognition in the visual cortex,” in *IJCAI*, 2005.

[205] N. Borges and G. G. L. Meyer, “Coping with training contamination in unsupervised distributional anomaly detection,” in *Conference on Information Sciences and Systems*, 2009.

[206] D. M. Rocke and D. L. Woodruff, “Identification of outliers in multivariate data,” *Journal of the American Statistical Association*, vol. 91, no. 435, 1996.

[207] W. Leow and R. Li, “The analysis and applications of adaptive-binning color histograms,” *CVIU*, vol. 94, no. 1-3, pp. 67–91, 2004.

[208] Y. Rubner, J. Puzicha, C. Tomasi, and J. Buhmann, “Empirical evaluation of

dissimilarity measures for color and texture," *CVIU*, vol. 84, no. 1, pp. 25–43, 2001.

[209] C. Shannon, "A mathematical theory of communications," *Bell System Technical Journal*, vol. 27, no. 2, pp. 632–656, 1948.

[210] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[211] M. Swain and D. Ballard, "Color indexing," *IJCV*, vol. 7, no. 1, pp. 11–32, 1991.

[212] H. Voorhees and T. Poggio, "Computing texture boundaries from images," *Nature*, vol. 333, no. 6171, pp. 364–367, 1988.

[213] J. Puzicha, T. Hofmann, and J. Buhmann, "Non-parametric similarity measures for unsupervised texture segmentation and image retrieval," in *CVPR*, 1997.

[214] D. Geman, S. Geman, C. Graffigne, and P. Dong, "Boundary detection by constrained optimization," *PAMI*, vol. 12, no. 7, pp. 609–628, 1990.

[215] B. Efron, "Bootstrap methods: Another look at the jackknife," *Annals of Mathematical Statistics*, vol. 7, no. 1, pp. 1–26, 1979.

[216] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Eurospeech*, 1997.

[217] N. Borges and G. G. L. Meyer, "Trimmed KL divergence between Gaussian mixtures for robust unsupervised acoustic anomaly detection," in *Interspeech*, 2009.

[218] R. A. Maronna, D. R. Martin, and V. J. Yohai, *Robust Statistics: Theory and Methods*, ser. Probability and Statistics. John Wiley and Sons, Inc., 2006.

[219] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, pp. 461–464, 1978.

[220] G. L. Alexanderson, "Euler and Kőnigsberg's bridges: A historical view," *American Mathematical Society*, vol. 43, no. 4, pp. 567–573, October 2006.

[221] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. pp. 167–256, 2003.

[222] R. M. Durbin, "Studies on the development and organisation of the nervous system of C. elegans," Ph.D. dissertation, Cambridge University, 1987.

[223] S. Wasserman and K. Faust, *Social network analysis : methods and applications.* Cambridge University Press, 1998.

[224] S. McCrystal, "It takes a network: the new front lines of modern warfare," *Foreign Policy*, March/April 2011.

[225] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *Phys. Rev. E*, vol. 64, Jul 2001.

[226] L. Goldwasser and J. Roughgarden, “Construction and analysis of a large caribbean food web,” *Ecology*, vol. 74, no. 4, pp. pp. 1216–1233, 1993.

[227] M. Faloutsos, P. Faloutsos, and C. Faloutsos, “On power-law relationships of the internet topology,” in *SIGCOMM*, 1999, pp. 251–262.

[228] D. J. Watts and S. Strogatz, “Collective dynamics of ‘small-world’ networks,” *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[229] C. Koch and G. Laurent, “Complexity and the nervous system,” *Science*, vol. 284, no. 5411, pp. 96–98, 1999.

[230] J. Grothendieck, A. Gorin, and N. Borges, “Social correlates of turn-taking behavior,” in *ICASSP*, 2009.

[231] A. Scala, L. A. N. Amaral, and M. Barthélémy, “Small-world networks and the conformation space of a short lattice polymer chain,” *Europhysics Letters*, vol. 55, no. 4, p. 594, 2001.

[232] H. Jeong, B. Tombar, R. Albert, Z. N. Oltyai, and A.-L. Barabási, “The large-scale organization of metabolic networks,” *Nature*, vol. 407, no. 6804, pp. 651–654, 2000.

[233] L. Adamic, “The small world web,” in *Research and Advanced Technology for Digital Libraries*, ser. Lecture Notes in Computer Science. Springer, 1999, vol. 1696, pp. 852–852.

[234] J. Abello, M. Resende, and S. Sudarsky, "Massive quasi-clique detection," in *Theoretical Informatics.* Springer, 2002, vol. 2286, pp. 598–612.

[235] J. Goldenberg, B. Libai, S. Solomon, N. Jan, and D. Stauffer, "Marketing percolation," *Physica A: Statistical Mechanics and its Applications*, vol. 284, no. 1-4, pp. 335 – 347, 2000.

[236] J. Coleman, E. Katz, and H. Menzel, "The diffusion of an innovation among physicians," *Sociometry*, vol. 20, no. 4, pp. pp. 253–270, 1957.

[237] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Phys. Rev. Lett.*, vol. 86, Apr 2001.

[238] X.-J. Xu, X. Zhang, and J. F. F. Mendes, "Impacts of preference and geography on epidemic spreading," *Phys. Rev. E*, vol. 76, Nov 2007.

[239] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," *Physics Reports*, vol. 424, no. 4-5, pp. 175 – 308, 2006.

[240] M. A. Porter, J.-P. Onnela, and P. J. Mucha, "Communities in Networks," *Notices of the American Mathematical Society, Vol. 56, No. 9, 2009*, 2009.

[241] B. P. Olding and P. J. Wolfe, "Inference for graphs and networks: Extending classical tools to modern data," Jun. 2009, submitted for publication.

[242] D. J. Watts, *Six Degrees: The Science of a Connected Age.* W. W. Norton and Company, Feb. 2003.

[243] A.-L. Barabasi, *Linked: How Everything Is Connected to Everything Else and What It Means.* Plume, Apr. 2003.

[244] E. D. Kolaczyk, *Statistical Analysis of Network Data: Methods and Models.* Springer, Mar. 2009.

[245] R. Larson, *Elementary linear algebra.* Cengage Learning, 2012.

[246] S. H. Unger, "GIT-a heuristic program for testing pairs of directed line graphs for isomorphism," *Commun. ACM*, vol. 7, pp. 26–34, January 1964.

[247] D. Conte, P. Foggia, C. Sansone, and M. Vento, "Thirty years of graph matching in pattern recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, 2004.

[248] B. Jenner, J. Köbler, P. McKenzie, and J. Torán, "Completeness results for graph isomorphism," *Journal of Computer and System Sciences*, vol. 66, no. 3, pp. 549–566, 2003.

[249] E. H. Sussenguth, "A graph-theoretic algorithm for matching chemical structures," *Journal of Chemical Documentation*, vol. 5, no. 1, pp. 36–43, 1963.

[250] M. Zaslavskiy, F. Bach, and J. Vert, "Global alignment of protein–protein inter-

action networks by graph matching methods," *Bioinformatics*, vol. 25, no. 12, pp. 1259–1267, 2009.

[251] J. W. Raymond and P. Willett, "Maximum common subgraph isomorphism algorithms for the matching of chemical structures," *Journal of Computer-Aided Molecular Design*, vol. 16, pp. 521–533, 2002, 10.1023/A:1021271615909.

[252] E.K. and Wong, "Model matching in robot vision by subgraph isomorphism," *Pattern Recognition*, vol. 25, no. 3, pp. 287 – 303, 1992.

[253] H.-Y. Kim and J. H. Kim, "Hierarchical random graph representation of handwritten characters and its application to hangul recognition," *Pattern Recognition*, vol. 34, no. 2, pp. 187 – 201, 2001.

[254] J. Rocha and T. Pavlidis, "A shape analysis model with applications to a character recognition system," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 16, no. 4, pp. 393 –404, apr 1994.

[255] V. Cantoni, L. Cinque, C. Guerra, S. Levialdi, and L. Lombardi, "2D object recognition by multiscale tree matching," *Pattern Recognition*, vol. 31, no. 10, pp. 1443 – 1454, 1998.

[256] T. Lourens, "A biologically plausible model for corner-based object recognition from color images," Ph.D. dissertation, University of Groningen, 1998.

[257] M. Pelillo, K. Siddiqi, and S. Zucker, "Matching hierarchical structures using

association graphs," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 21, no. 11, pp. 1105 –1120, nov 1999.

[258] P. Showbridge, M. Kraetzl, and D. Ray, "Detection of abnormal change in dynamic networks," in *Information, Decision and Control, 1999. IDC 99. Proceedings. 1999*, 1999, pp. 557 –562.

[259] A. Pande, M. Gupta, and A. Tripathi, "Design pattern mining for gis application using graph matching techniques," in *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on*, vol. 3, July 2010, pp. 477 –482.

[260] D. H. Fisher, "Knowledge acquisition via incremental conceptual clustering," *Machine Learning*, vol. 2, pp. 139–172, 1987, 10.1007/BF00114265.

[261] J. R. Ullmann, "An algorithm for subgraph isomorphism," *J. ACM*, vol. 23, pp. 31–42, January 1976.

[262] J. McGregor, "Backtrack search algorithms and the maximal common subgraph problem," *Software: Practice and Experience*, vol. 12, no. 1, pp. 23–34, 1982.

[263] L. Torresani, V. Kolmogorov, and C. Rother, "Feature correspondence via graph matching: Models and global optimization," *Computer Vision–ECCV 2008*, pp. 596–609, 2008.

[264] S. Ndiaye and C. Solnon, “Cp models for maximum common subgraph problems,” *Principles and Practice of Constraint Programming–CP 2011*, pp. 637–644, 2011.

[265] B. Park, K. Lee, and S. Lee, “A novel stochastic attributed relational graph matching based on relation vector space analysis,” in *Advanced Concepts for Intelligent Vision Systems*. Springer, 2006, pp. 978–989.

[266] W.-H. Tsai and K.-S. Fu, “Error-correcting isomorphisms of attributed relational graphs for pattern analysis,” *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 9, no. 12, pp. 757 –768, December 1979.

[267] B. McKay, “Practical graph isomorphism,” *Congressus Numerantium*, vol. 30, pp. 45–87, 1981. [Online]. Available: http://cs.anu.edu.au/~bdm/nauty/PGI

[268] B. D. McKay. nauty user’s guide. [Online]. Available: http://cs.anu.edu.au/~bdm/nauty/nug.pdf

[269] T. Miyazaki, “The complexity of McKay’s canonical labeling algorithm,” *Groups and Computation II*, 1996.

[270] B. Messmer and H. Bunke, “Efficient subgraph isomorphism detection: a decomposition approach,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 12, no. 2, pp. 307 –323, 2000.

[271] K. Shearer, H. Bunke, and S. Venkatesh, “Video indexing and similarity retrieval by largest common subgraph detection using decision trees,” *Pattern Recognition*, vol. 34, no. 5, pp. 1075 – 1091, 2001.

[272] B. Messmer and H. Bunke, “A decision tree approach to graph and subgraph isomorphism detection,” *Pattern Recognition*, vol. 32, no. 12, pp. 1979 – 1998, 1999.

[273] I. Koch, “Enumerating all connected maximal common subgraphs in two graphs,” *Theoretical Computer Science*, vol. 250, no. 1, pp. 1–30, 2001.

[274] D. Eppstein, “Subgraph isomorphism in planar graphs and related problems,” in *Proceedings of the sixth annual ACM-SIAM symposium on discrete algorithms.* Society for Industrial and Applied Mathematics, 1995, pp. 632–640.

[275] H. Bunke, “On a relation between graph edit distance and maximum common subgraph,” *Pattern Recognition Letters*, vol. 18, no. 8, pp. 689 – 694, 1997.

[276] H. Bunke and K. Shearer, “A graph distance metric based on the maximal common subgraph,” *Pattern Recognition Letters*, vol. 19, no. 3-4, pp. 255 – 259, 1998.

[277] W. Tsai and K. Fu, “Subgraph error-correcting isomorphisms for syntactic pattern recognition,” *Systems, Man and Cybernetics, IEEE Transactions on*, no. 1, pp. 48–62, 1983.

[278] A. Sanfeliu and K. Fu, "A distance measure between attributed relational graphs for pattern recognition," *Systems, Man and Cybernetics, IEEE Transactions on*, no. 3, pp. 353–362, 1983.

[279] A. Dumay, R. van der Geest, J. Gerbrands, E. Jansen, and J. Reiber, "Consistent inexact graph matching applied to labelling coronary segments in arteriograms," in *Pattern Recognition, 1992. Vol.III. Conference C: Image, Speech and Signal Analysis, Proceedings., 11th IAPR International Conference on*, 1992, pp. 439–442.

[280] A. Rosenfeld, R. Hummel, and S. Zucker, "Scene labeling by relaxation operations," *Systems, Man and Cybernetics, IEEE Transactions on*, no. 6, pp. 420–433, 1976.

[281] L. Davis, "Shape matching using relaxation techniques," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 1, pp. 60–72, 1979.

[282] S. Ullman, "Relaxation and constrained optimization by local processes," *Computer Graphics and Image Processing*, vol. 10, no. 2, pp. 115–125, 1979.

[283] W. Christmas, J. Kittler, and M. Petrou, "Structural matching in computer vision using probabilistic relaxation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 17, no. 8, pp. 749 –764, August 1995.

[284] B. Huet and E. Hancock, "Shape recognition from large image libraries by

inexact graph matching," *Pattern Recognition Letters*, vol. 20, no. 11, pp. 1259–1269, 1999.

[285] H. Almohamad and S. Duffuaa, "A linear programming approach for the weighted graph matching problem," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 15, no. 5, pp. 522 –525, may 1993.

[286] B. Luo and E. Hancock, "Structural graph matching using the em algorithm and singular value decomposition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 10, pp. 1120–1136, 2001.

[287] S. Medasani, R. Krishnapuram, and Y. Choi, "Graph matching by relaxation of fuzzy assignments," *Fuzzy Systems, IEEE Transactions on*, vol. 9, no. 1, pp. 173–182, 2001.

[288] M. Van Wyk, T. Durrani, and B. Van Wyk, "A rkhs interpolator-based graph matching algorithm," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 988–995, 2002.

[289] S. Kosinov and T. Caelli, "Inexact multisubgraph matching using graph eigenspace and clustering models," *Structural, Syntactic, and Statistical Pattern Recognition*, pp. 455–473, 2002.

[290] S. Umeyama, "An eigendecomposition approach to weighted graph matching

problems," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 10, no. 5, pp. 695 –703, sep 1988.

[291] L. Xu and I. King, "A pca approach for fast retrieval of structural patterns in attributed graphs," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 31, no. 5, pp. 812 –817, Oct 2001.

[292] M. Carcassoni and E. Hancock, "Weighted graph-matching using modal clusters," in *Computer Analysis of Images and Patterns.* Springer, 2001, pp. 142–151.

[293] D. G. Corneil and D. G. Kirkpatrick, "A theoretical analysis of various heuristics for the graph isomorphism problem," *SIAM Journal on Computing*, vol. 9, no. 2, pp. 281–297, 1980.

[294] H. Pao, G. A. Coppersmith, and C. E. Priebe, "Statistical inference on random graphs: Comparative power analyses via monte carlo," *Journal of Computational and Graphical Statistics*, vol. 20, no. 2, pp. 395–416, 2011.

[295] S. Fortin, "The graph isomorphism problem," University of Alberta, Tech. Rep., 1996.

[296] L. C. Freeman, "Centrality in social networks: conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1979.

[297] J. Scott, "Social network analysis: developments, advances, and prospects," *Social Network Analysis and Mining*, vol. 1, pp. 21–26, 2011.

[298] V. E. Krebs, "Mapping networks of terrorist cells," *Connections*, vol. 24, no. 3, pp. 43–52, 2002.

[299] P. Bonacich, "Factoring and weighting approaches to status scores and clique identification," *Journal of Mathematical Sociology*, vol. 2, no. 1, pp. 113–120, 1972.

[300] C. Priebe, J. Conroy, D. Marchette, and Y. Park, "Scan statistics on enron graphs," *Computational and Mathematical Organization Theory*, vol. 11, pp. 229–247, 2005.

[301] N. Borges, G. Coppersmith, G. Meyer, and C. Priebe, "Anomaly detection for random graphs using distributions of vertex invariants," in *Conference on Information Sciences and Systems*, 2011.

[302] L. Von Collatz and U. Sinogowitz, "Spektren endlicher grafen," in *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg*, vol. 21, no. 1. Springer, 1957, pp. 63–77.

[303] A. Mowshowitz, "The characteristic polynomial of a graph," *Journal of Combinatorial Theory, Series B*, vol. 12, no. 2, pp. 177–193, 1972.

[304] A. Schwenk, “Almost all trees are cospectral,” *New directions in the theory of graphs*, pp. 275–307, 1973.

[305] R. Gelbart, “Use of the spectrum in graph isomorphism,” Master’s thesis, The University of British Columbia, 1976.

[306] P. Erdős and A. Rényi, “On random graphs,” *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.

[307] B. Bollobás, *Random Graphs (Cambridge Studies in Advanced Mathematics)*, 2nd ed. Cambridge University Press, 2001.

[308] R. van der Hofstad, “Random graphs and complex networks,” November 2011, Lecture Notes.

[309] H. Westergaard, *Contributions to the History of Statistics.* P. S. King, 1932.

[310] D. W. Scott, “On optimal and data-based histograms,” *Biometrika*, vol. 66, no. 3, pp. 605–610, 1979.

[311] R. Bellman, *Adaptive control processes.* Princeton University Press, 1961, vol. 4.

[312] M. P. Wand, “Data-based choice of histogram bin width,” *The American Statistician*, vol. 51, no. 1, pp. pp. 59–64, 1997.

[313] H. A. Sturges, “The choice of a class interval,” *Journal of the American Statistical Association*, vol. 21, no. 153, pp. 65–66, 1926.

[314] C. Moler, “Matrix computation on distributed memory multiprocessors,” *Hypercube Multiprocessors*, vol. 86, pp. 181–195, 1986.

[315] K. Rayner, “Eye movements in reading and information processing: 20 years of research.” *Psychological bulletin*, vol. 124, no. 3, p. 372, 1998.

[316] S. Martinez-Conde, S. L. Macknik, and D. H. Hubel, “The role of fixational eye movements in visual perception,” *Nature Reviews Neuroscience*, vol. 5, no. 3, pp. 229–240, 2004.

[317] S. Ganapathy, “Signal analysis using autoregressive models of amplitude modulation,” Ph.D. dissertation, Johns Hopkins University, 2012.

[318] S. Ganapathy, S. Thomas, and H. Hermansky, “Comparison of modulation features for phoneme recognition,” in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on.* IEEE, 2010, pp. 5038–5041.

[319] S. Ganapathy, P. Motlicek, and H. Hermansky, “Autoregressive models of amplitude modulations in audio compression,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1624–1631, 2010.

[320] S. Ganapathy, J. Pelecanos, and M. K. Omar, “Feature normalization for speaker verification in room reverberation,” in *Acoustics, Speech and Signal*

Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011, pp. 4836–4839.

[321] M. Cooke, *Modelling auditory processing and organisation.* Cambridge University Press, 1993, vol. 7.

[322] P. Green, M. Cooke, and M. Crawford, "Auditory scene analysis and hidden markov model recognition of speech in noise," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1. IEEE, 1995, pp. 401–404.

[323] M. Cooke, A. Morris, and P. Green, "Missing data techniques for robust speech recognition," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 2. IEEE, 1997, pp. 863–866.

[324] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech communication*, vol. 34, no. 3, pp. 267–285, 2001.

[325] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 1. IEEE, 1998, pp. 121–124.

[326] J. Jung, K. Kim, and M. Kim, "Advanced missing feature theory with fast score

calculation for noise robust speaker identification," *Electronics letters*, vol. 46, no. 14, pp. 1027–1029, 2010.

[327] B. Raj and R. Stern, "Missing-feature approaches in speech recognition," *Signal Processing Magazine, IEEE*, vol. 22, no. 5, pp. 101–116, 2005.

[328] J. Ming, T. Hazen, and J. Glass, "Combining missing-feature theory, speech enhancement, and speaker-dependent/-independent modeling for speech separation," *Computer Speech & Language*, vol. 24, no. 1, pp. 67–76, 2010.

[329] M. Van Segbroeck and H. Van Hamme, "Advances in missing feature techniques for robust large-vocabulary continuous speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 123–137, 2011.

[330] Q. Wang, S. R. Kulkarni, and S. Verdu, "Divergence estimation for multidimensional densities via k-nearest-neighbor distances." *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2392–2405, 2009.

[331] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the thirtieth annual ACM symposium on Theory of computing.* ACM, 1998, pp. 604–613.

[332] A. Gionis, P. Indyk, R. Motwani *et al.*, "Similarity search in high dimensions via hashing," in *VLDB*, vol. 99, 1999, pp. 518–529.

Vita

Nash M. Borges earned the Bachelor of Science in Electrical Engineering degree in 2002 from Tufts University, graduating *summa cum laude* with a second major in Mathematics. He earned a Master of Science degree in Electrical and Computer Engineering from Johns Hopkins University in 2004.

Nash is currently employed as a Senior Research Scientist with the Department of Defense. His research interests are in data mining, machine learning, and the fusion of speech and natural language processing with geo-spatial, behavioral, and social network analyses.