Detection of Nonstationary Noise and Improved Voice Activity Detection in an Automotive Hands-free Environment

by

Stephen William Laverty

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Master of Science

in

Electrical Engineering

May 2005

APPROVED:

Dr. Donald R. Brown, Major Advisor

Dr. Brian King, Committee Member

Dr. Fred J. Looft, Department Head

Dr. John McNeill, Committee Member

Abstract

Speech processing in the automotive environment is a challenging problem due to the presence of powerful and unpredictable nonstationary noise. This thesis addresses two detection problems involving both nonstationary noise signals and nonstationary desired signals. Two detectors are developed: one to detect passing vehicle noise in the presence of speech and one to detect speech in the presence of passing vehicle noise. The latter is then measured against a state-of-the-art voice activity detector used in telephony. The process of compiling a library of recordings in the automobile to facilitate this research is also detailed.

ACKNOWLEDGEMENTS

I owe a great deal of thanks to very many people for making my graduate work possible and enjoyable. I'd like to start by thanking my parents for their unwavering support and generous encouragement throughout my academic career and entire life. I'm certain I could not have gotten this far without it.

Next I'd like to thank Rick Brown for his commitment, encouragement, assistance and company over an intense three years of electrical and computer engineering. I owe two of those three years to him convincing me that I should go to graduate school and, despite a few rough spots, I'm very glad he did. I don't think it's possible for me to quantify everything I've learned from him over that time or my appreciation for it.

Bose Corporation deserves my thanks for their generosity in supporting me for my first year of graduate school. I'd specifically like to thank Jeff Faneuff for making my relationship with Bose possible. He and Vasu Iyengar also contributed much valuable advice and feedback on this work.

I owe Jeremy Slater and David Feinzeig thanks for their support, their company, and their assistance with this project. They've made many late nights in the lab rather enjoyable and were anything but hesitant to aid in the data collection effort.

Last, but certainly not least, I'd like to thank my committee, Brian King and John McNeill, for their time and their valuable feedback.

Contents

Chapter 1

Introduction

The past decade has brought about a proliferation in both number and complexity of accessories in new automobiles as well as the pervasive use of mobile phones. All of these new accessories increase driver load. Driver load refers to the amount of attention required from the driver, inclusive of tasks other then driving. Elevated driver load plays an important role in safety [GB$^+$97] and the load of controlling accessories in the vehicle is attributed to upwards of one tenth of automobile accidents caused by distractions inside the cabin, exclusive of cell phones [WT95].

Driver load can be reduced by hands-free speech acquisition [FGW94]. Speech recognition provides a less load intensive way of interacting with these vehicle systems [FGW94] by eliminating the need to find and manipulate controls. It also provides an alternative method for performing dialing and other call control operations related to the cellular phone, which are responsible for some portion of the excluded cellular phone related accidents. Conversational audio can also be provided by the hands-free speech acquisition system, avoiding the distraction of holding a handset or mounting a headset while driving. The primary drawback of hands-free speech acquisition is its relatively low signal to noise ratio (SNR).

The safety benefits of the hands-free acquisition system are very desirable but performance must be maintained despite the reduced signal to noise ratio (SNR). Performance of the acquisition system is crucial to the acceptance of this technology, since recognition accuracy is strongly dependent on signal quality, and inaccurate recognition may lead users to revert to traditional visio-tactile methods for modifying equipment settings.

Given the increased safety provided by using a cabin-integrated speech acquisition system, it seems reasonable that drivers would want that system to work under as broad a set of conditions as possible. Hands-free speech systems can already achieve an acceptable quality of speech under specific conditions, namely having the windows up. The windows-down case introduces external, or environmental, noise into the cabin which was previously attenuated greatly by the windows. The environmental noise tends to have a strong non-stationary element, much of which may be passing vehicle noise. Therefore attenuation of passing vehicle noise from audio acquired in the automobile enhances its usability both for conversation and speech recognition under a substantially expanded range of conditions. Signal processing can be used to attenuate this unwanted noise.

Noise reduction research targeting hands-free speech acquisition in the automobile has been pursued for many years. Previous work has focused on a number of different aspects of this problem. Reducing the impact of specific sources of vehicle noise is often a consideration, such as in [WB95]. [OVP92] and [WG92] both address the totality of noise under different driving conditions. It is unclear, however, whether passing vehicle noise is included in their experiments. Other studies, such as [AK05], are concerned only with quality of speech for recognition purposes and do not consider the perceptual quality of speech for conversational purposes.

Noise reduction for hands-free speech acquisition systems in the automobile continues to be an area of active research. Recent work in [Fan02] focused on noise reduction in the automotive environment using a microphone array. Highly nonstationary noise, such as passing vehicle noise, was specifically noted as problematic and left unaddressed. A recent development in microphone array-based noise reduction in the automobile is presented in [Coh04]. This system is designed to deal with highly nonstationary noise in an automotive environment. Testing of this algorithm in [Bro04] revealed, however, that it was not effective in mitigating passing vehicle noise specifically.

This thesis focuses on the problem of passing vehicle noise. The detection of passing vehicle noise and the detection of speech in the presence of passing vehicle noise are addressed specifically. While detection does not provide mitigation of the noise directly, the detectors can be applied to mitigation schemes, such as [Coh04], to potentially improve mitigation performance.

1.1 Thesis Organization

This thesis is comprised of four major sections:

- Background Material
- Data Acquisition and Analysis
- Detection of Passing Vehicle Noise
- Detection of Speech in the Presence of Passing Vehicle Noise

Background material is provided in Chapter 2 and consists of two pieces. The first discusses assumptions and assertions about the problem by exploring the expected operating environment. The second addresses specific techniques from the literature that are applied in later chapters.

Chapter 3 relates information about the recordings used in this project. First, the system used to obtain the multichannel recordings of a variety of real-world passing vehicle events is discussed in detail. This includes what equipment was used, how it was configured, and how it was connected as well as a characterization and verification of the system. Second, the data gathered during these recording sessions is analyzed and a simple model useful for simulating passing vehicle noise is developed.

The problem of detecting passing vehicles is explored in Chapter 4. The solution is presented in two steps: (i) the selection and evaluation of features and (ii) the development of an optimal classifier. A number of features are investigated. Each is described in detail and sample feature data is generated from a test sample. This sample data is then analyzed in the context of detecting passing vehicle noise. After promising features are identified, the task of using them to classify whether passing vehicle noise is present is addressed. Because only one scalar feature is identified, only the straight-forward case of unidimensional feature classification is discussed. The chapter concludes with the description of a viable detector and some basic performance evaluations of that detector.

Chapter 5 parallels Chapter 4 in the exploration of features and the application of classification techniques, but addresses the problem of detecting speech when passing vehicle noise is present instead of detecting the passing vehicle noise. To avoid repetition the detailed description of the features explored in Chapter 4 is omitted but each feature is reevaluated as to its ability to differentiate whether speech is present during a passing vehicle event. Additional features more relevant to speech are also explored in this chapter in a similar manner and more complex multidimensional feature classification techniques are applied. The chapter concludes with the presentation of an improved detector and its performance is compared to both G.729 and the GSM VAD.

1.2 Thesis Contributions

The major contributions of this thesis are as follows:

- Chapter 3
  - An acquisition system for producing recordings inside the vehicle cabin is described in detail.
  - A characterization of passing vehicle noise is provided and a basic model for passing vehicle noise is specified.

- Chapter 4
  - Statistics relating a variety of features to detection of passing vehicle noise are given.
  - The results of applying thresholds to unidimensional features for classification are shown.
  - A viable detector for passing vehicle noise is detailed and its performance is evaluated.
- Chapter 5
  - Statistics relating a variety of features to the detection of speech during passing vehicle events are given.
  - The results of applying thresholds to unidimensional features for classification are shown.
  - The results of applying discriminant analysis to multidimensional features for classification are shown.
  - A viable detector of speech in the presence of passing vehicle noise is described. Performance is compared to two industry standard voice activity detectors.

Chapter 2

Background

This chapter highlights background material needed in Chapters 3, 4, and 5. This includes common signal processing and classification techniques used in later chapters as well as information about prior data collection efforts. To begin, a model of the operating environment is defined and the sources of sound are discussed to provide a better understanding of the signals present.

2.1 System Model

To understand how the passing vehicle noise detection and passing vehicle noise tolerant voice activity detection problems relate to methods of signal processing we need to have an idea of the composition of the signals to be processed and how they are structured. This begins with understanding the physical configuration of the acoustical space and the sources of sound in this space.

A basic model of the vehicle in open space is shown in Figure 2.1. In this model each microphone in the array is connected to both the source of speech and the sources of noise by a number of acoustical paths. Since these noise sources, the speaker, and the microphones are fixed relative to the vehicle, these acoustical paths are constant over time.

While Figure 2.1 takes into account many of the prevalent noise sources in the vehicle,

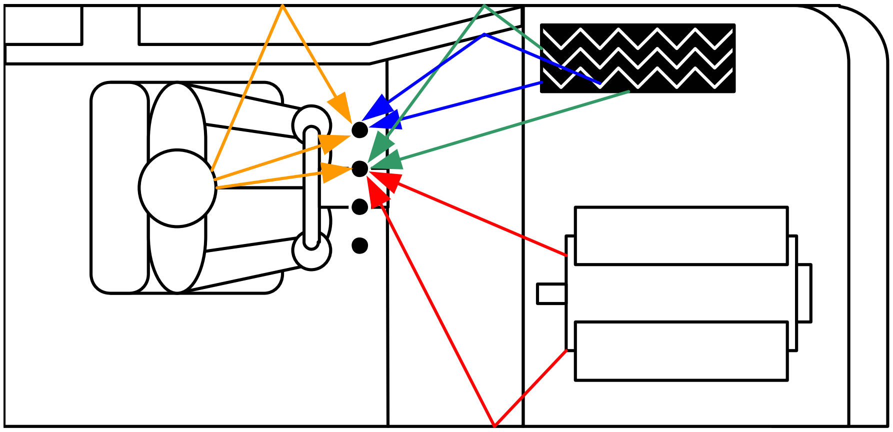

Figure 2.1: Basic physical model of acoustical paths coupling speech and noise sources to microphones in the array.

such as engine noise and road noise, it does not describe environmental noise such as passing vehicle noise. An illustration of acoustical paths in the case of passing vehicle noise is shown in Figure 2.2. It is clear that, as the location of the passing vehicle changes over time, the acoustical paths between noise sources in the passing vehicle and the microphones also change.

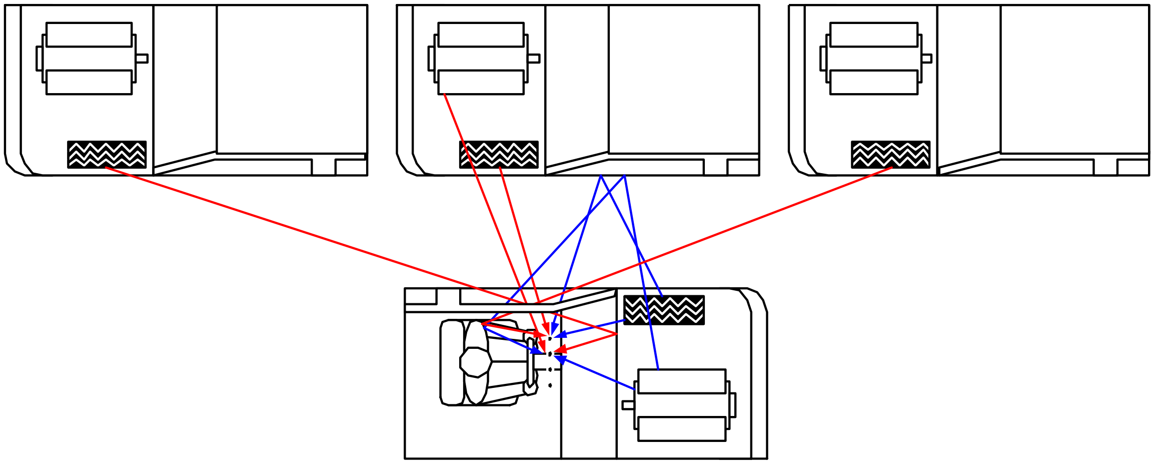

Figure 2.2: Sound propagation paths at different moments in time as a vehicle passes.

The physical model described in Figures 2.1 and 2.2 can be transformed into a conceptual model better suited to signal processing. Figure 2.3 illustrates a simple conceptual model that can be applied to the environment described above. In this model speech is transformed through M systems and received at M microphones and additive noise is present at each of these microphones as well. We note that this is the same model adopted in [Coh04].

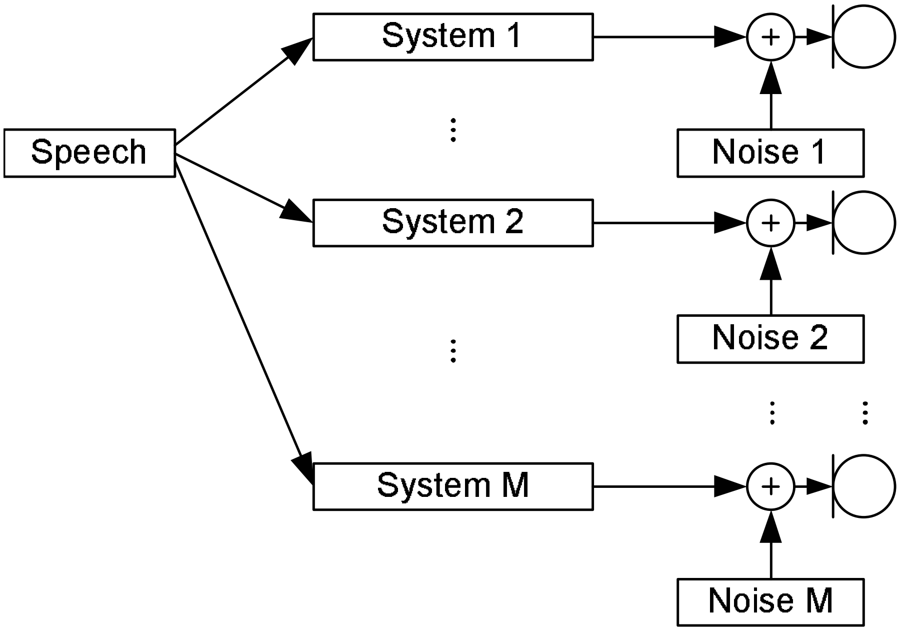

Figure 2.3: Conceptual model of speech and noise sources in the vehicle as described in [Coh04].

The speech portion of the conceptual model maps directly onto the physical model, where each of the M systems represents the net effect of all of the acoustical paths connecting the speech source to each microphone. While the physical model enumerates the sources of noise and considers the systems which couple each source to each microphone, the conceptual model considers only a single noise source added to the signal received at each microphone. This is the combined effect of all the noise sources, both stationary and nonstationary, at that microphone.

The nature of the signals present in the model can be examined to determine what signal processing techniques might be relevant. As in [Coh04], the desired signal, speech, is assumed to be nonstationary. The noise, however, which [Coh04] treats as a combination of stationary and nonstationary noise, deserves further examination. When no passing vehicle is present, as in Figure 2.1, the noise received at the microphones is approximately stationary, as is demonstrated later in Section 3.2. When a vehicle passes, however, as shown in Figure 2.2, the noise observed at the microphones is nonstationary, which is also demonstrated in Section 3.2. The collection of experimental data, described in the following section, allows the properties of these noises to be verified.

2.2 Data Collection

There have been two prior data collection efforts closely related to the effort described in Chapter 3. These data collection efforts were both performed using four microphones arranged as a broadside array attached to the driver's visor. The microphones used in both cases were identical.

The first effort, described in Appendix A, was performed by Punit Prakash. This effort involved separate recordings of passing noise and speech. These recordings suffered from several problems including aliasing, dropped samples, and mechanical microphone saturation. The second effort, described in [MPC04], was performed by a team of undergraduate students at WPI. This effort addressed the shortcomings of the previous effort but the team did not record any passing vehicle events.

Neither of these data collection efforts produced recordings appropriate for evaluation of the algorithms used or detectors developed in this thesis. The acquisition systems developed in this prior work did, however, serve as appropriate guides for acquiring new data. The data collection effort that produced the recordings used in this thesis is presented in Chapter 3. These recordings are then used to verify the assumptions regarding stationarity present in the conceptual model as well as for the evaluation of the detection schemes presented in Chapters 4 and 5.

2.3 Passing Vehicle Noise Detection

The first of two detection problems addressed in this thesis is the passing vehicle noise detection problem. This problem consists of detecting the presence of passing vehicle noise while background noise is present and speech may or may not be present. The problem can be approached using the general framework shown in Figure 2.4, where distinguishing features are first created based on the data and then those features are classified. This framework is drawn from classification theory [Jam85].

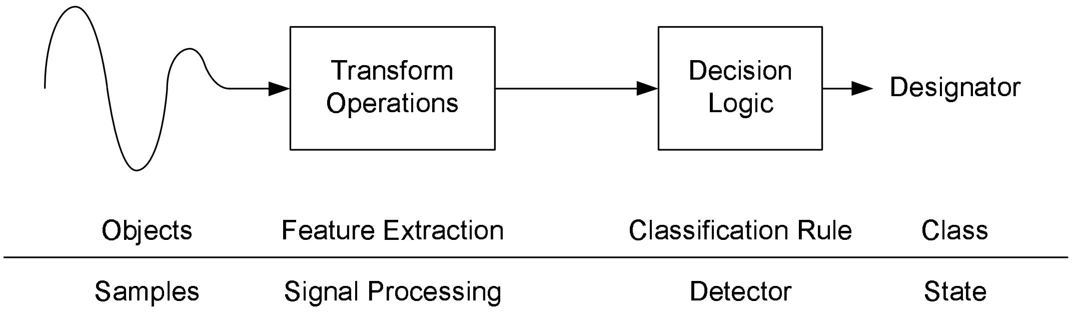

Figure 2.4: A classification model for detection.

2.3.1 Feature Extraction, Classification, and Detection

The detection problem in signal processing is similar to classification as seen in statistics. This is shown in Figure 2.4 and a good introduction is provided in [Jam85]. This classification model is made up of two key steps. The first, *feature extraction*, is analogous to signal processing or preprocessing steps. The second, *classification* via a classification rule, is analogous to a detector.

Features, in general, refer to the properties of an item. For physical objects in space, features might consist of color, mass, and volume. While slightly less apparent, one can consider features of a signal. Such features might be instantaneous power, or amplitude. Both of these features are functions of only one sample of the signal. These simple features may suffice in many situations, but often the salient characteristics of a signal exist in the correlation between samples of the signal. This leads to more complex features such as time-averaged power, the presence of certain frequencies, or estimated pitch parameters.

Classification of an item is generally performed based on one of more of the item's features. Returning to the physical analogy, objects might be classified by any or all of the previously enumerated features such as those that are or are not red, those that are greater than or less than ten kilograms or those that are greater than ten kilograms but not red. Applying this idea to the signal properties already listed, classes might consist of regions where the power is greater than 10dB or the estimated pitch is between 100Hz and 300Hz.

This two stage approach to detection problems is applied for both the passing vehicle noise detection problem and the passing vehicle noise tolerant voice activity detector problem. To address the first stage, feature generation, for the passing vehicle noise detection problem two signal processing techniques from the literature are used in the creation of distinguishing features. Both are single-channel, as opposed to multichannel, techniques as they prove powerful enough in the case of passing vehicle noise. The first is the recursive least squares linear predictor (RLS). The second is the short-time Fourier transform (STFT).

2.3.2 Feature Extraction Using Recursive Least Squares

The first technique from the literature used to generate a distinguishing feature is the recursive least squares linear predictor (RLS). RLS is applied to separate the predictable component of the microphone signal from the unpredictable component, with the objective of enhancing just the transient/nonstationary components of the signal, namely speech and passing vehicle noise.

The recursive least squares linear predictor predicts the next sample of an input sequence based on previous samples. The prediction is simply a linear combination of previous samples. This weighted moving average is a FIR filter as shown in Figure 2.5.

While the basic structure is that of an FIR filter, the coefficients or weights are contin-

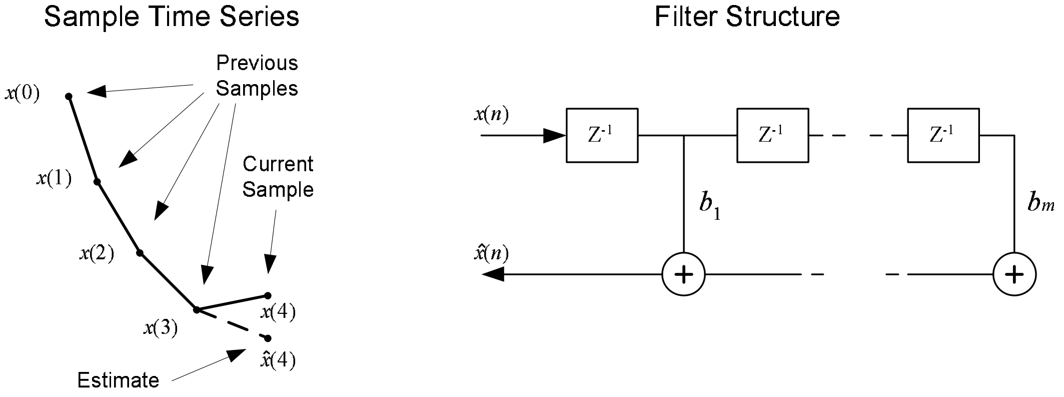

Figure 2.5: Filter structure of RLS and a sample time sequence.

uously changing. The recursive portion of the algorithm is concerned with recalculating the optimum tap weights given each new observation. Optimum is defined as minimizing the cost function [Vas00]

$$J = \sum_{i=0}^{n-1} [\lambda^{n-i} \varepsilon^2(i)]$$

where ϵ is the difference between the *the prediction* $\hat{x}(i)$ and the actual sample value $x(i)$, i.e.,

$$\varepsilon(i) = x(i) - \hat{x}(i) = x(i) - \sum_{m=1}^{M} b_m x(i-m)$$

and n is the current sample number. This cost is computed over all previous samples. The λ provides a weighting factor that exponentially decays the further back in time a sample is, deemphasizing data exponentially relative to its age [Hay01].

2.3.3 Feature Extraction Using Short-Time Fourier Transform

While RLS can be effective at isolating transients, the detectors developed in Chapter 4 also used spectral analysis to develop useful distinguishing features. One powerful technique used in [Coh04] is the short-time Fourier transform (STFT).

The STFT is concerned with what frequencies are present in a signal and in what proportion when only a short interval of that signal is considered. To facilitate considering a short interval, a window centered at the time of interest is first applied to the signal. Next the Fourier transform is applied to the windowed signal. This two step procedure leads to the analysis formula

$$STFT(\tau, f) = \int_{-\infty}^{\infty} s(t)w(t-\tau)e^{-2\pi jft}dt$$

where $s(t)$ is the signal being analyzed, $w(t)$ is the window function being applied, and τ is the time around which the signal is considered. While the explanation above considers the product of the signal and the window to be a signal that the Fourier transform is applied to, the associativity of multiplication allows the product of window function and the Fourier transform kernel to be considered a new kernel of a general integral transform applied to $s(t)$ [Hop01]. This combined kernel is considered the kernel of the STFT. Like RLS, the STFT is applied in order to generate a more distinguishing feature before classification is applied.

2.3.4 Classification of Extracted Features

The usefulness of a feature for classification of a given signal can be quantified statistically and visualized in a number of ways. Three evaluations will be given for each feature examined in both detection sections, each of which serves a different purpose. A plot of the feature over time serves to locate where incorrect classifications are made and facilitates determining under what conditions these failures occur. The conditional distributions (i.e. $f(x|A)$, $f(x|A^c)$ where A represents passing vehicle noise being present) of the random feature values (x) serve as a very general statistical evaluation. The receiver operator characteristic serves as an excellent method for comparison.

Given the framework shown in Figure 2.4, classification theory is concerned with the methods by which classification rules are constructed to produce optimum results for a given set of feature data. Much of this classification literature can be leveraged to solve the detection problems addressed in this thesis. In the case of the passing vehicle noise detector's unidimensional feature, classification is reduced to a fairly simple problem. The multidimensional feature vector produced by the passing vehicle noise tolerant voice activity detector provides a good opportunity to gain performance by leveraging classification theory.

2.4 Passing Vehicle Noise Tolerant Voice Activity Detection

The second detection problem addressed in this thesis is the detection of speech while passing vehicle noise is present. Like the previous problem, well known techniques from the literature are applied to the data for the purpose of creating distinguishing features and classifying those features. In contrast to the prior problem, multichannel techniques are explored since single channel techniques alone are unable to produce acceptable results as shown in Chapter 5. The first such multichannel technique addressed, beamforming, is used to help create distinguishing features. Specifically it is used to amplify the speech and attenuate the passing vehicle noise.

2.4.1 Feature Extraction Using Beamforming

Beamformers are *multichannel* techniques designed to amplify or attenuate signals based on their direction of arrival. In the acoustical case, beamforming takes advantage of the similarities and geometric relationship of the systems coupling any single source to all microphones and the differences between that set of systems and the set that couples a different source. This is done without regard for the specifics of the sources. Adaptive beamformers address linear time-varying models by allowing the set of systems relating a source to the microphones to change over time. Proper adaptation may place requirements on or require knowledge of properties of the sources, losing some of the desirable decoupling from specifics of the sources.

Leveraging the spatial information made available by the use of arrays is discussed in [JD93]. Descriptions of delay-and-sum, superdirective, and adaptive Griffiths-Jim beamformers is provided in [Lee02]. Applying beamforming to passing vehicle noise has been addressed in a prior thesis [Fan02] where the author determined that these techniques alone were only marginally successful. Chapter 5 demonstrates that while the beamformer alone offers little advantage, it can improve the performance of the voice activity detector when combined with other feature extraction techniques.

2.4.2 Classification Using Discriminant Analysis

Unlike beamforming, the second technique addressed, discriminant analysis, is used for developing rules for making decisions based on those features. Again, the classification model described above and in [Jam85] is used to address the problem of passing vehicle noise tolerant voice activity detection. As this problem proves more challenging, a multidimensional feature is required to achieve higher performance unlike in the previous problem. Classification theory is particularly helpful when having to classify these multidimensional features. Two features are created in Chapter 5, hence a two dimensional feature vector must be classified. Two simple techniques for performing such a classification, Linear and Quadratic Discriminant Analysis, are discussed in [Han81].

The linear discriminant is concerned with finding the optimum linear combination of the elements of the feature vector. Optimum in this case is the result of a likelihood ratio test (LRT) assuming the distributions are normal. In the 2-D case at hand this produces a mapping where the score contours are straight lines. Thus selecting a score threshold selects a straight line parallel to these contours as the decision boundary.

The quadratic discriminant is concerned with finding the optimum quadratic combination of the elements of the feature vector. Like the linear discriminant it also finds an optimum assuming normal distributions but, given the quadratic form, allows for circular or hyperbolic decision regions. In the 2-D case the coefficients of each term of an equation of the form $y^2+xy+x^2+x+y = c$ are determined based on a LRT of two normals. These two discriminant techniques are applied in Chapter 5 to a two-dimensional feature vector to achieve better results than with any single scalar feature alone.

Chapter 3

Data Collection and Analysis

This chapter documents the data collection effort and the subsequent analysis of selected portions of the data collected. This work was necessitated by the lack of readily available data applicable to this problem, as discussed in Section 2.2. The acquisition and analysis of sample data for this thesis consists of three logical parts. First, the acquisition system design, construction, and verification is documented. Next, the tests performed are described. Lastly, the acquired data is characterized to produce a model by which passes can be synthesized.

3.1 Acquisition System for Recordings

The acquisition system used for collecting the data was constructed from [MPC04], a block diagram of which is shown in Figure 3.1. The microphone array is connected to a "bias box" that provides the DC bias required for the microphones to operate. Next the audio signals from the bias box as well as the signal from a speech reference boom microphone travel to an eMagic 6|2m USB recording device. The eMagic is then connected to a laptop running Cakewalk Sonar v3 which is used to perform the multichannel recordings.

Initial recordings showed occasional electrical saturation in the particular vehicles chosen, which were older and/or less expensive than those used in [MPC04] and, as such, it seems

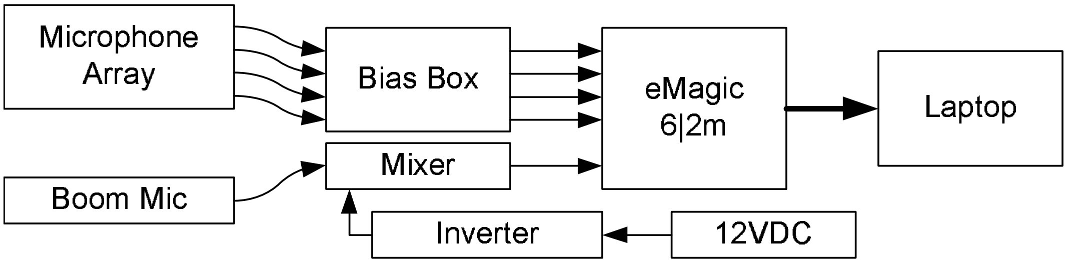

Figure 3.1: Acquisition system.

reasonable to assume they might suffer from louder wind and road noise. This electrical saturation was overcome by modifying the bias box to include an adjustable attenuation circuit as shown in Figure 3.4 and Figure 3.5 since the input scale of the eMagic is fixed. To improve the fidelity of the recordings the bias circuit was also modified to match the manufacturer's recommendation as shown in Figure 3.3. The power supply was replaced with two nine volt batteries as shown in Figure 3.2 providing cleaner power than that available from the vehicle.

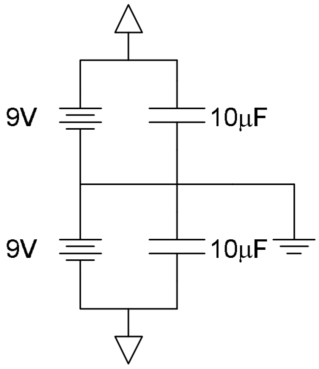

Figure 3.2: Power supply circuit.

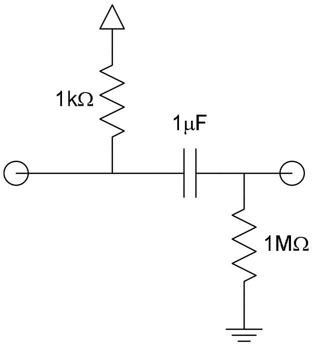

Figure 3.3: Microphone bias circuit.

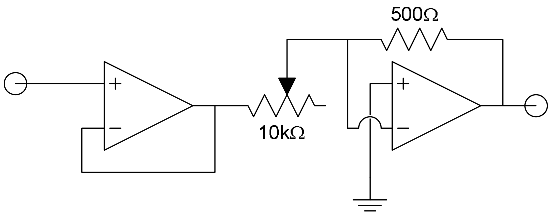

Figure 3.4: Attenuation circuit.

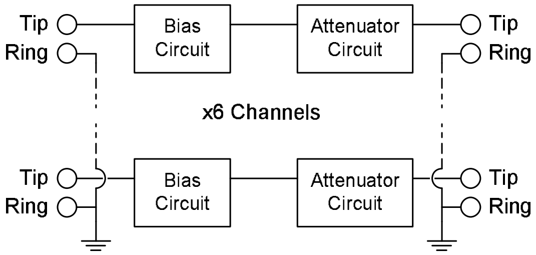

Figure 3.5: Bias box overview.

The attenuation circuit shown in Figure 3.4 is a concatenation of a buffer and an inverting amplifier. The LM741 operational amplifier was used for the buffer as well as the amplifier. The buffer isolates the circuit from dependence on the source impedance. The output of the buffer is then connected to the inverting amplifier. The inverting amplifier has a fixed feedback resistor of 500Ω and an adjustable input resistor implemented with a 20-turn 10kΩ potentiometer.

The attenuation circuit was characterized to document any frequency dependent effects. The potentiometer was adjusted to the setting used for recording (approximately 3kΩ) for the characterization. The characterization was performed by driving the circuit with a fifteen bit maximum length sequence, which represents all frequencies equally up to one half of the sampling rate, and applying a Fourier transform to the output. This signal was output via one of the eMagic's two outputs and connected to the input of the attenuation circuit as well as directly connected to one of the inputs on the eMagic, providing a reference signal. The output of the attenuation circuit was then connected to another input of the eMagic. The result can be seen in Figure 3.6. The frequency response of the attenuation circuit is flat with a maximum deviation of one decibel.

The in-vehicle recordings with the improved bias/attenuation circuit appear to be free of saturation and of a similar quality to the previous system. In recordings of silence, harmonics of 120Hz are slightly visible. This is presumed to be due to the inexpensive modified sine

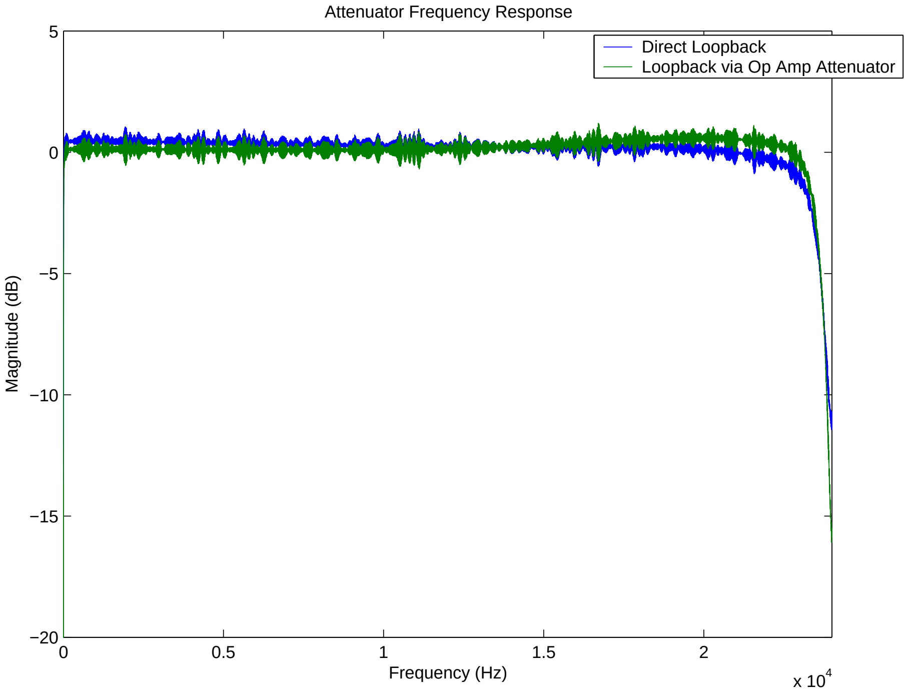

Figure 3.6: Attenuation circuit characterization. The green curve represents the frequency response of the attenuation circuit as set during recording. The blue curve shows the frequency response of the test equipment looped back into itself with no device under test.

wave inverter that powers the mixer. While this data is likely to be usable for most, if not all, applications, a replacement of the mixer with a passive balanced to unbalanced adapter will alleviate the need for the inverter in the future as the mixer was only needed to adapt the XLR microphone to the RCA input on the eMagic.

3.1.1 Recordings Produced

The set of recordings produced by this new effort is composed of separate samples of passing vehicle noise, engine noise, and speech. The passing vehicle noise recorded by this effort is *uncontrolled* unlike that of the effort described in Appendix A. The passing vehicle noise was recorded both while driving and while parked. These four types of recordings were made in two vehicles: a 2001 Hyundai Accent and a 1993 Dodge Intrepid. The latter was accompanied by video recording to provide additional documentation.

The recordings were performed in and around Worcester, Massachusetts as seen in Figure 3.7. The parked vehicle passing noise tests were performed on Pleasant Street across from Huntley Street (marked by “Start” in the figure) facing toward the rotary (south east). The passing vehicle noise recordings taken while driving were performed while driving from the intersection of Huntley Street and Pleasant Street (marked by “Start” in the figure) to just past the runway on Mulberry Street (marked by “End” in the figure and actually just outside of Worcester in Leicester, Massachusetts) or vice-versa. The speech and vehicle noise recordings were performed on Mulberry Street (marked by “End” in the figure).

All tests were conducted with the windows rolled up as well as with the windows rolled all the way down. Both the parked passing vehicle noise recordings and the speech recordings were conducted with the car off and parked on the side of the road. The passing vehicle noise recordings tests consisted of driving the marked route. The parked passing vehicle noise tests simply relied on the presence of traffic, of which there was plenty, traveling in two lanes and opposite directions. The speech tests consisted of reading two Harvard Phonetically Balanced Sentence lists [Kee04]. The vehicle noise tests consisted of accelerating to fifty

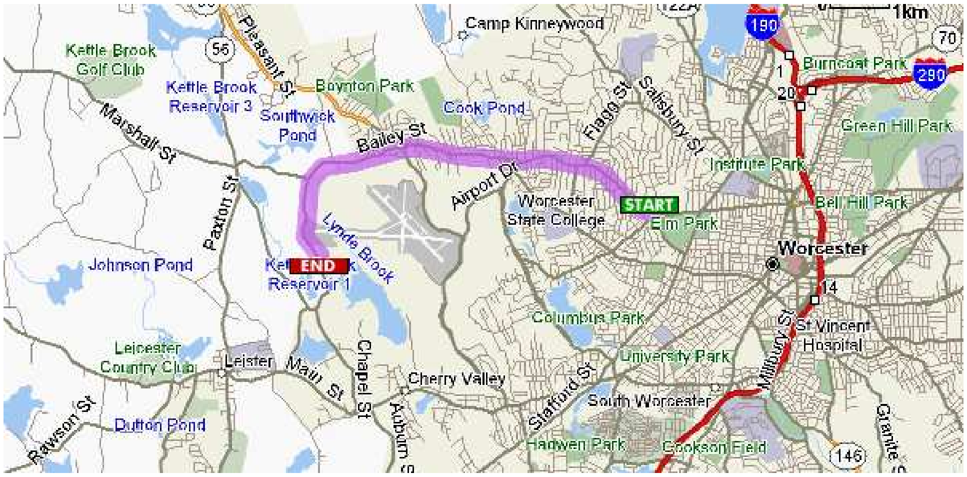

Figure 3.7: Locations and route for recordings. Driving noise recordings were performed along the route indicated between "Start" and "End". Parked noise recordings were conducted at the point marked "Start". Speech recordings were performed at the point marked "End".

miles per hour and then coasting down to thirty miles per hour. The passing vehicle noise recording taken while driving relied on traffic present along the marked route, which was in a single lane traveling in the opposite direction for the majority of the course, with the exception of a short stretch where there were two lanes in each direction.

These four types of recordings were performed for two configurations of two vehicles. This yielded a set of sixteen recordings. These recordings are presented in the structure of the final archive produced in Figure 3.8.

3.2 Analysis of Passing Vehicle Noise Recordings

The collected data were analyzed to produce a single-channel model of the signals present both for the purpose of simulation and to verify assumptions made in Chapter 2. A simulation was then constructed from the model to verify its validity. This analysis consisted of two parts. The first is the analysis of the background driving noise. The second is the analysis of the passing vehicle noise.

Directory and File Names

hyundai	winup	drive.wav	Driving between rotary and airport (passing cars).
		speech.wav	Car stationary and off with speaker reading sentence lists.
		coast.wav	Acceleration to 50 mph and coast down to 30 mph with no other vehicles present.
		stationary.wav	Parked by rotary as traffic passes.
	windown	drive.wav	Driving between rotary and airport (passing vehicles).
		speech.wav	Car stationary and off with speaker reading sentence lists.
		coast.wav	Acceleration to 50 mph and coast down to 30 mph with no other vehicles present.
		stationary.wav	Parked by rotary as traffic passes.
intrepid	winup	drive.wav	Driving between rotary and airport (passing vehicles).
		speech.wav	Car stationary and off with speaker reading sentence lists.
		coast.wav	Acceleration to 50 mph and coast down to 30 mph with no other vehicles present.
		stationary.wav	Parked by rotary as traffic passes.
	windown	**drive.wav**	Driving between rotary and airport (passing vehicles).
		speech.wav	Car stationary and off with speaker reading sentence lists.
		coast.wav	Acceleration to 50 mph and coast down to 30 mph with no other vehicles present.
		stationary.wav	Parked by rotary as traffic passes.

Figure 3.8: Directory tree for summer 2004 recordings and brief descriptions of the files. Bold indicates samples used in examples in this report.

3.2.1 Driving Noise Stationarity Analysis

To begin the analysis of driving noise, we asked the question of whether the driving noise is stationary. Stationarity was evaluated using the reverse arrangements test as described in [BP88] and applied in [CCC+05]. This test hypothesizes that the data are stationary and provides a test statistic z_a to evaluate the hypothesis. The data must be windowed in order to generate z_a. The choice of the window size is constrained by the window needing to be longer than one period of the lowest frequency we are concerned with but also short enough to expose momentary nonstationarities. A window length of 25 milliseconds was chosen as it is safely below the fundamental frequency of a male voice and still fairly short.

The reverse arrangements test was applied to the first 30 second interval not containing passes found in the driving data. The stationary hypothesis was then evaluated with a 5% significance level. The hypothesis was not rejected by the test. The p-value, which represents what percentage of time series from a stationary process would produce more extreme values of z_a (5% of which would be so extreme that they would be falsely classified as nonstationary), is 84% ($p = 0.84$). This suggests that it is very likely that the data is, in fact, stationary. This cannot, however, be proven without a doubt statistically and represents only this single sample.

3.2.2 Driving Noise Model

Given that the data is likely to be stationary, the model consists of generating a filter whose frequency response approximated the observed spectrum of driving noise in the recording. The observed spectrum was obtained by averaging many frames of the spectrogram of the recording excluding those containing the passing vehicle or any obvious anomalies. An autoregressive filter approximation was then produced using the Yule-Walker method. The very low frequency portion (< 150Hz) was then manually tuned to reduce its contribution to overall signal power. The filter magnitude, which is also the spectrum of the signal generated by filtering white noise with it, is shown in Figure 3.9.

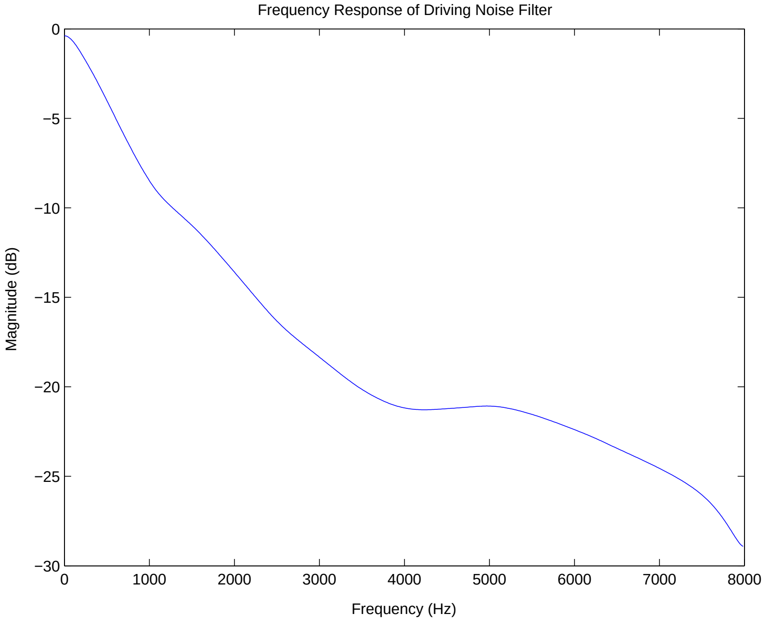

Figure 3.9: Smoothed average spectrum of background driving noise.

3.2.3 Passing Vehicle Noise Stationarity Analysis

Like the driving noise analysis, the passing vehicle noise model also began with determining its stationarity. Again a 30 second sample was chosen for analysis. To achieve conditions as similar as possible to the driving noise sample, the passing vehicle noise sample was chosen as an adjacent 30 seconds. This 30 second sample contained approximately the same background noise as well as three passing events occurring at approximately 3 seconds, 5 seconds, and 19 seconds.

The reverse arrangements test of [BP88] and [CCC^{+}05] was then applied to this sample using the same 25 millisecond window size and 5% significance level. The hypothesis that the data is stationary is rejected. Further, the value of the test statistic, z_a, indicates that there is less than a one in 10^{33} chance that a stationary process generated the passing vehicle noise sample data. The strong result of the reverse arrangements test is complemented by a plot produced by averaging the variance of 52 passes with their peaks at $t = 0$ as shown in Figure 3.10. The trend present clearly indicates the nonstationarity present in the passes.

3.2.4 Passing Vehicle Noise Model

The passing vehicle noise is modeled in a similar way to the driving noise although a method to introduce and withdraw it is also required given its nonstationarity. The passing vehicle noise filter was modeled based on averaging a small number of frames from the height of the pass. The spectrum was also modeled on approach and recession so as to verify that there were no unexpected changes. While the spectrum was not simply scaled based on distance it did maintain a similar shape through the entire event, as can be seen in Figure 3.11 where blue denotes approach and red denotes recession. The Yule-Walker method was then applied to the difference between this spectrum and the steady-state spectrum since the passing vehicle noise is added on top of the background noise. The filter magnitude generated by the Yule-Walker fit can be seen in Figure 3.12. The noise is introduced and withdrawn by multiplication in the time domain by an exponentiated Hanning window. The Hanning window provides for a smooth transition and the exponentiation serves to steepen the transition region as shown in Figure 3.13.

3.2.5 Combined Driving and Passing Vehicle Noise Model

These two models were then combined to form a laboratory simulation of the passing vehicle noise recordings. The model used for this simulation is shown in Figure 3.14. This model simulates driving noise by simply filtering a white noise source with a filter whose frequency response matches the average spectrum of driving noise as shown in Figure 3.9. The passing vehicle noise is simulated in a similar way, based on Figure 3.12, but the variation in the noise over time is accomplished by multiplying the time series produced, by the window shown in Figure 3.13.

The laboratory noise was then compared to an actual noise sample. This was first done visually, via a spectrogram. The spectrogram is shown in Figure 3.15. The sample recording is shown on the left, and the laboratory sample (including speech) is shown on the right. With the exception of not being quite as intense in the lower frequencies and a lack of the

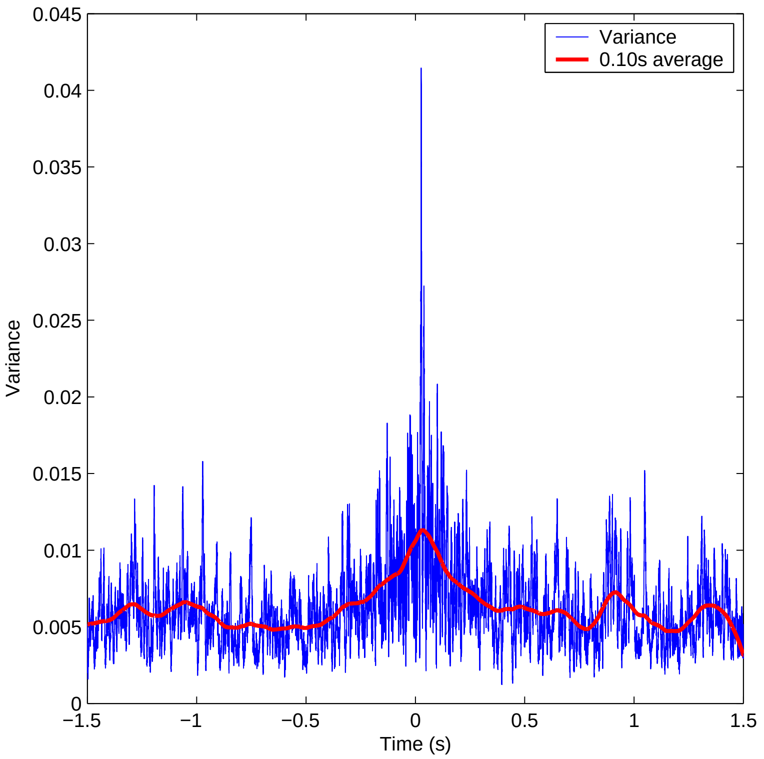

Figure 3.10: Time series of variance of passing vehicle noise.

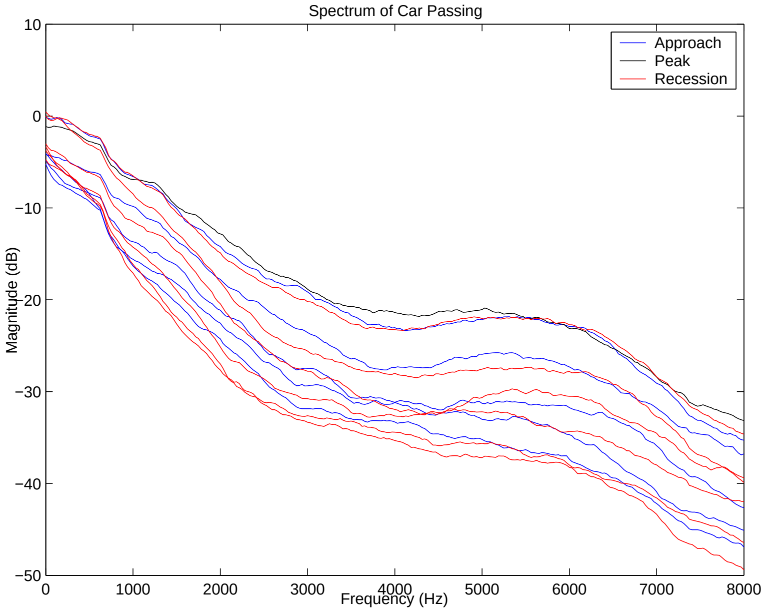

Figure 3.11: Smoothed spectrum of passing noise at different moments in time.

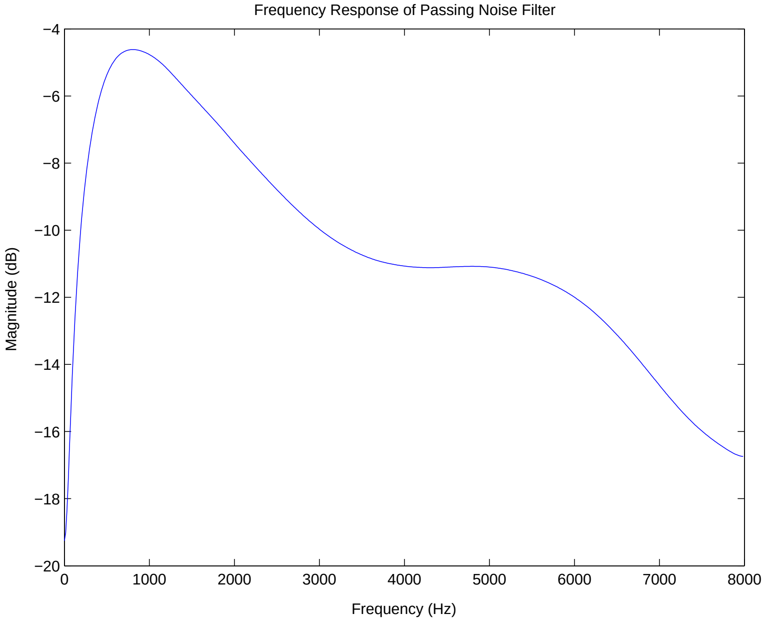

Figure 3.12: Smoothed spectrum of pass at peak minus background noise.

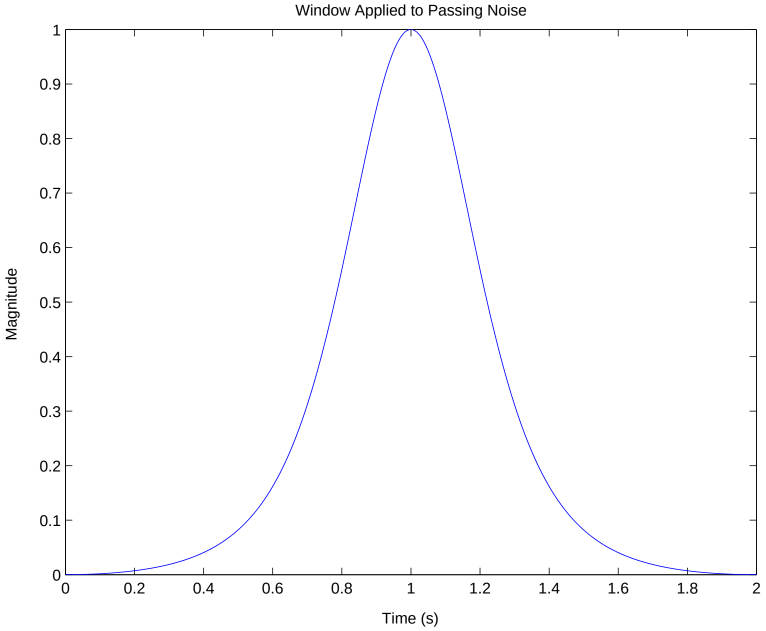

Figure 3.13: Window used to introduce and conclude the pass.

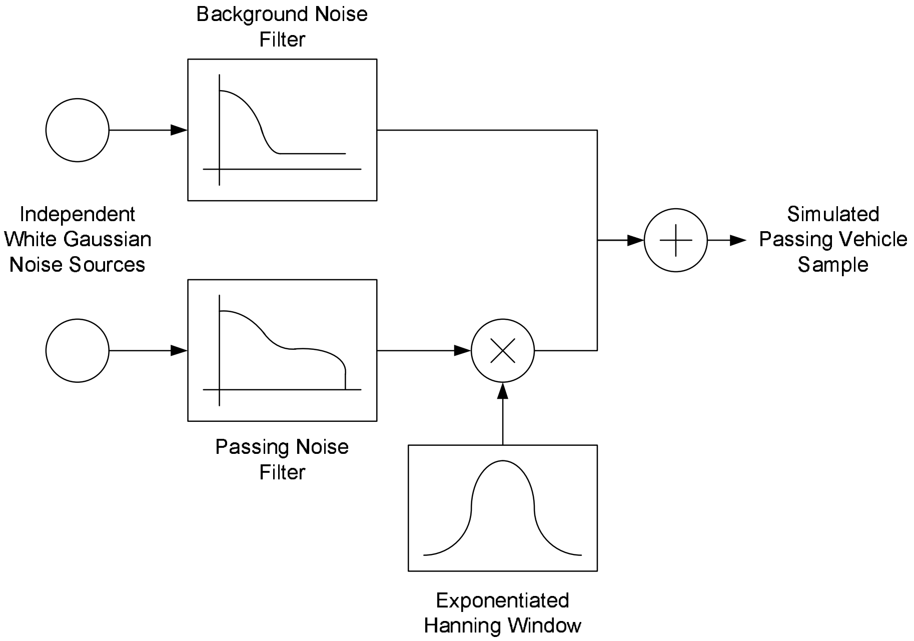
Background Noise Filter
Independent White Gaussian Noise Sources
Simulated Passing Vehicle Sample
Passing Noise Filter
Exponentiated Hanning Window

Figure 3.14: Block diagram of passing vehicle noise model.

vague banding present in the original signal below 1.5kHz, the simulation appears to be a fairly good fit. Next, informal subjective listening tests were performed. The five subjects all felt that the samples were very similar and one of them, not knowing the first one played was the simulation, mistook it for the actual recording.

A very interesting result pertains to the claim that a Doppler shift is audible in the original. Many claim to hear this, although clearly no Doppler shift is present in the simulation, and this claim is retracted after hearing the simulation and claiming that the same Doppler shift is present. This is probably due to the Doppler shift claim being based on the auditory perception of an approaching and then receding object. Given the results of listening to the simulation, the perception of the approach and recession is most likely based on the amplitude profile of the introduction and withdrawal of the noise and not a shift in frequency. This is supported by the very similar spectral profiles of approach and withdrawal in the original recording, as seen in Figure 3.11. While Doppler shift estimation techniques are available in literature such as [RZB97], these listening tests suggest that there appears to be little need to apply them to this problem.

The verification of this model, both visually, via the spectrogram, and via listening tests, indicates that it is a fairly good simplification of the phenomena present. The structure of this model is then used in Chapters 4 and 5 to guide the development of features via signal processing.

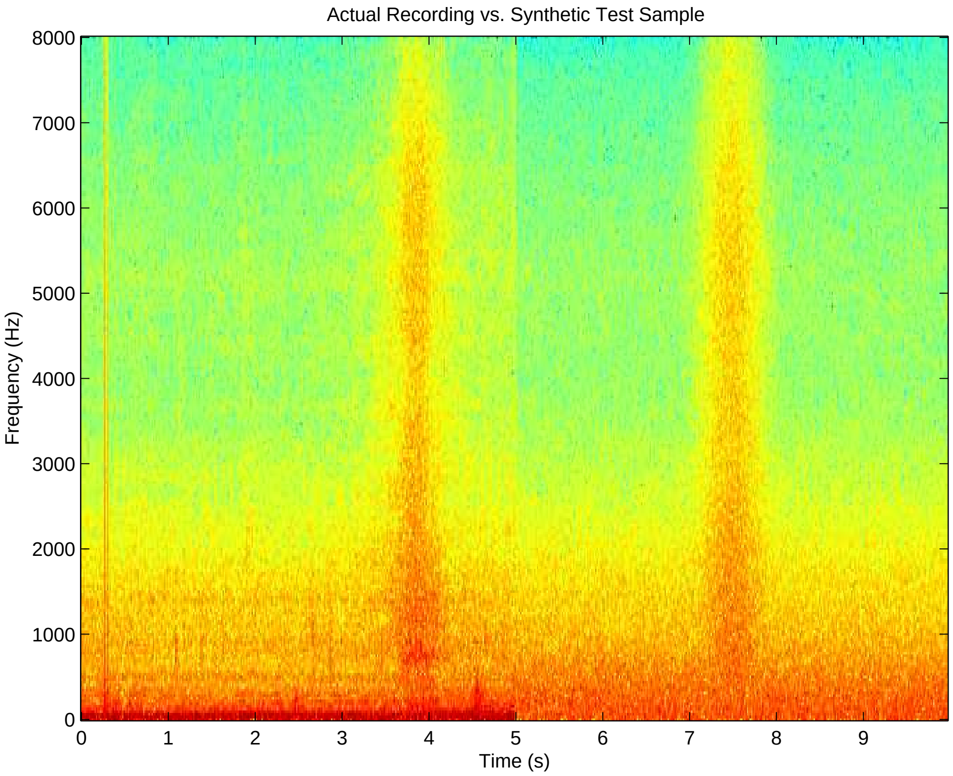

Figure 3.15: Actual (left) passing vehicle noise vs. simulated noise (right).

Chapter 4

Passing Vehicle Noise Detection

Passing vehicle noise detection is the first of two detection problems addressed in this thesis. This section describes the evolution and performance of a passing vehicle noise detector that is robust in the presence of speech. This implies our first two out of three goals: good detection performance, that is, maximum true detections and minimum false detections, and robustness, that is the likelihood of the technique generalizing well to other samples. Our third goal, which in practice is often at odds with the first two, is low processing delay or latency.

The development of the detector, guided by these goals, begins with the description of the test sample. Next, a series of features are described and evaluated in the context of passing vehicle noise detection using the test sample. The development concludes with the combining of features using classification.

4.1 Sample Used for Evaluations

An appropriate test sample is required for the evaluation of the detection scheme. To effectively evaluate the detection scheme the sample must include not only the signal to be processed but the truth as to whether passing vehicle noise is actually present. A single test sample is used consistently throughout this section. The use of the same sample facilitates easy comparison of techniques.

The sample is composed of separate speech and passing vehicle noise samples combined through addition. As noted in Chapter 3, these recordings were taken in identical acoustical environments making such superposition a sound practice. The sample of speech being separate allows for the detailed classification of speaker state. In this case a determination of whether speech is present was made for every 10ms window. Likewise this separation allows for a more accurate classification as to whether passing vehicle noise is present or not in each 10ms window of the noise sample. A spectrogram of this sample is seen in Figure 4.1.

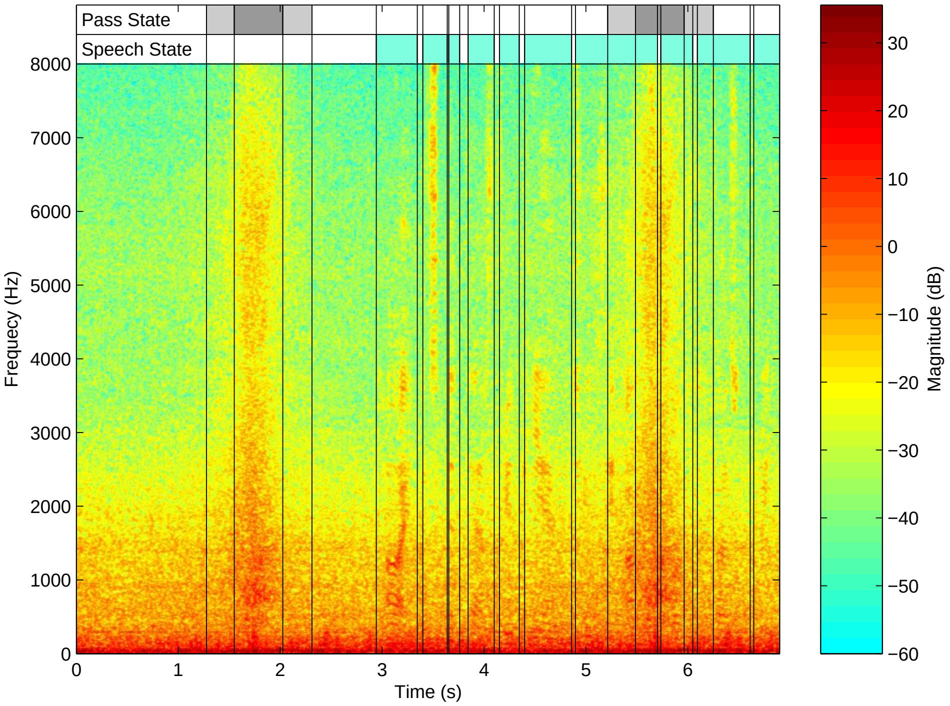

Figure 4.1: Spectrogram of test sample used for passing vehicle noise detector evaluation and development.

4.2 Development of Distinguishing Features

This section describes the incremental development of several scalar features used to detect passing vehicle noise. An outline of this development is provided in Figure 4.2. Ultimately, the test sample described in Section 4.1 is classified perfectly. Because this level of performance can be achieved with scalar features based on single-channel data, multichannel signal processing techniques and multidimensional classification are both unnecessary.

Each feature begins with a general description, optionally followed by a detailed explanation and concludes with an evaluation using three different types of plots. These plots all follow the same format and are described in detail in only the first feature to avoid needless repetition.

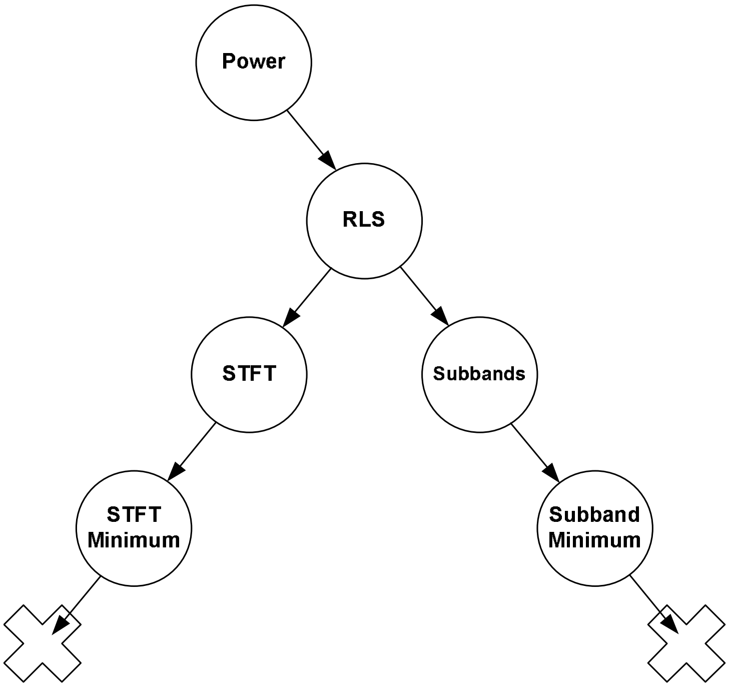

Figure 4.2: An outline of the development of and relationship between features leading up to the passing vehicle noise detector.

4.2.1 Power

The first feature examined in the context of detecting passing vehicle noise and a classic feature in *voice* activity detection is signal power. Instantaneous signal power is evaluated first, followed by a variety of time averages. These results are then compared.

Instantaneous power is found by simply squaring each sample of the recorded signal, as seen in Figure 4.3. The first plot of this feature's performance, shown in Figure 4.4, is a time series plot. This plot is augmented with the true content or state of the signal as shown via background shading. The shading consists of two levels of gray for passing vehicle noise. The darker gray indicates where we consider the event to be present. The lighter gray indicates a margin of transition where the vehicle noise has begun to appear but we do not consider it fully present. For the purposes of evaluating the detector, any activation outside both gray regions is a false detection and lack of activation within the dark gray region is a missed detection but the behavior of the detector within the light gray region is not of concern. The true presence of speech is also shown using teal shading. Examining the time series plot for instantaneous power reveals that it is probably not a particularly distinguishing feature. No clear trend related to the gray shading is visible in the time series data.

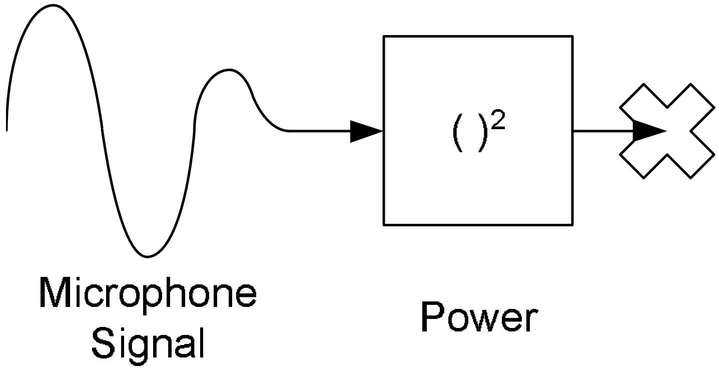

Figure 4.3: Structure of the power feature.

Next, a statistical profile of the instantaneous power feature is shown in Figure 4.5. This plot shows the estimated probability distribution function (PDF) of the feature values conditioned on which gray region they fall in. The PDF is estimated by interpolating the points of a histogram. Three conditional distributions are overlayed and each is normalized individually to facilitate easy visual comparison. The estimated distribution for feature values when no pass is present (outside both gray regions) is shown in blue. The estimated distribution when a pass is present (inside the dark gray region) is shown in red. The estimated distribution of feature values in the margin between (inside the light gray region) is shown in gray. This last distribution does not directly play a role in the accuracy of detection but is included for reference. This plot reveals a slight difference in mean instantaneous power between regions where a pass is occurring and regions where it is not. It is also clear that this difference in mean is dwarfed by the large variance of both and that there is no separation for low values of power, although they are quite prevalent. This is due to the fact that sound is an AC signal and since many high frequencies are present in the test sample there are frequent zero crossings. So while the plot makes it clear that higher amplitudes are achieved while passing vehicle noise is present, many samples are present between this amplitude and zero due to the constant zero crossings. The massive overlap in the distributions that results prevents any clear separation of states based on the value of the instantaneous power.

The ability, or lack of ability, of instantaneous power to differentiate between the passing vehicle noise present state and the passing vehicle noise absent state is quantified by a receiver operating characteristic (ROC) curve, shown in Figure 4.8. The plot shows the relationship between the percent true positive detections, that is when the detector believes passing vehicle noise is present and in reality it is, versus percent false positive detections, that is when the detector believes that passing vehicle noise is present but in reality it is not, for all threshold settings. It is important to note that while selecting a threshold corresponding to any point on the ROC curve provides the noted level of performance when used with the sample it was generated from, it may not, and likely will not, generate the same level of performance for another sample. Since there is a different detection result for choosing a threshold between any two consecutive unique feature values the curve appears quite smooth, despite being generated from and evaluated on a single test sample.

This plot demonstrates that instantaneous power can provide 40% true positive detections with a rate of only 20% false positive detections, which is surprisingly good considering the appearance of the time series. This level of accuracy is, however, fairly low. For comparison, *perfect detection*, at least for this test sample, would correspond to a vertical line along the left edge of the plot, indicating that up to 100% true positives can be detected while not incurring any false positives and then a horizontal line across the top of the plot showing that as we raise the threshold beyond its optimal point we will begin incurring false positive detections until all feature values are below the threshold. The opposite case is illustrated by the diagonal line extending from the bottom left to the top right. Randomly making a detection decision with no regard for the data or phenomenon would result in an ROC curve along this line. Unfortunately the plot demonstrates that instantaneous power is closer to this latter case of random selection than the former case of perfect detection.

The instantaneous power results suffer from the variance overtaking the difference in means. A reduction in variance can be achieved through time averaging. Time averages over one hundredth, one fiftieth, one twentieth, and one tenth of a second are produced via Figure 4.6. The time series for these features are shown at top of Figure 4.7. A trend related to the region colors is now more apparent. The separation is now great enough to clearly demonstrate the effect of a zero false positive detector, which is shown via two horizontal lines. The uppermost line indicates the maximum amplitude of the feature values over time. The lowermost line indicates the maximum amplitude of the feature values outside of the gray regions. Therefore the detector can be active for any amplitude that falls between these lines and achieve a zero false positive rate. The activation of such a detector is shown via a red region.

A trend is also apparent in the conditional distributions. The longer the moving average, the more tightly the power values cluster around their centers of mass. While this achieves better separation, it also increases processing delay. A comparison of the classification performance of these features is provided in Figure 4.8. Here the substantial gains in detection performance available at the expense of latency are quite apparent.

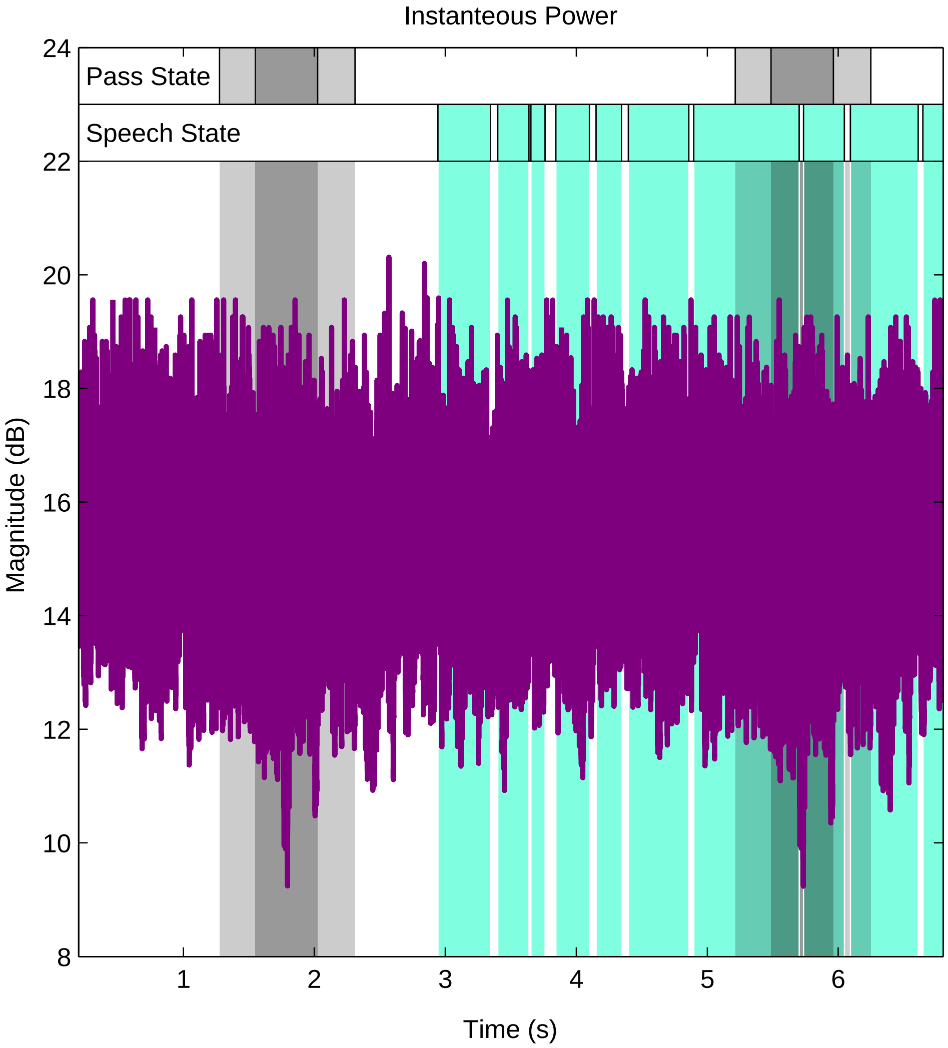

Figure 4.4: Instantaneous power over time.

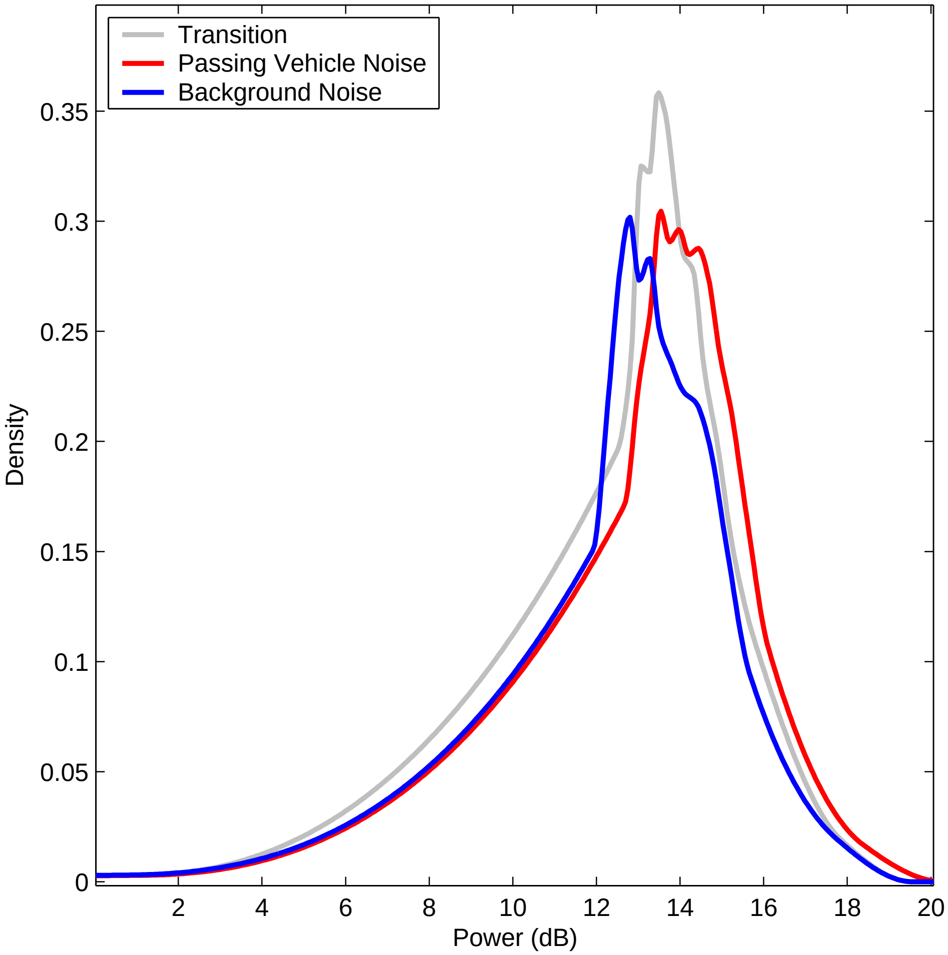

Figure 4.5: Conditional distribution of instantaneous power.

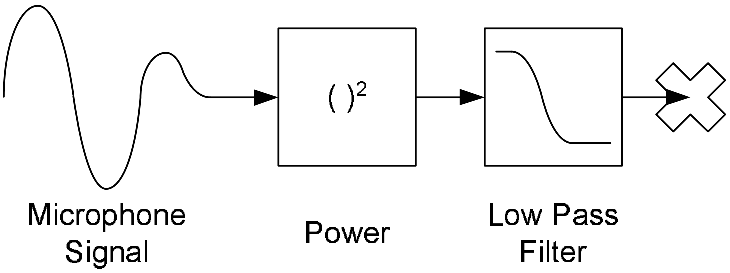

Figure 4.6: Structure of average power feature.

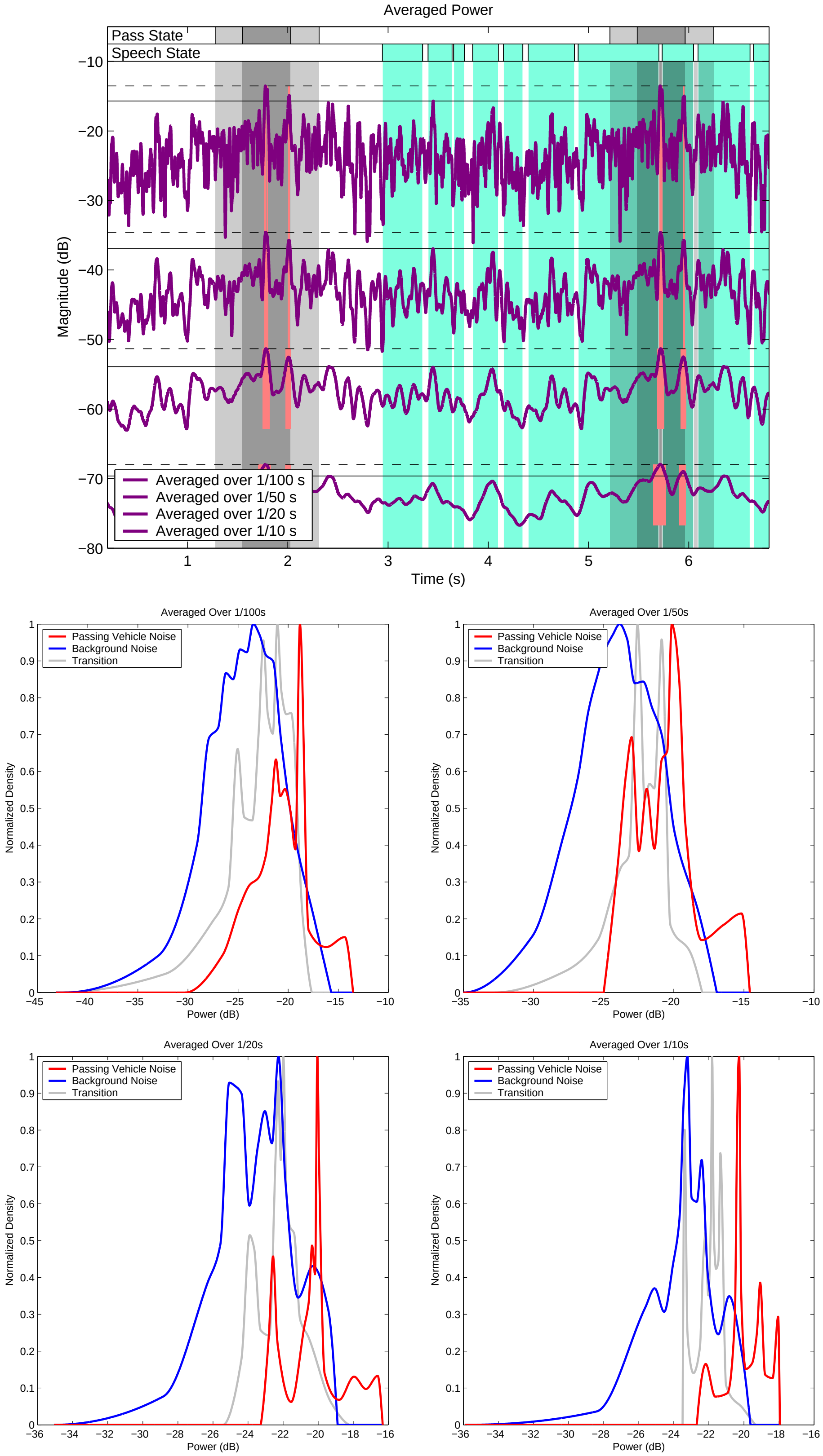

Figure 4.7: Analysis of average power for four averaging intervals.

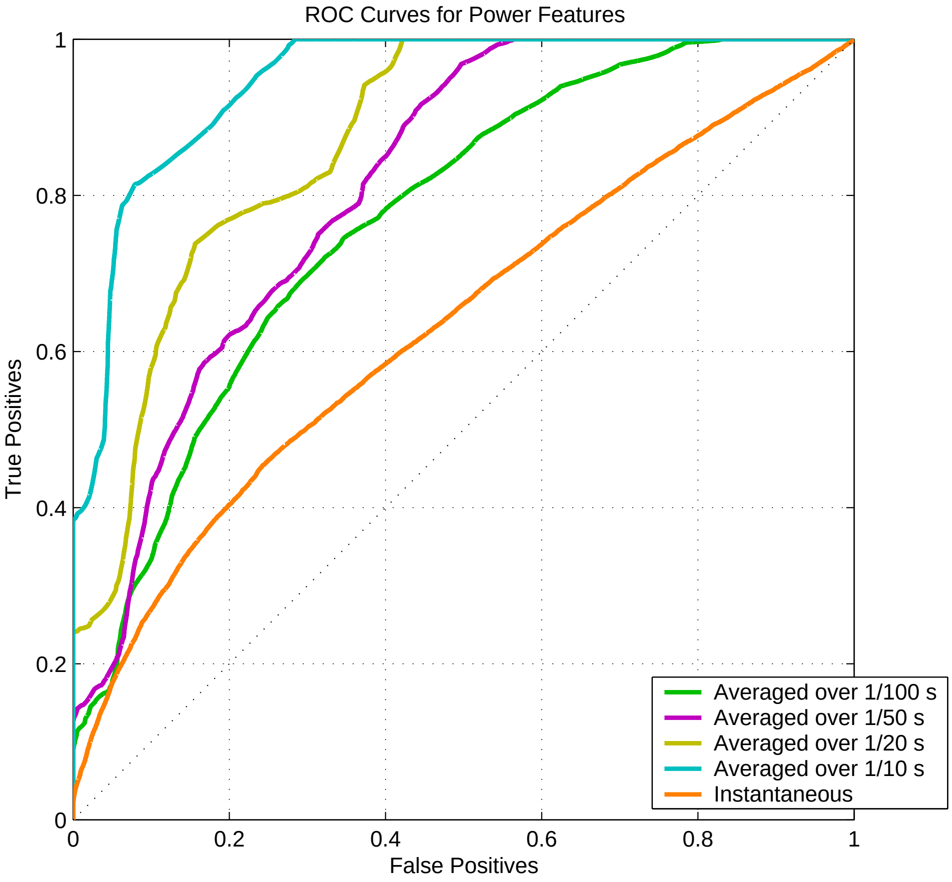

Figure 4.8: Receiver operating characteristic plots for instantaneous and averaged power.

4.2.2 Recursive Least Squares

One problem with the average power technique, in addition to the processing latency, is that it does not directly exploit the nonstationary nature of the passing vehicle noise. The recursive least squares prediction filter, as mentioned in Chapter 2, is able to provide some differentiation based on nonstationarity.

The recursive least squares prediction filter attempts to predict future inputs based on previous inputs. The signal consisting of these predictions then represents the 'predictable' component of the signal and the signal composed of the difference between our prediction and the actual signal then represents the 'unpredictable' portion of our signal. What is considered predictable is both a function of the structure of predictor and the value of its parameters.

This filter seems well suited to at least one of our two problem signal components. While the predictability of speech is questionable, one would expect background noise in an automobile to be fairly predictable given that much of it is made either by moving parts, whose speed of rotation cannot change quickly due to inertia, or at least by parts whose behavior is synchronized to one or more of these rotating parts.

RLS is applied to the data as a preprocessing step and the prediction error is used to extract the 'unpredictable' portion of the signal as shown in Figure 4.9. RLS provides two axes of parametric adjustment. The filter is evaluated over a variety of filter orders while fixing the forgetting factor at 0.98. These results are shown in Figure 4.10. Examining the time series we see substantial background noise reduction as well as some reduction in the speech portion of the signal. This is attributable to the continuous adaptation of the filter which allows it to react to fast changes in the signal's statistics. The ROC plot shown in Figure 4.11 demonstrates a clear improvement in accuracy over the power features.

The ROC curve clearly demonstrates that increases in filter order do not significantly impact the accuracy given this test sample. This result is best explained using a plot of the frequency response of the filter generated by RLS over time, shown in Figure 4.12. Orders

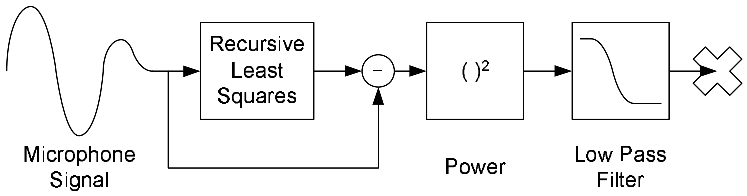

Figure 4.9: Structure of the recursive least squares feature.

of 2, 5, and 20 are shown. All three orders demonstrate a clear high pass filter behavior with a similar cutoff frequency that is consistent over time. While some additional fitting is apparent in the higher order filters, the additional bands are comparatively small and short-lived as well as occurring at frequencies where the signal is much less powerful. This assures that the impact of this additional fitting is quite small.

Next we explore the effect of changing the forgetting factor. Figure 4.13 shows four choices of forgetting factor from slowest adaptation (not very 'forgetful') at a λ of 0.99999 to extremely fast adaptation (very 'forgetful') at a λ of 0.80. Slow adaptation occurs with the high values of λ. Here the filter is not adapting quickly enough to attenuate the unvoiced speech. As we increase the speed of adaptation we see a substantial decrease in amplitude of the speech peaks. While its unclear from the time and conditional distribution plots whether we have passed an optimum, very close scrutinization of the ROC plot shown in Figure 4.14 indicates that the best performance is achieved at $\lambda \approx 0.98$ for this test sample.

Preprocessing with RLS improves detection accuracy without the latency of longer averaging. This comes at an increase in computational complexity, however.

4.2.3 Subband Power

In addition to separation by prediction, the signal can also be separated by frequency. While this could be used *instead of* predictive separation it can also be used *in addition* to predictive separation. An examination of a spectrogram of the RLS prediction error signal as seen in Figure 4.15 reveals spectral characteristics that can be exploited. Most notably the power

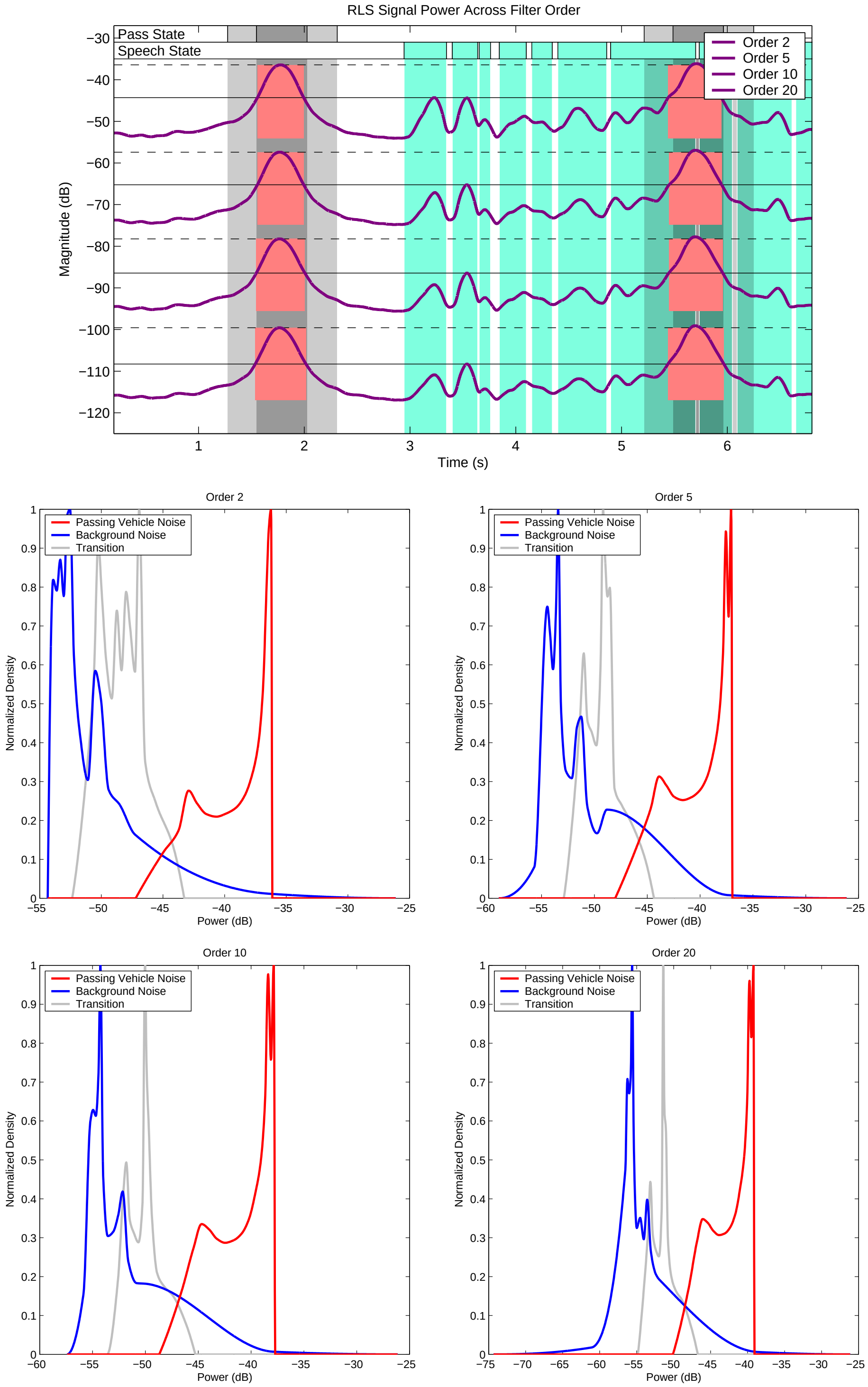

Figure 4.10: Recursive least squares across filter order with a forgetting factor of 0.98.

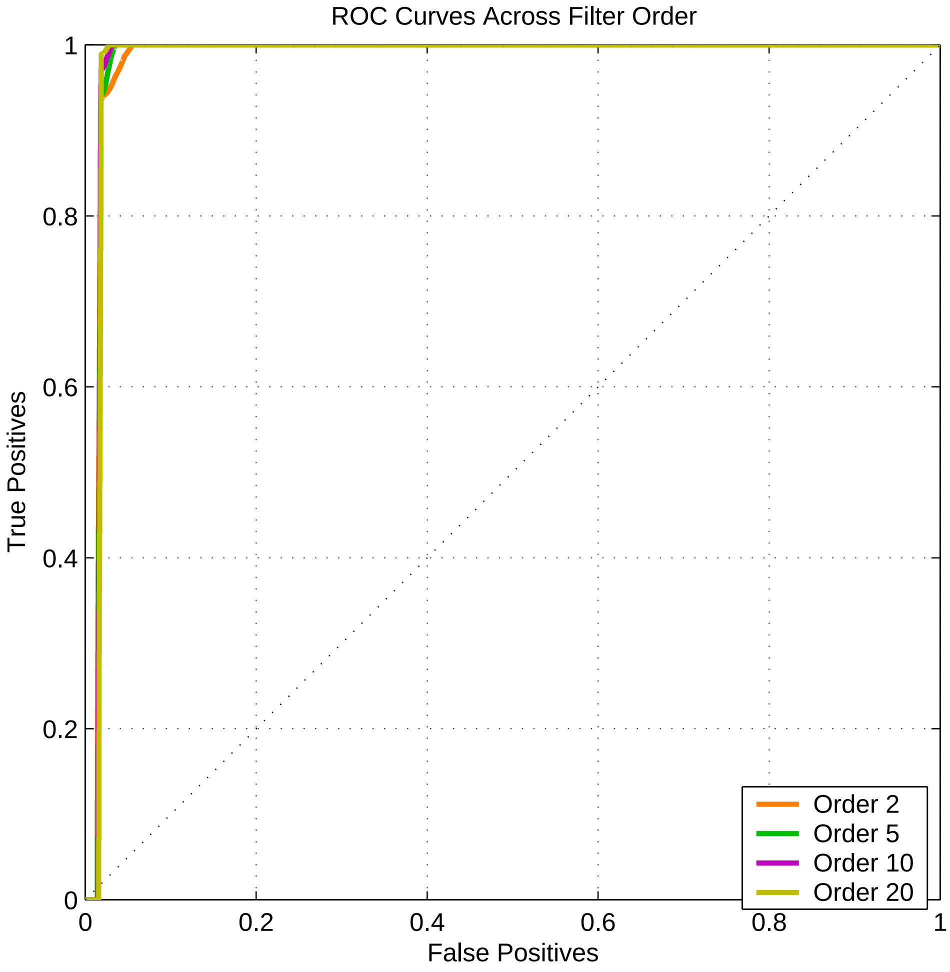

Figure 4.11: Receiver operating characteristic for recursive least squares across filter order with a forgetting factor of 0.98.

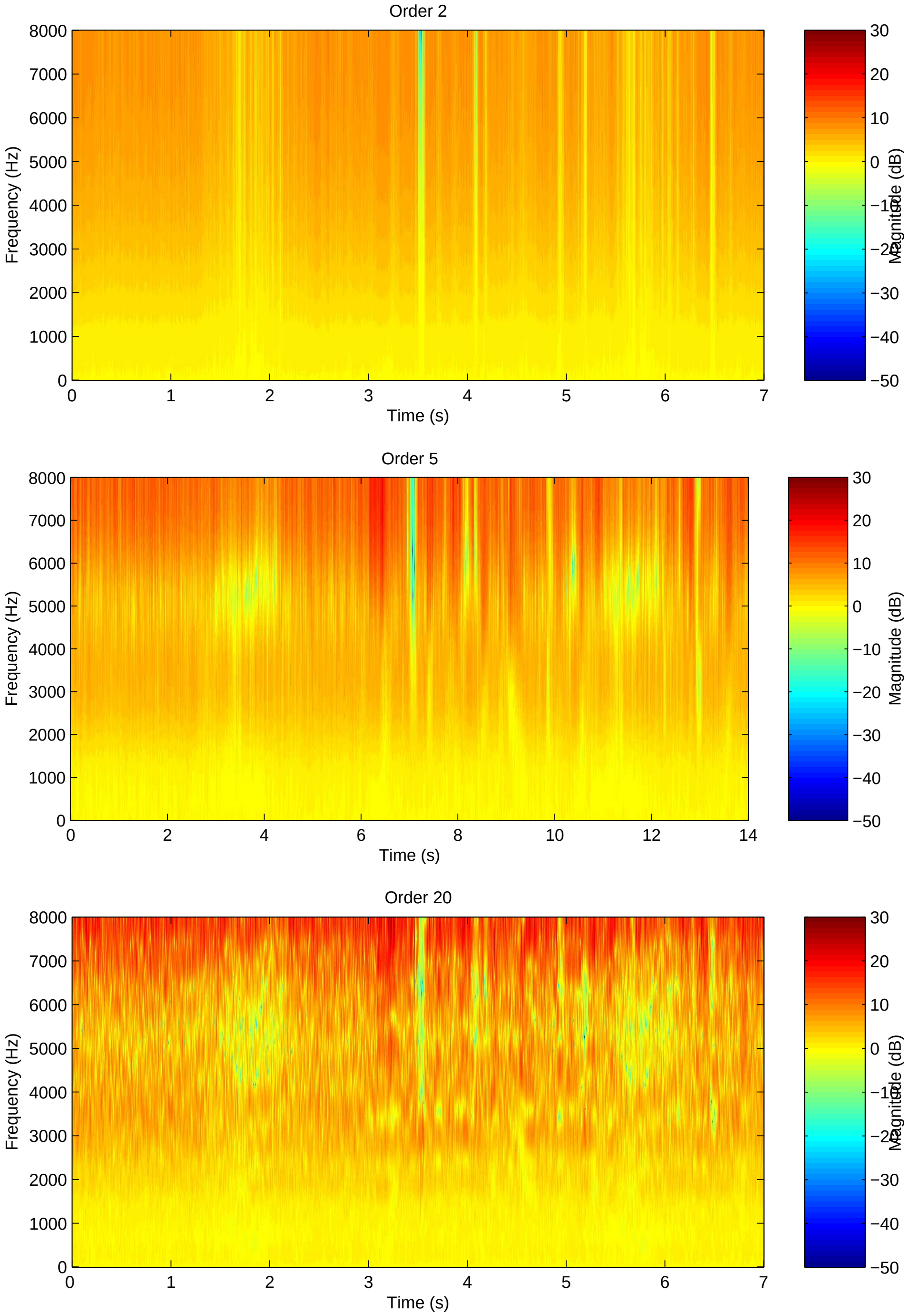

Figure 4.12: Frequency response of filter generated by RLS over time for three orders and a forgetting factor of 0.98.

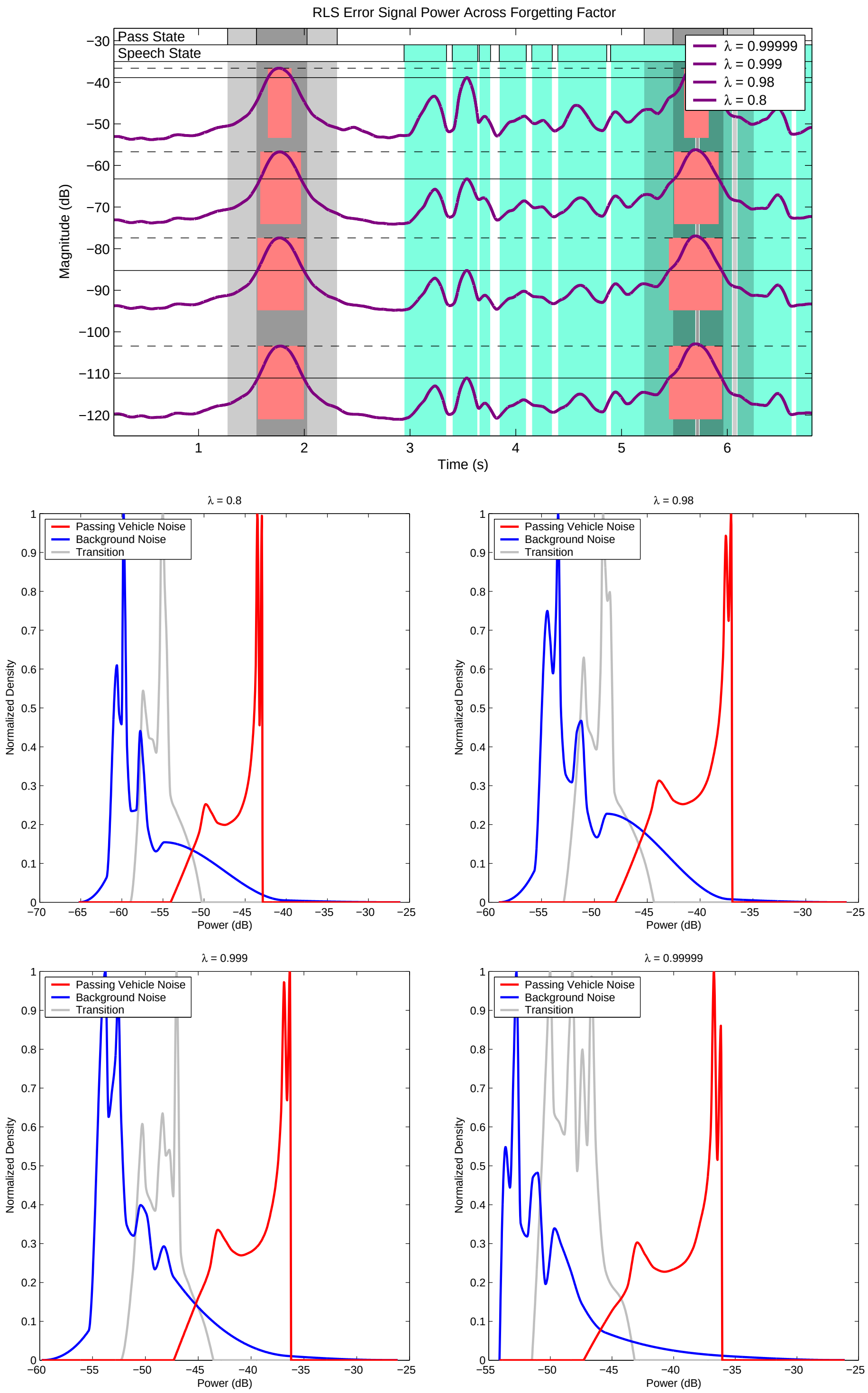

Figure 4.13: Recursive least squares across forgetting factor for an order of two.

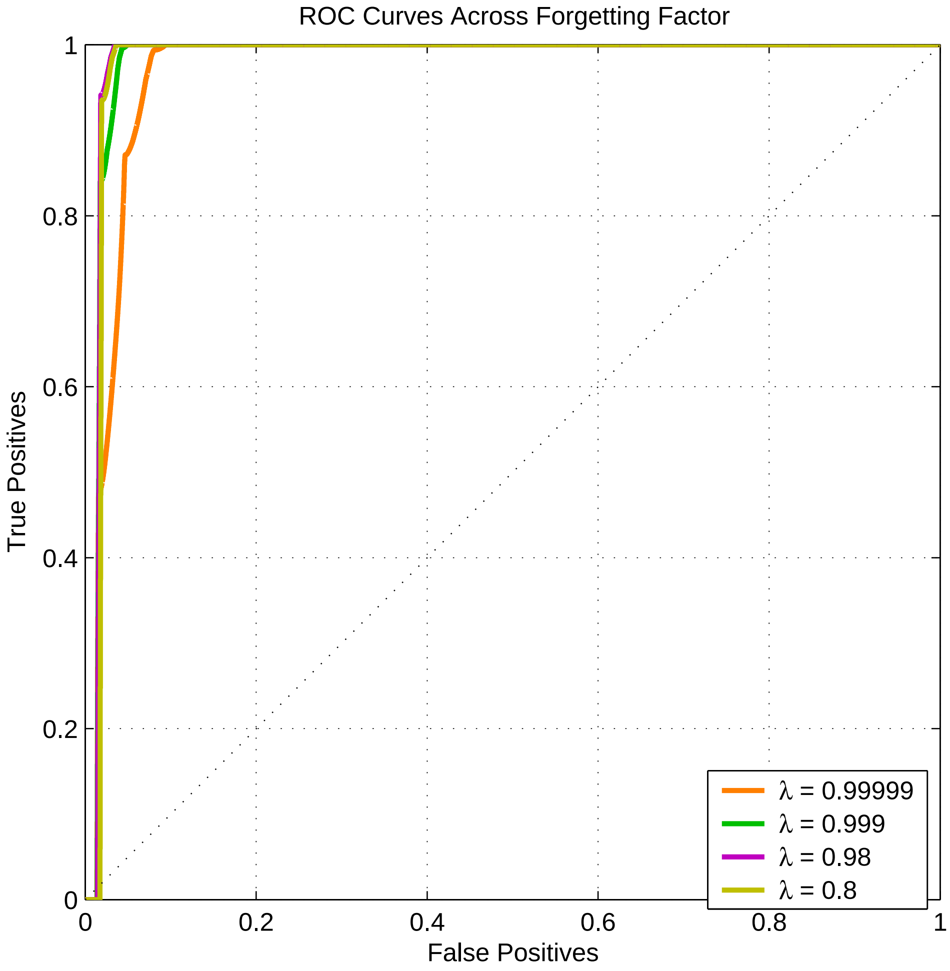

Figure 4.14: Receiver operating characteristic for recursive least squares across forgetting factor for an order of two.

of the pass is much more evenly distributed across the spectrum than the speech.

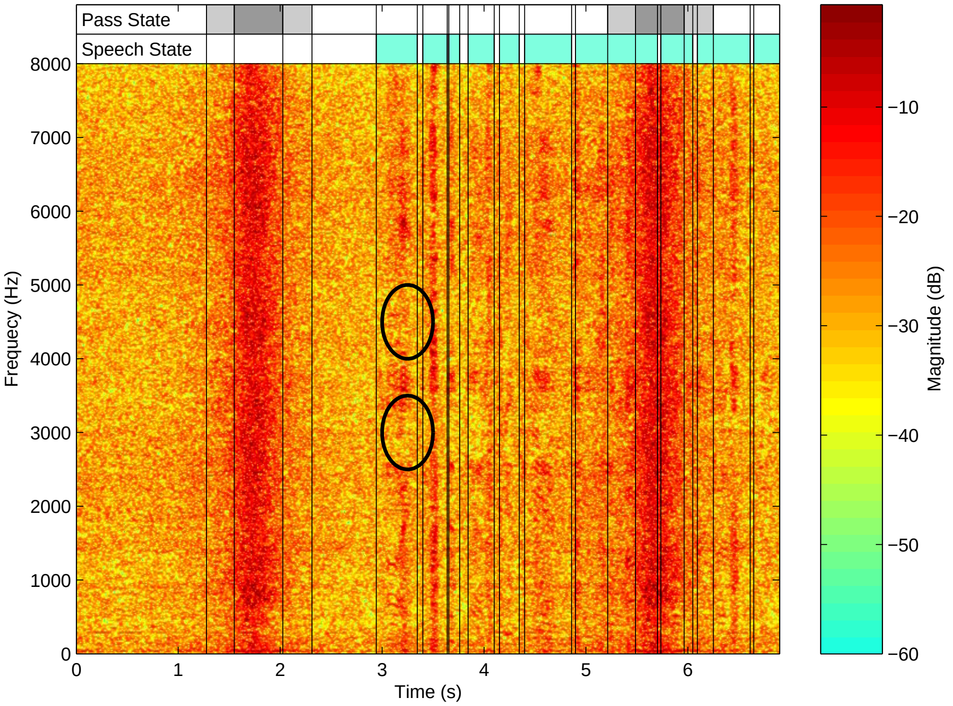

Figure 4.15: Spectrogram of RLS error signal.

The subband power feature is implemented as a set of six adjacent bandpass filters. These filters are each 1kHz wide. The first has its low band edge at 1kHz and the last has its upper band edge at 7kHz. This forms a filter bank with center frequencies of 1.5kHz, 2.5kHz, 3.5kHz, 4.5kHz, 5.5kHz, and 6.5kHz. This structure is shown in Figure 4.16.

The time series shown in Figure 4.17 reveals that there are substantial gains to be made with these features. The passes are very consistent across the majority of the spectrum as evidenced by the fact that the power is nearly identical in all six bands. Speech, however, is clearly not consistent across the six bands as the power in each peak is quite different.

Many of the conditional distributions shown in Figure 4.18 are quite favorable. A closer look reveals that the conditional distribution for the 3kHz-4kHz band has no overlap between feature values during passing and feature values not during passing, thus achieving perfect separation for this test sample, even if only by a small margin. This is evident on the ROC plot shown in Figure 4.19 as the 3.5kHz line goes straight to 1.0 without diverging from the

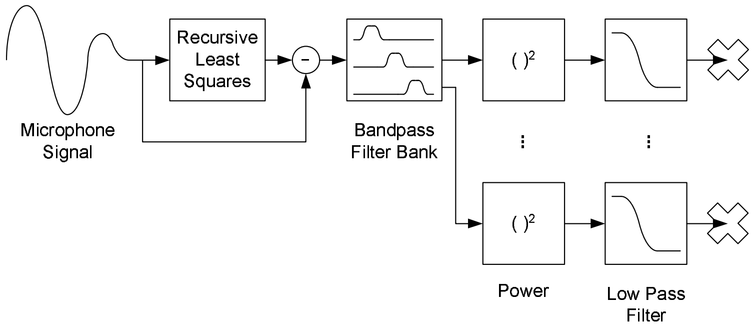

Figure 4.16: Structure of subband features.

vertical axis. It is important to recall that this performance is specific to this test sample and given the dependence on a feature of speech being present at a certain frequency, this result is less likely to produce similar results on other test samples than previous features. While we cannot appropriately evaluate the robustness of this feature without many more test samples, the selection of an arbitrary frequency range motivates the development of a feature more likely to be robust.

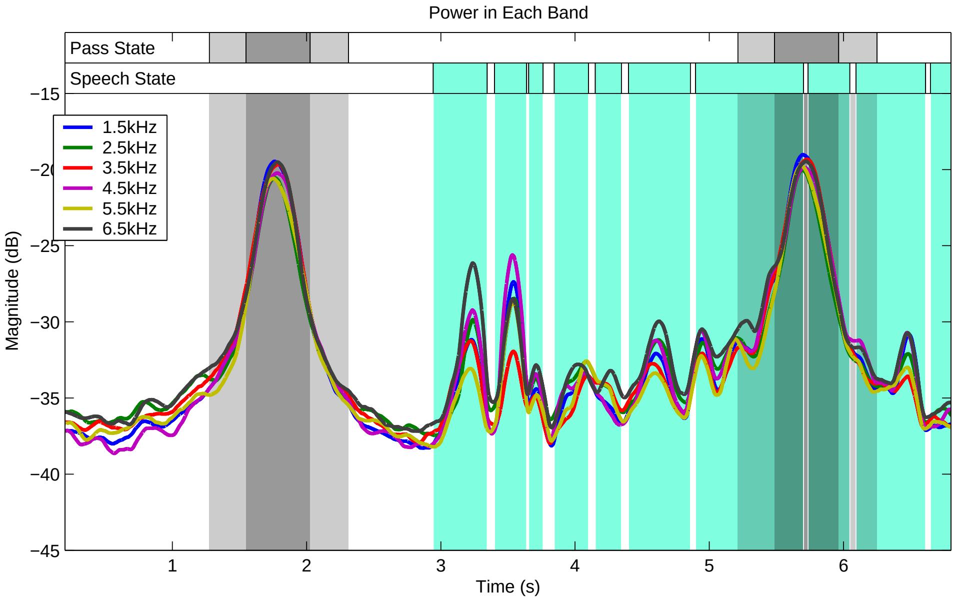

Figure 4.17: Power in different bands for the subband features as specified in Figure 4.16.

Given the perfect separation achieved with this feature we cannot improve further on

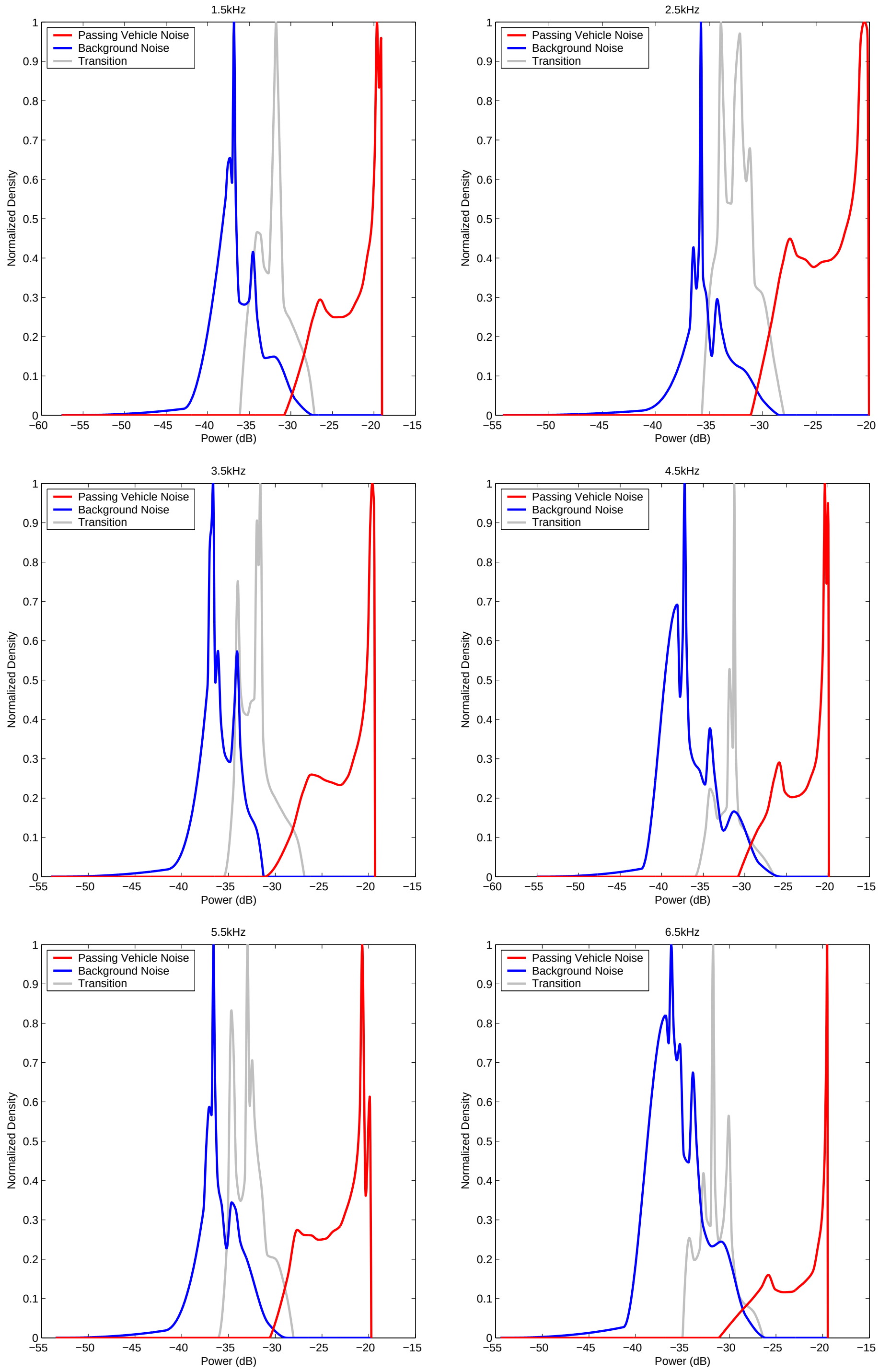

Figure 4.18: Power in different bands for the subbands features as specified in Figure 4.16.

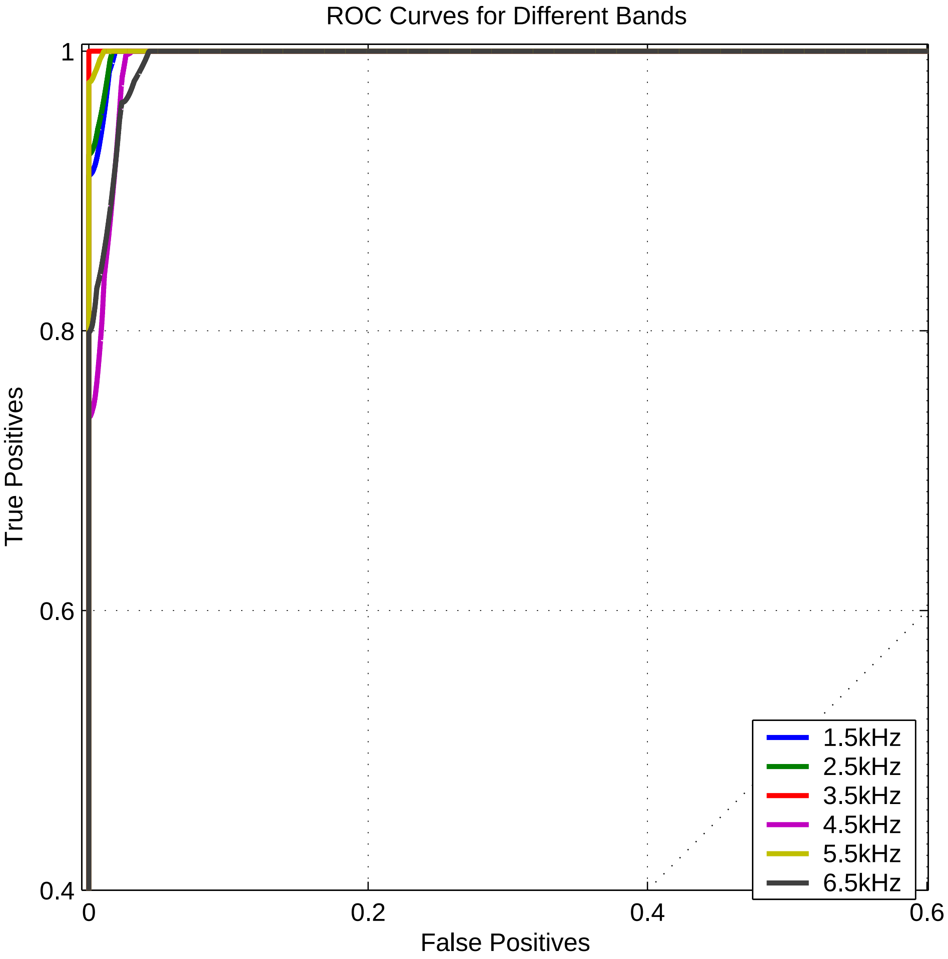

Figure 4.19: Receiver operating characteristic for power in different bands of the subband features as specified in Figure 4.16.

the accuracy of this feature. We can, however, address robustness, that is, increasing the likelihood that the detector will perform well on other data sets, and improve latency.

4.2.4 Subband Minimum

Greater robustness can be introduced into the subband technique by using the output of more than one filter. If we reexamine the time series in Figure 4.18 we see that not only is the speech substantially less powerful in some bands but we see also that which band it is least powerful in varies. This can be leveraged by computing the power in a number of bands and then using the minimum of those values as shown in Figure 4.20. In the case of the pass they should all be fairly similar, given the model discussed in Section 3.2, so selecting the minimum has little impact. In the case of the speech, however, the uneven distribution should produce bands that have a much lower power than others and hence produce a substantially lower minimum.

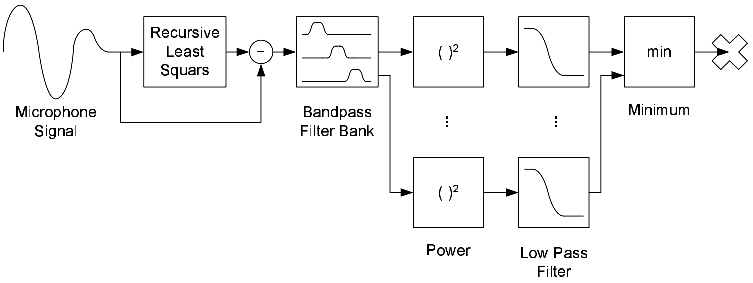

Figure 4.20: Structure of subband minimum feature.

The minimum should not hurt the accuracy as long as all bands are greater than or approximately equal to the 5.5kHz band during the pass, which appears to be the case in Figure 4.18. The minimum should attenuate the peaks seen in the speech regions even further. This is shown in Figure 4.21. The separation between the conditional PDFs is maintained as seen in Figure 4.22 and the minimum should serve to shift the center of mass of the pass-not-present PDF further down. Figure 4.23 clearly shows that there is no degradation in accuracy. While the perfect separation achieved here is specific to this test sample and separation is unlikely to be perfect for all samples, this feature should generalize better than selecting only a single band from the previous feature as it allows for frequency shifts in the properties of the speech.

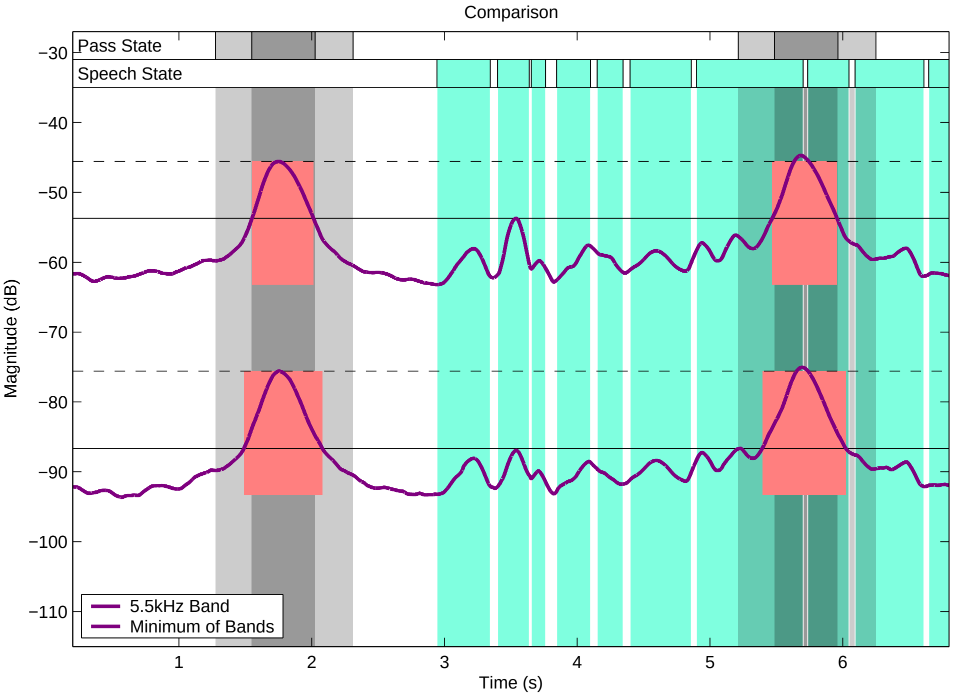

Figure 4.21: Minimum power in any band over time.

The minimum power across all subbands technique maintains the accuracy of the previous technique and boosts robustness. The bandpass filter may, however, introduce unacceptable processing latency. In the following section, we can examine a similar method that allows for decreased latency.

4.2.5 Short-Time Fourier Transform

The output of a short-time Fourier transform can often replace the output of a filter bank. The bin values of the STFT over time represent the downsampled output of a filter bank whose filters have abutting band edges. This is the case with the previous feature. This provides a low latency and computationally efficient implementation of the same idea. While

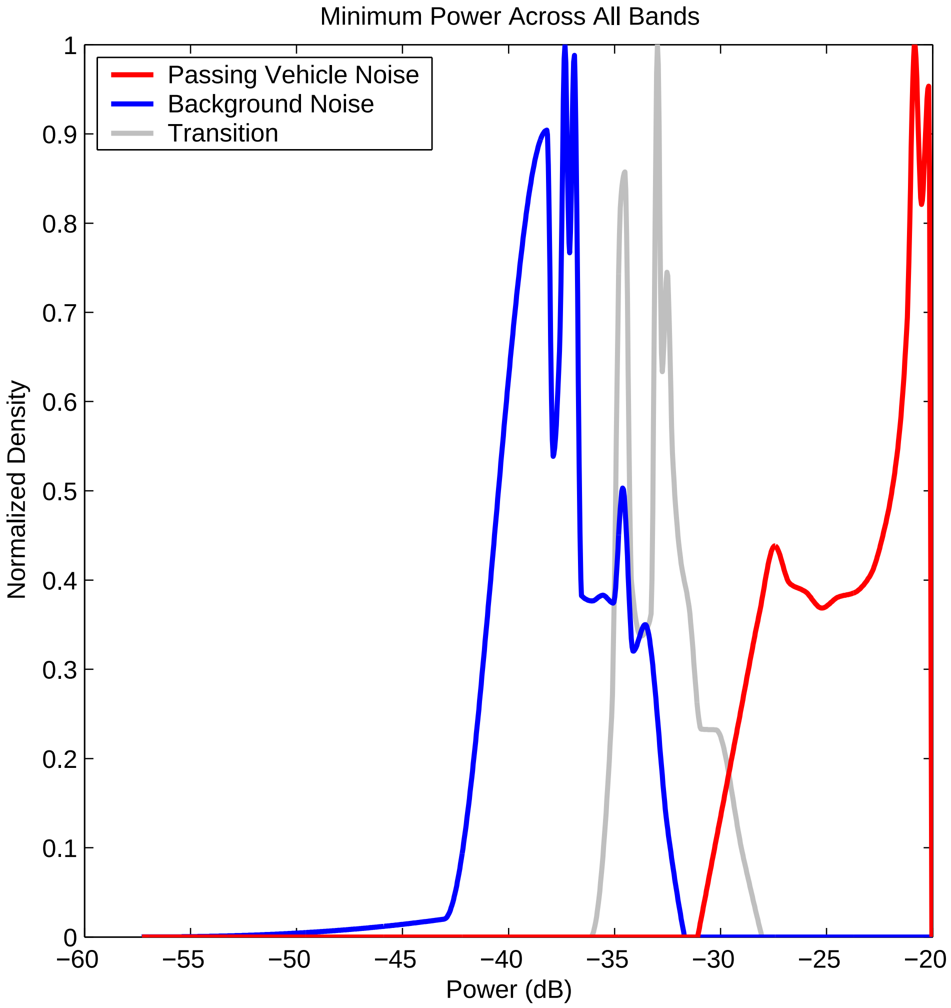

Figure 4.22: Conditional distribution of minimum power in any band.

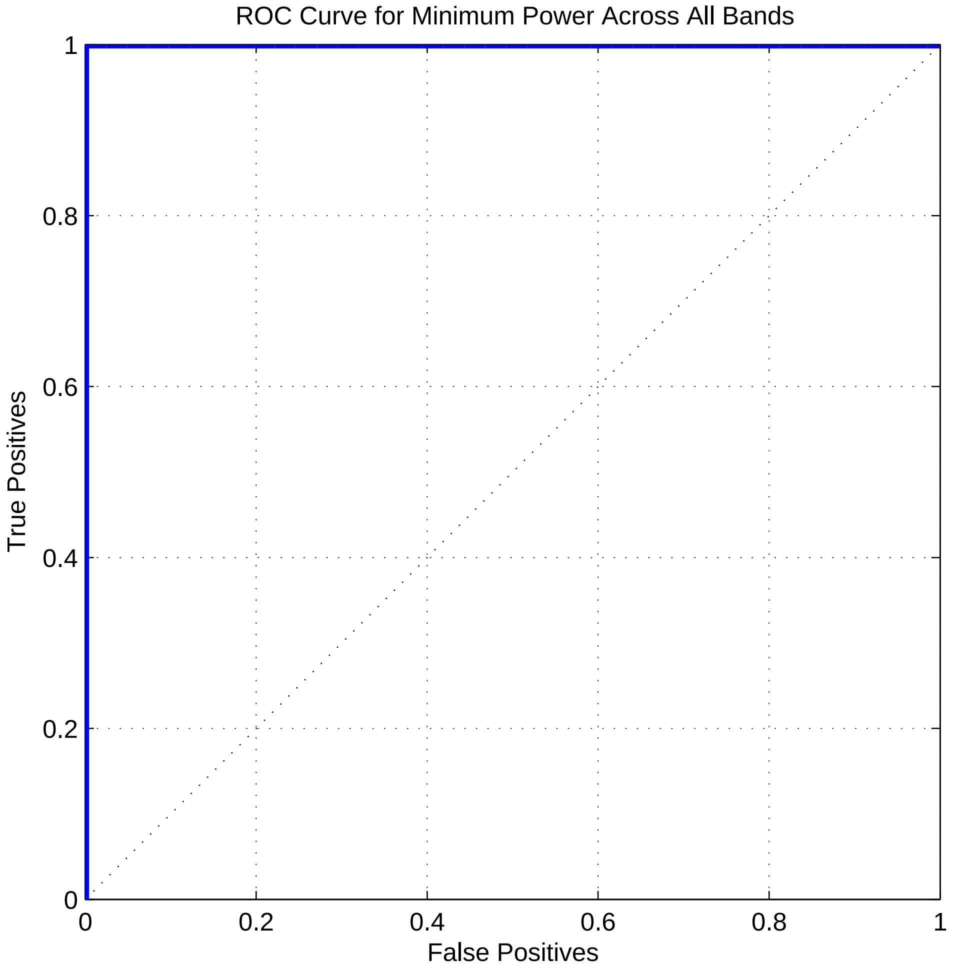

Figure 4.23: Receiver operating characteristic of the minimum power in any band.

a direct substitute could have been made, the bands were shifted slightly to accommodate using the output of an FFT more efficiently. Efficiency is obtained by choosing a power of two length and not having to interpolate frequency bins.

The STFT is applied using a periodic Hanning window with 50% overlap between windows. Only 16 bins are used when computing the STFT which is sufficient for generating 5 values over the frequency range of interest. Larger bin counts, up to 128, were also tested but did not yield the performance seen here. This is due to the reduction in averaging over frequency which provides increased variance and hurts separability, similar to what was seen in the power feature.

The STFT bin outputs resemble the filter bank outputs as can be seen in Figures 4.25, 4.26, and 4.27. The structure is also analogous to the filter bank as can be seen in Figure 4.24. It is notable that none of these features are perfect in the context of this test sample, unlike the filter bank implementation, but two are still quite good and the slight compromise in accuracy for the reduction in latency may be attractive in many applications.

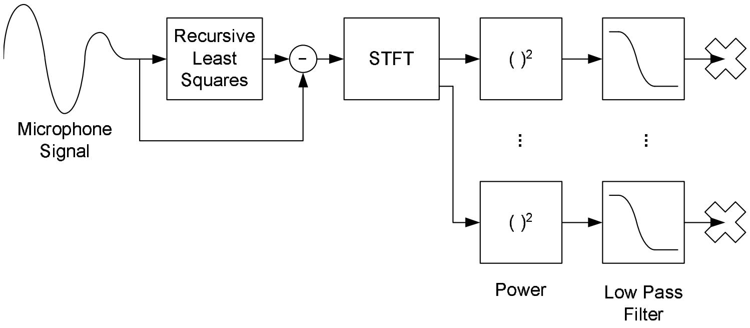

Figure 4.24: Structure of STFT feature.

4.2.6 STFT Minimum

In much the same way the subband feature of Section 4.2.3 was improved by applying a minimum in Section 4.2.4, the STFT feature can be improved with a minimum power across all bins as seen in Figure 4.28 and as described in Chapter 2. Such a minimum is illustrated

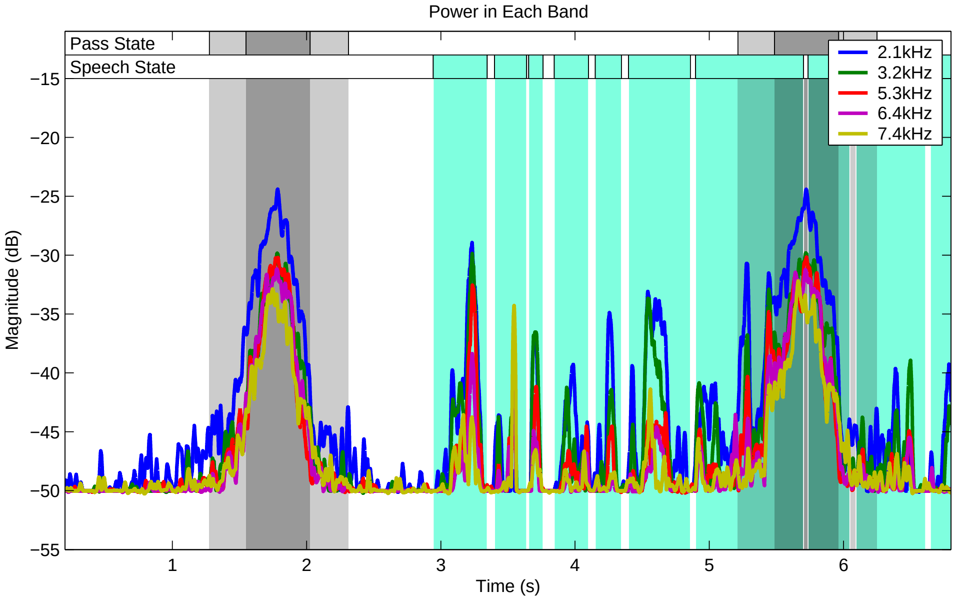

Figure 4.25: Power in different STFT bins.

in Figure 4.29. We can see from Figures 4.30 and 4.31 that we achieve near perfect detection on this test sample with this feature.

4.3 Classification

The final step in constructing a detector, according to the framework established in Chapter 2, is to design a classifier to transform the available feature values into a detection result. In this case we need to generate a detection result of passing vehicle noise present or passing vehicle noise not present. As stated in Section 2, the classifier simply defines decision regions in a space whose dimensionality is equal to the total dimensionality of the features under consideration. In this case our feature development has led to two implementations of only one unidimensional feature.

While there are a variety of schemes for designing decision regions in two or more dimensions, this one dimensional case simplifies to dividing a one dimensional space, or a 'number line' into two classes: passing noise present or not present. This situation is further simplified

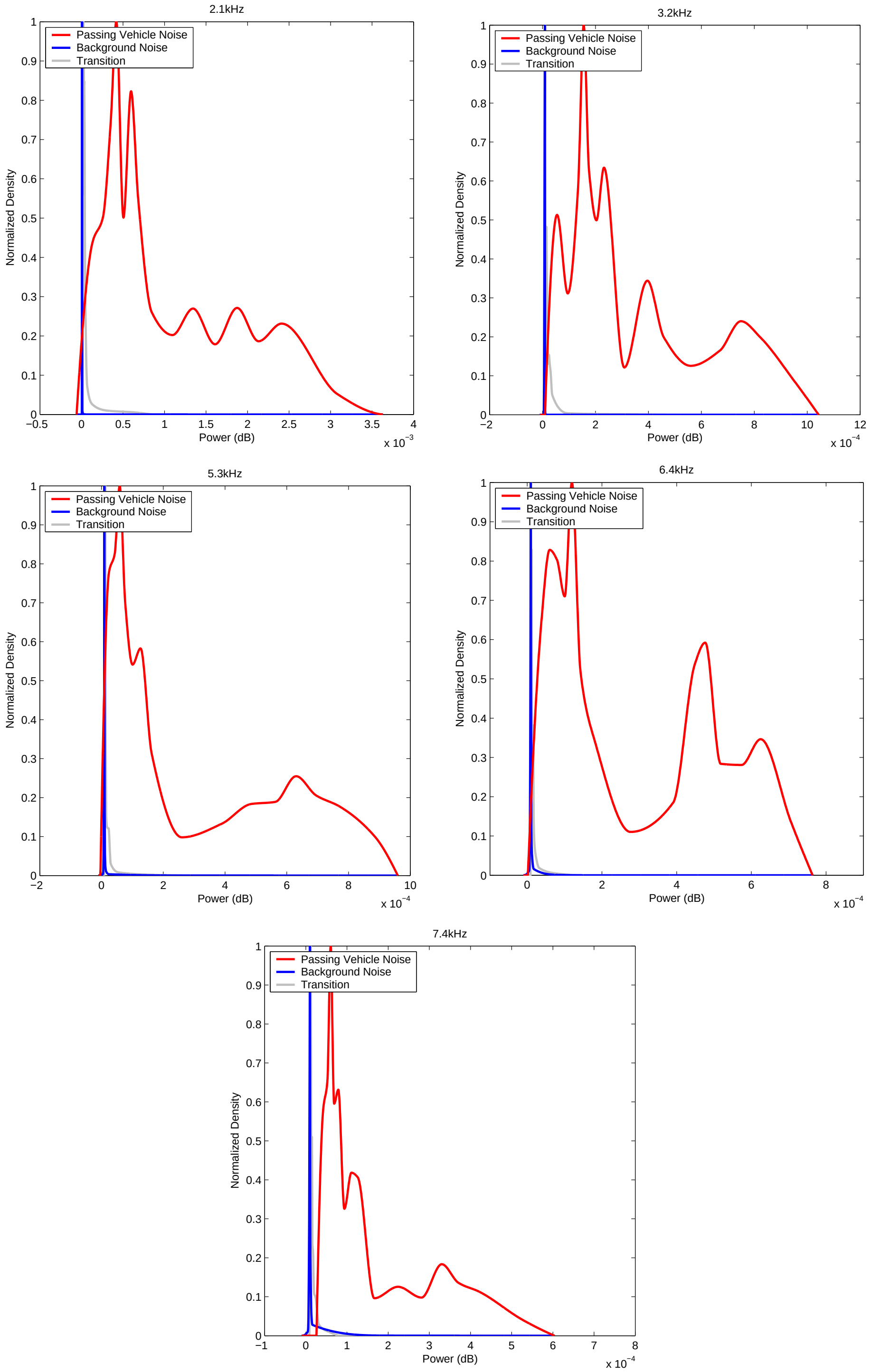

Figure 4.26: Power in different STFT bins.

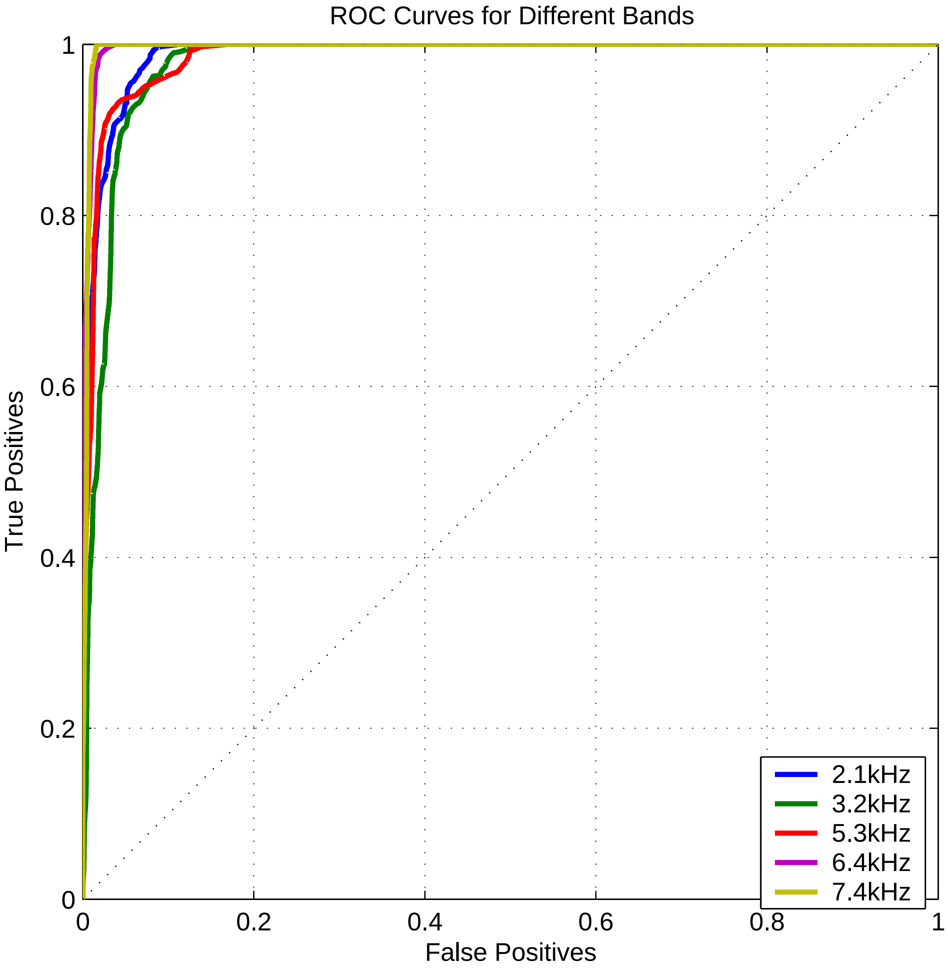

Figure 4.27: Receiver operating characteristic of the power in different STFT bins.

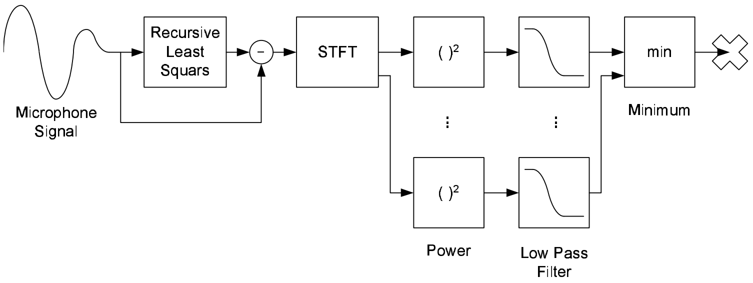

Figure 4.28: Structure of STFT minimum feature.

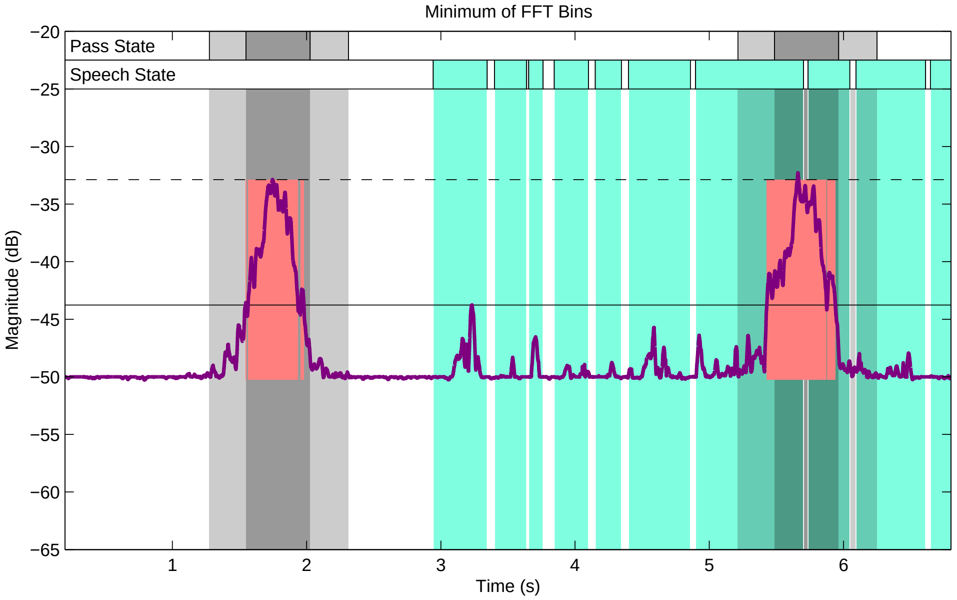

Figure 4.29: Minimum power in any STFT bin over time.

by the fact that the mode or modes present in each conditional distribution are not in any way interleaved as can be seen in Figure 4.22. This allows a single division of the space to be made. This is a conventional threshold-based detection scheme and its result, when applied to the specific test sample shown in Figure 4.1, is clearly shown in Figures 4.23 and 4.31. To revisit an earlier evaluation in this context: with the choice of the correct threshold perfect, detection accuracy can be achieved on this test sample using a simple threshold-based detection scheme when the 'Subband Minimum Power' feature is employed.

The high accuracy of this detector makes it a good choice for two applications in the automobile, assuming it generalizes to other test samples well. It can be used to activate noise mitigation measures only when the noise is present, helping to reduce potential speech distortion. Automatic speech recognition can also benefit by adjusting confidence or selecting an alternate profile to be used when passing vehicle noise is present as indicated by the detector output.

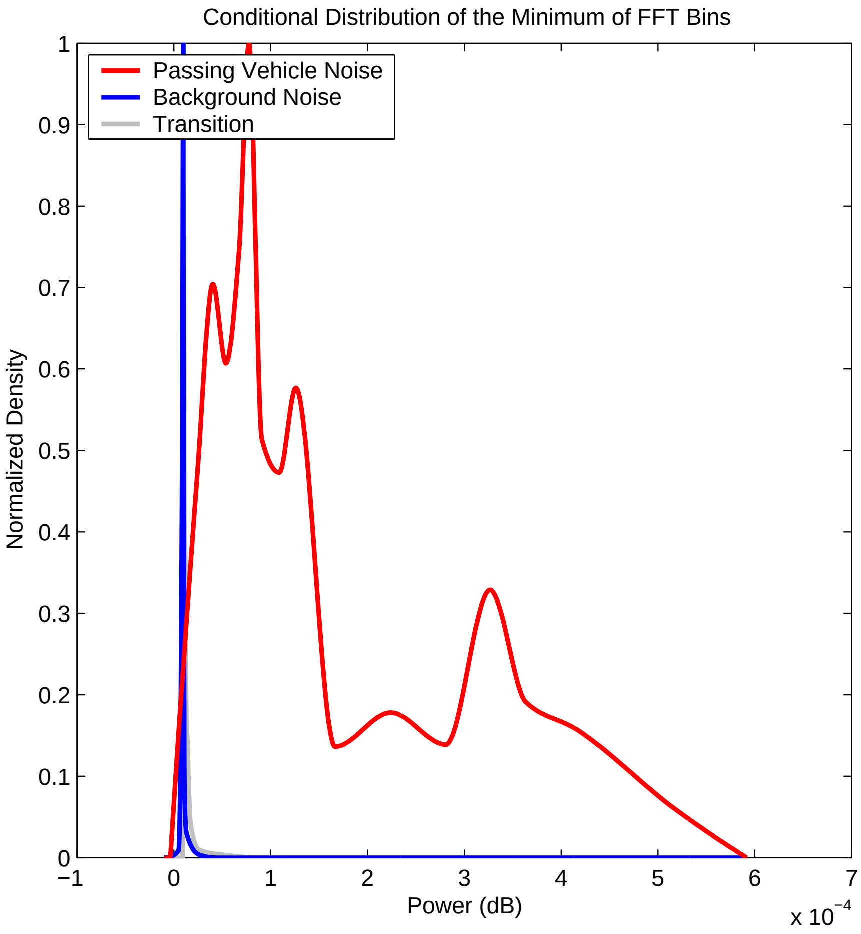

Figure 4.30: Conditional distribution of minimum power in any STFT bin.

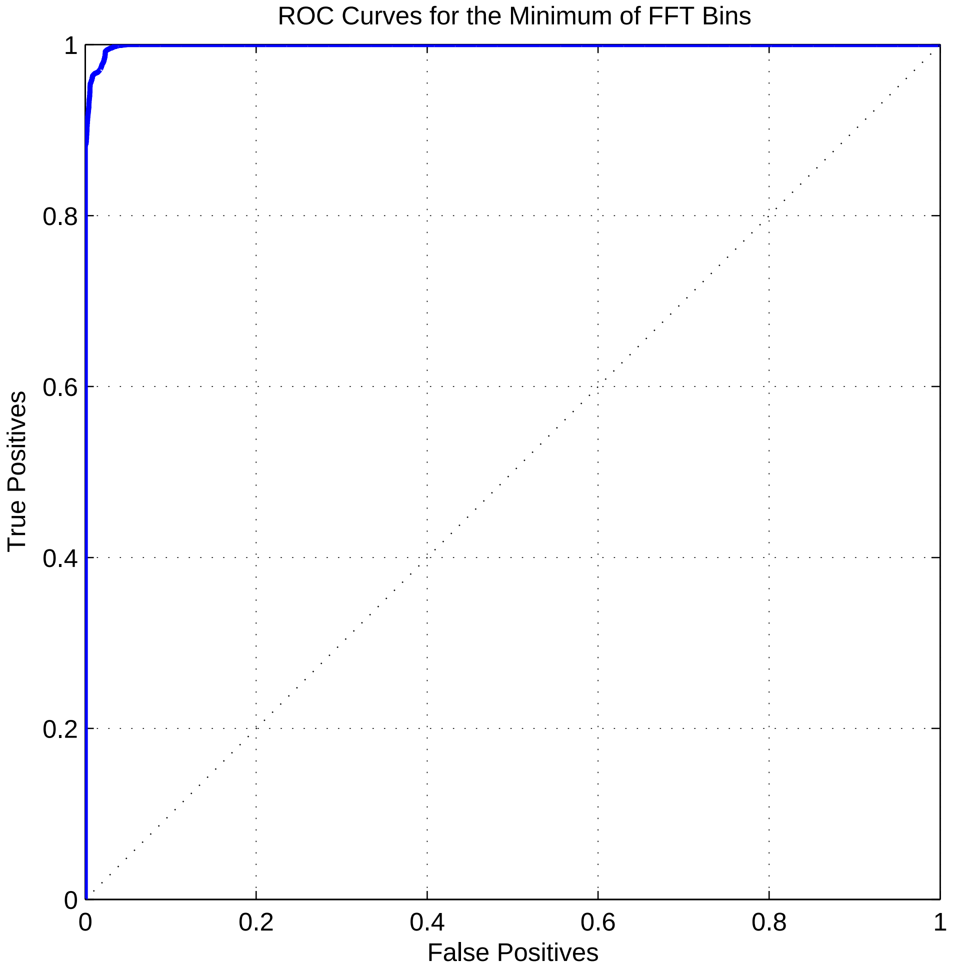

Figure 4.31: Receiver operating characteristic of the minimum power in any STFT bin.

Chapter 5

Speech Detection in the Presence of Passing Car Noise

The second detection problem addressed in this thesis is that of detecting whether speech is present during passing vehicle events or constructing a passing vehicle noise tolerant voice activity detector (PVNT-VAD). This is another problem of detecting the presence of one nonstationary signal in the presence of another nonstationary signal. This problem is different as far as which signal is our 'desired' signal, or the signal we wish to detect, and which signal is our nuisance nonstationary signal. Due to the presence of a third, fairly stationary component, we cannot simply invert the solution to the previous passing vehicle noise detection problem. This problem also suffers from the fact that speech does not share the relatively simple character of the passing vehicle noise, as described in Section 3.2, which was exploited in the previous solution.

The solution to the passing vehicle noise detection problem depended on predictive filtering. The predictive filtering used in the prior solution is not applicable in this detection problem. Figure 4.10 demonstrates that while this technique works well for finding passing vehicle noise, the speech is attenuated and more importantly the "pass mixed with speech" and the "pass without speech" events do not appear substantially different. The lack of applicability of the techniques in the passing vehicle noise detection problem is confirmed by reviewing Figure 4.18 where we see that while speech by itself might be easily differentiable from passing noise, the speech is again lost in the pass when mixed, regardless of examining specific bands. This review motivates a "fresh start" where new techniques are used but under the same methodology. The notable exception to the previous methodology is the deemphasis of time series plots since we will be focusing on short, discontinuous periods of time. The classification framework of Chapter 2 is applied again, beginning with feature creation and concluding with the development of a classifier. First,however, a test sample to be used for evaluation of both is defined.

5.1 Test Sample

An appropriate test sample is required to evaluate any scheme proposed for this detection problem. Like the test sample in the previous detection problem, a single test sample is used throughout the section. It is, again, composed of separate pass and speech samples to allow for knowing the true state of both and hence what the desired detector output is. This sample is composed of three passes in silence and three passes mixed with speech. This provides more data for evaluation of the detector output since we will only be evaluating the detector *during* passes. A spectrogram of this test sample can be seen in Figure 5.1.

5.2 Features

This section develops two features to be used for the detection of speech during passing vehicle events. An outline of this development is provided in Figure 5.2. Each feature begins with a general description, optionally followed by a detailed explanation and concludes with an evaluation using plots of conditional distribution of feature values and a receiver operating characteristic for the feature.

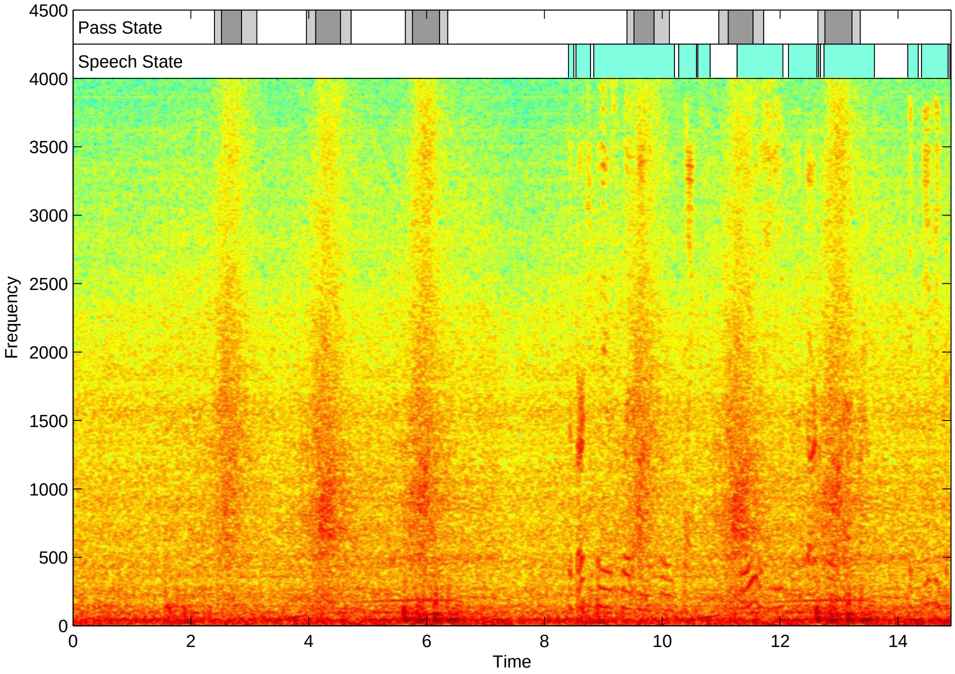

Figure 5.1: Spectrogram of test sample for PVNT-VAD development.

5.2.1 Power

The first feature examined in the context of the PVNT-VAD is signal power. While the results in previous chapter indicate that it may not be a terribly effective transformation by itself, it is nevertheless a transformation we wish to include and its effects should be noted. As a fairly good case for not employing instantaneous power has already been made, time-averaged power is applied immediately as seen in Figure 5.3.

The application of power to this detection problem is seen in Figure 5.4. The plots here follow the same general format as the previous chapter. The conditional distributions contain two curves. The red curve represents the distribution of feature values when only passing vehicle noise is present. The green curve represents the distribution of feature values when both passing vehicle noise and speech are present simultaneously. ROC curves are again generated from the conditional distributions, with the caveat that the thresholds are chosen and evaluated against only the test sample shown in Figure 5.1.

It is apparent from both the conditional distributions and the ROC curves that power is

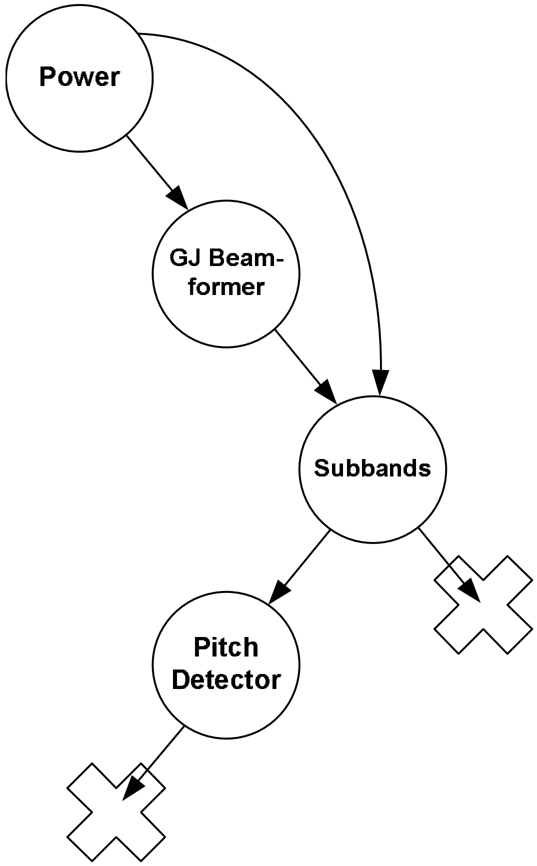

Figure 5.2: An outline of the development of and relationship between features leading up to the passing vehicle noise tolerant voice activity detector.

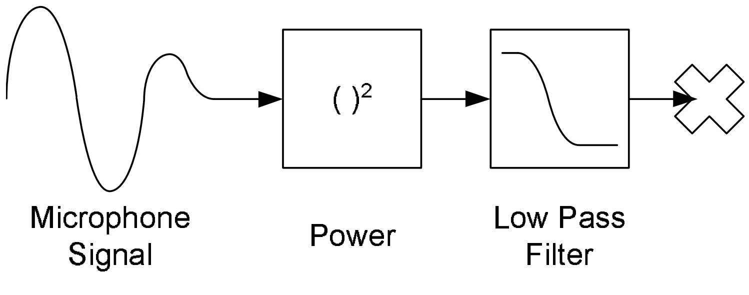

Figure 5.3: Structure of power feature.

barely able to differentiate whether or not speech was present during the pass. The speech needs to be made more powerful than the pass via preprocessing.

5.2.2 Beamforming

A beamformer amplifies some sources and attenuates others based on the geometry of the sources and sensors as described in Chapter 2. The beamformer to be used for this feature is the Griffiths-Jim adaptive beamformer (G-J BF). The direction of amplification, or look direction, is fixed and the direction of attenuation is adapted. The details for this algorithm can be found in [JD93].

As stated, the application of this technique requires observations from multiple sensors and geometric information. As described in Section 3, the multiple sensors requirement, which in the case of acoustic signals are microphones, is met as well as the requirement of knowing their geometry. Although speech is not statistically stationary, the speaker is spatially stationary relative to the sensors and is perpendicular to the array. The passing vehicle noise is moving relative to the array, so its attenuation requires the adaptation made available by the algorithm.

The results of applying this algorithm in the manner shown in Figure 5.5 to the test signal are shown in Figures 5.6 and 5.7. The ROC curve reveals that this feature is neither uniformly better nor uniformly worse than the power of the unprocessed signal. Overall power is not improved but power in specific bands can be examined.

5.2.3 Subband Power

While speech is clearly not more powerful than the passing vehicle noise over all frequencies, speech may be more powerful in certain frequency ranges. Speech can be split into two categories with different spectral characteristics: voiced and unvoiced. Voiced speech is most powerful over a range of relatively low frequencies (<1kHz). Power in unvoiced speech is spread over much of our available spectrum.

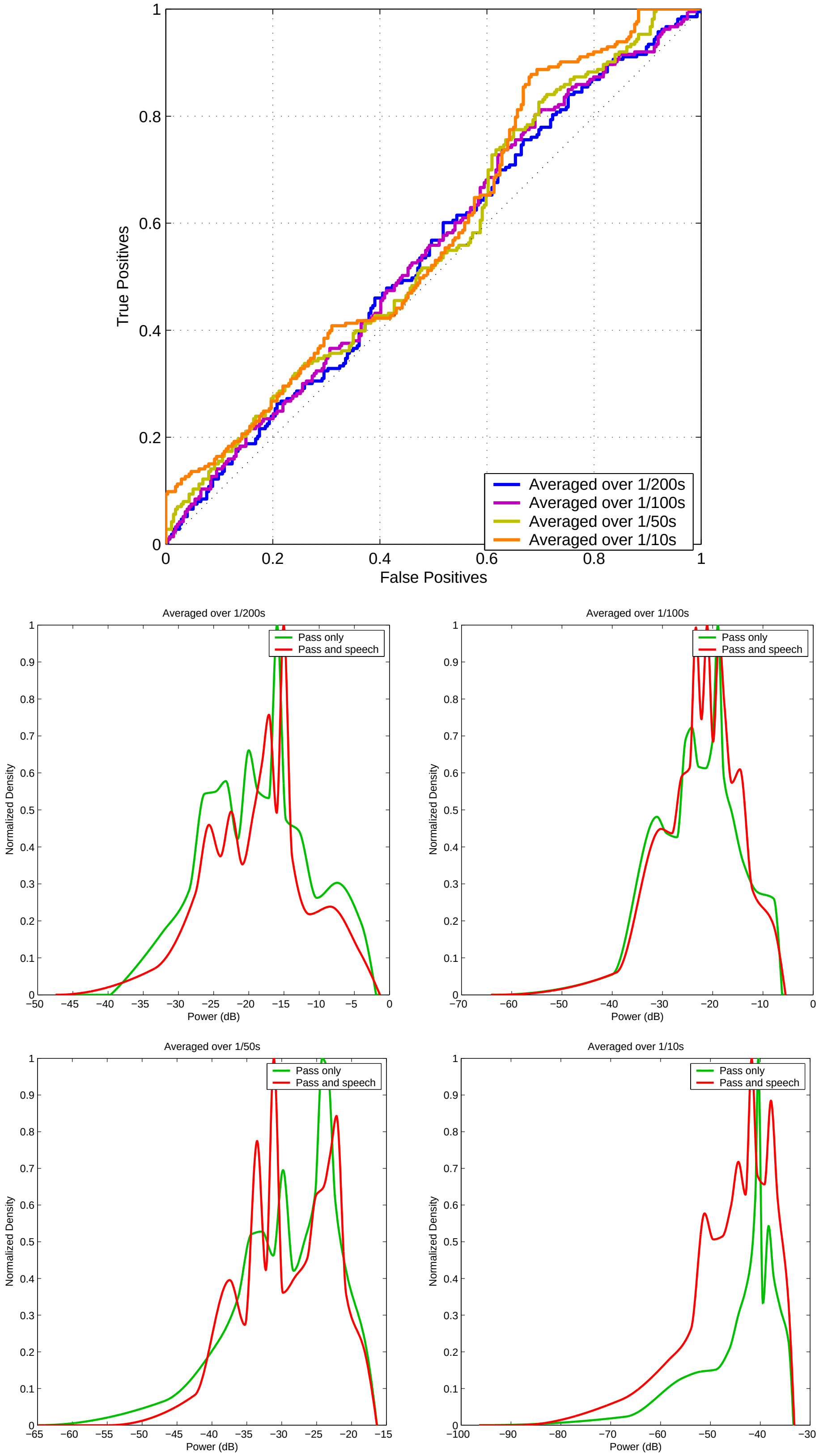

Figure 5.4: Bandpass filter power features.

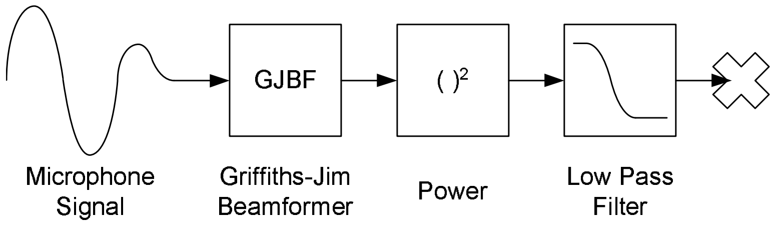

Figure 5.5: Structure of Griffiths-Jim beamformer feature.

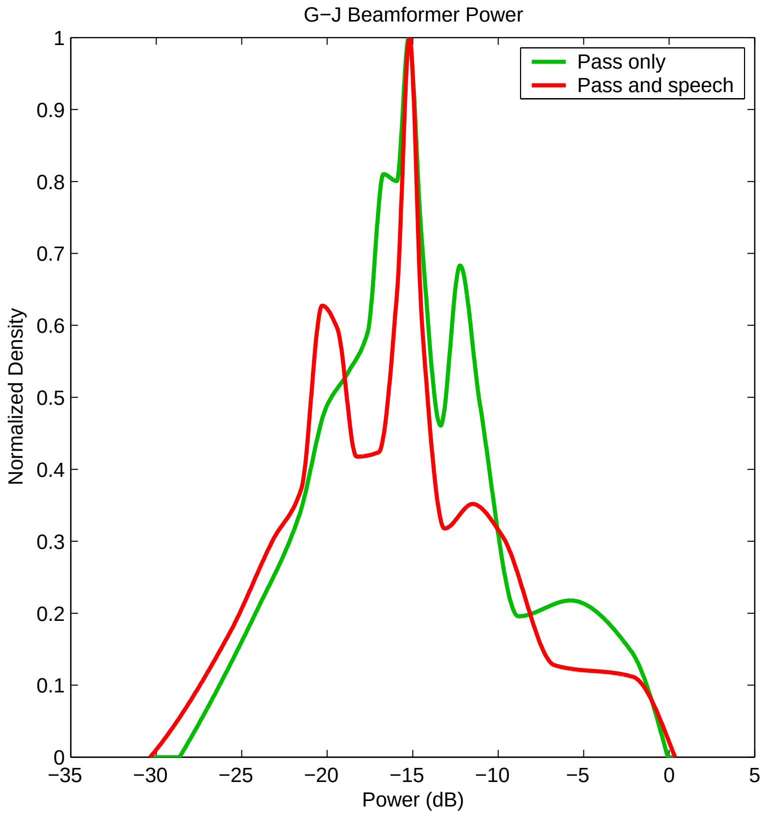

Figure 5.6: Conditional distribution of power after G-J BF.

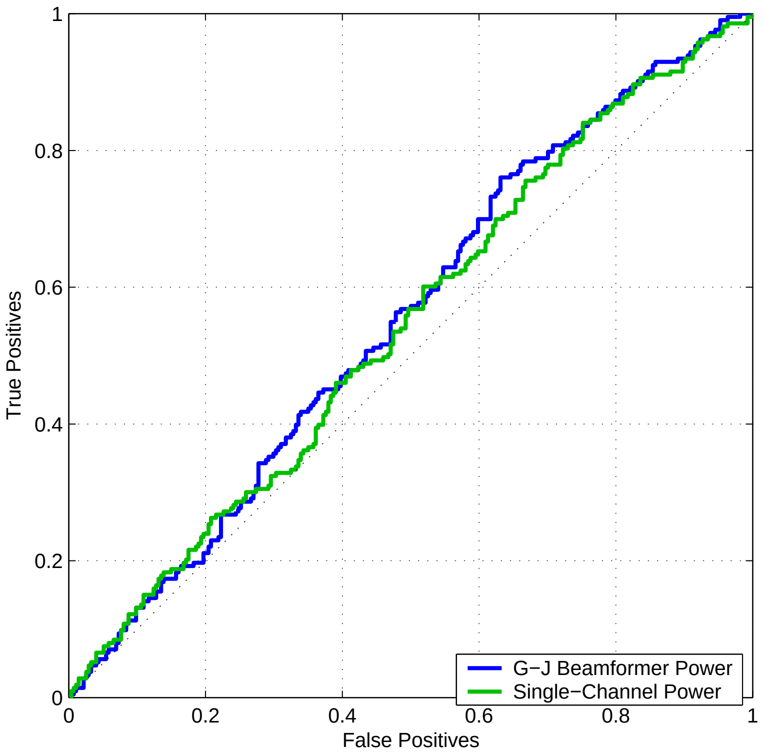

Figure 5.7: ROC of power after G-J BF.

Two bandpass filters, one for each type of speech, can be applied to either the original signal or the signal from the beamformer to focus on frequencies where speech is more powerful than the passing vehicle noise as shown in Figure 5.8. The voiced speech pass band is 320Hz-500Hz and the unvoiced pass band is 260Hz-3500Hz. These were determined empirically by testing a wide range of values for both upper and lower cutoff frequencies and determining which combination was most statistically favorable via ROC curves when applied to this test sample.

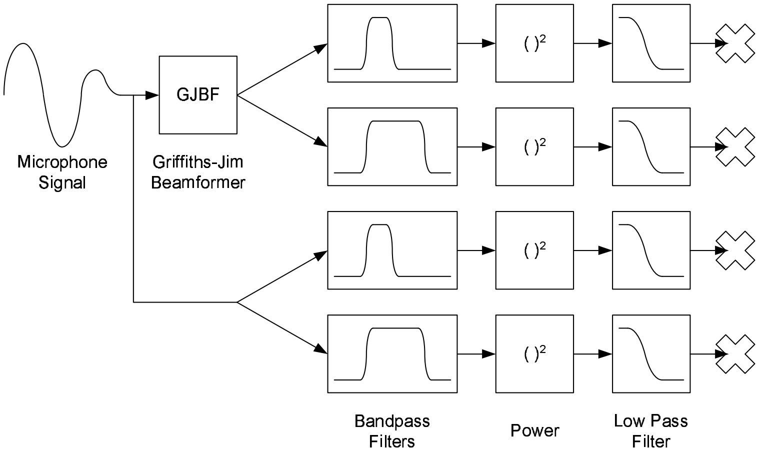

Figure 5.8: Structure of bandpass power feature.

The results of applying these filters to both the original signal and the signal from the G-J BF are shown in Figure 5.9. This improves accuracy greatly and also demonstrates that substantial gains are made by employing the G-J BF. While the band edges exploit some knowledge of the voiced speech, the harmonic structure present in the voiced speech has not yet been exploited.

5.2.4 Pitch Prediction

Pitch prediction is a form of predictive filtering based on estimating the pitch of voiced speech. This feature makes use of the estimate of the pitch rather than the prediction of the signal. The pitch estimate is bounded by the reasonable range of the pitch of a human

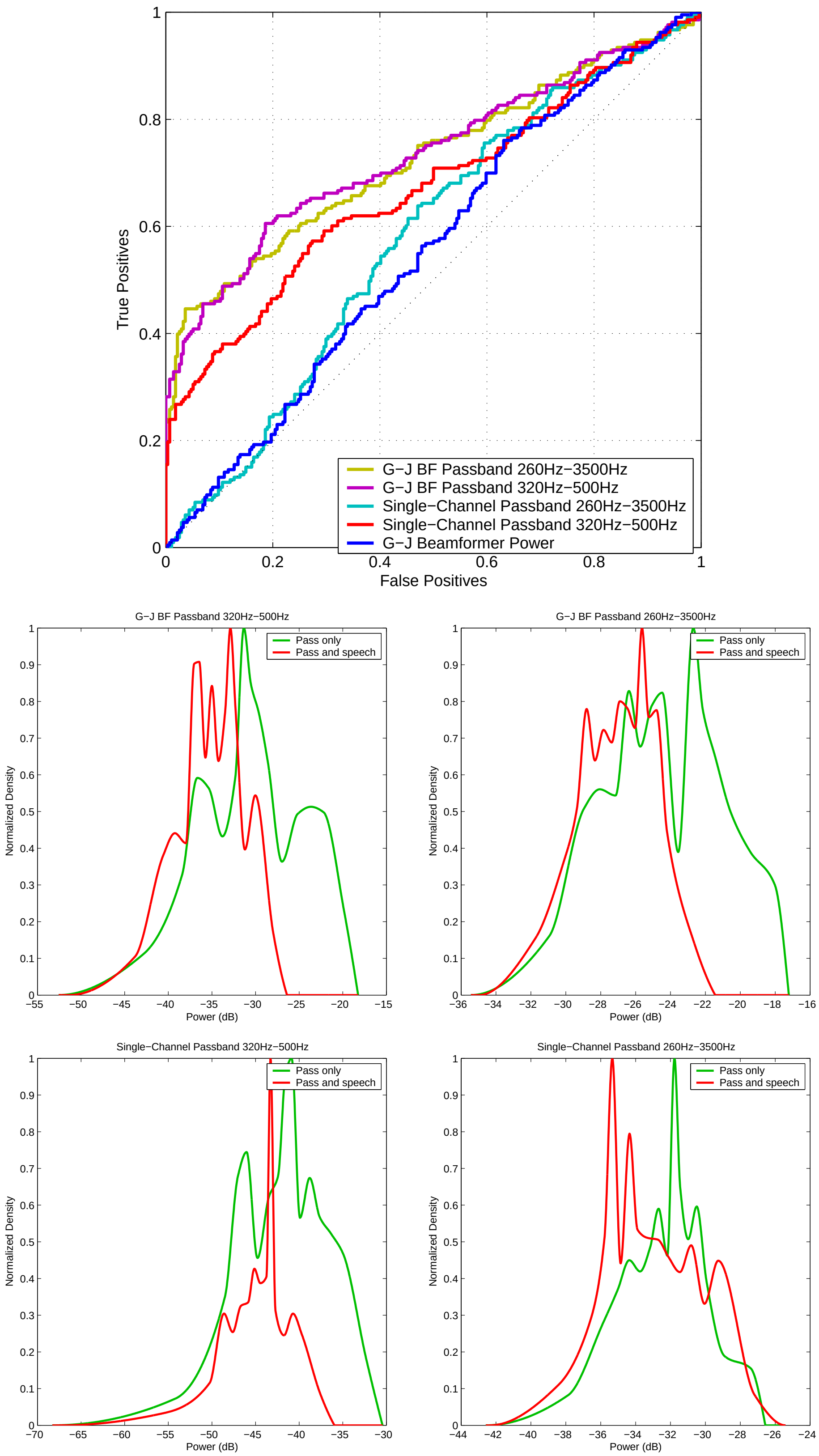

Figure 5.9: Bandpass filter power features.

voice.

A simple one tap pitch estimate is employed. This estimator calculates the autocorrelation of two consecutive 40ms frames. The estimate of frequency or pitch is taken as the inverse of the lag value that maximizes correlation subject to the range of period, and hence lag, defined by the human voice. The amplitude of the voiced speech is estimated via the actual value of the autocorrelation at that lag. This correlation value is used as the voiced speech feature.

This pitch detection scheme is applied to the voiced speech band as shown in Figure 5.10. The conditional distribution and ROC curve for the voiced speech feature, when applied to the test sample, are shown in Figures 5.11 and 5.12. Accuracy is similar to the voiced speech power feature at higher false positive values, however accuracy at lower false positive values is improved markedly. While accuracy is better than previous features, it is still far from optimal. Performance can, however, be further improved by exploiting diversity between the features.

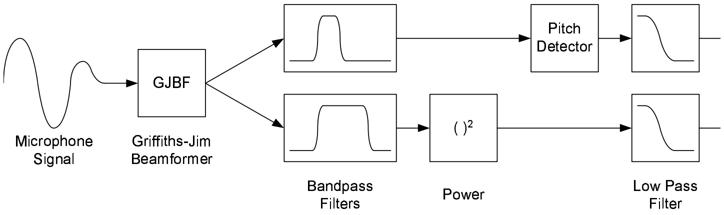

Figure 5.10: Structure of pitch feature.

5.3 Classifiers

The final step in constructing the passing vehicle noise tolerant voice activity detector is selecting a classifier to map one or more of the features above to a detection result. This detection result indicates whether speech is present during a passing vehicle event. Since feature development concluded with two terminal features, the classifier maps a 2-D feature

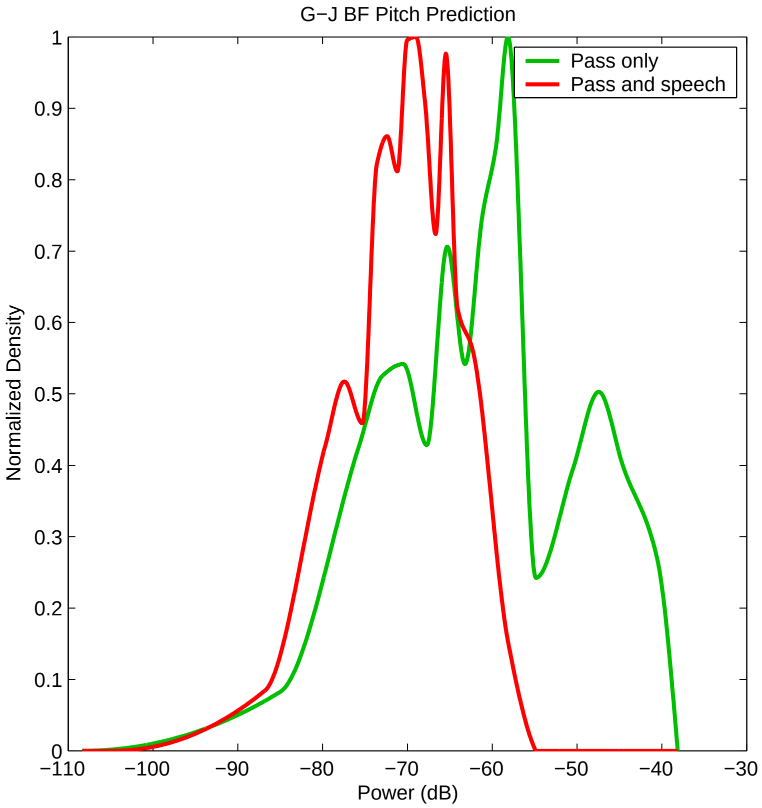

Figure 5.11: Conditional distribution of estimated pitch power feature.

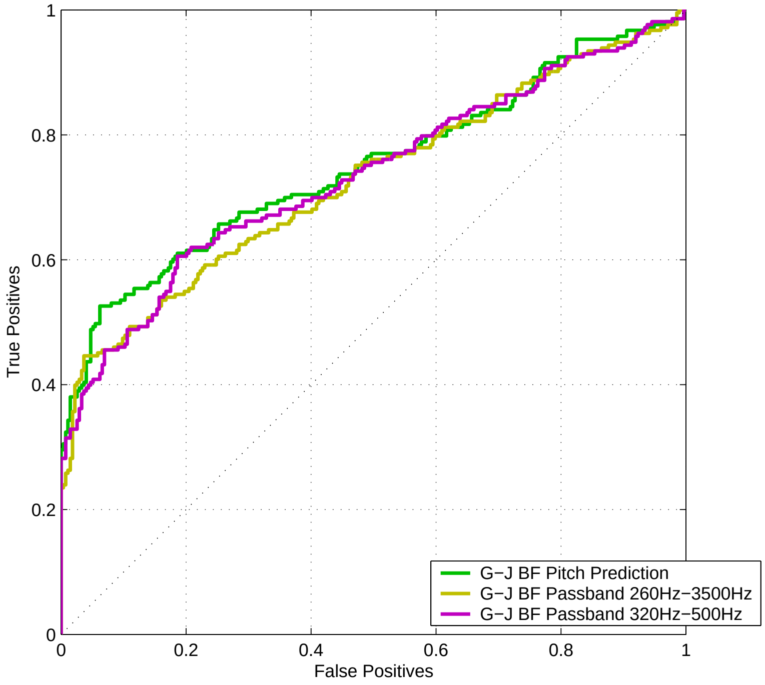

Figure 5.12: ROC of estimated pitch power feature.

vector to a 1-D result space. In this case, the concept of the threshold is broadened to the idea of the decision region in a 2-D space.

Two basic techniques from classification theory are applied to this problem. Both produce mappings from a 2-D feature vector to a 1-D score space and are introduced in Chapter 2. A threshold applied to this score space produces a classification or detection result. This threshold mapped back into the 2-D feature space defines the borders between decision regions.

5.3.1 Linear Discriminant Analysis

The linear discriminant is concerned with finding the optimum linear combination of the elements of the feature vector. Optimum in this case is the result of a likelihood ratio test (LRT) assuming the distributions are normal. In the 2-D case at hand this produces a mapping where the score contours are straight lines. Thus selecting a score threshold selects a straight line parallel to these contours as the decision boundary.

The linear discriminant must be optimized on a different sample than it is applied to in order to generate meaningful results. To this end, another sample composed of different portions of the same speech and noise samples used to generate Figure 4.1 was used for finding the optimal combining coefficients using linear disrcriminant analysis. The original test sample shown in Figure 4.1 was then evaluated using these coefficients.

A scatter plot of the 2-D feature data along with labeled score contours is shown in Figure 5.13. The corresponding conditional distributions of score are shown in Figure 5.14 and an ROC curve based on score is shown in Figure 5.15.

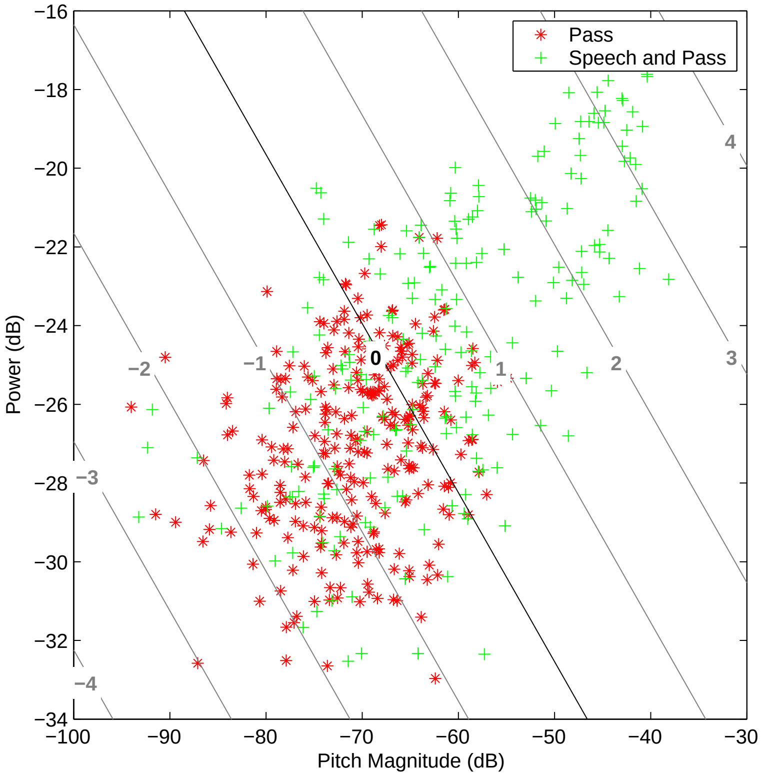

Figure 5.13: Scatter plot of feature data with linear discriminant score contours.

The scatter plot clearly shows that perfect separation will not be achieved with these features due to the substantial overlap of some of the speech over most of the pass. The ROC curve does, however, demonstrate an marked improvement made by combining both

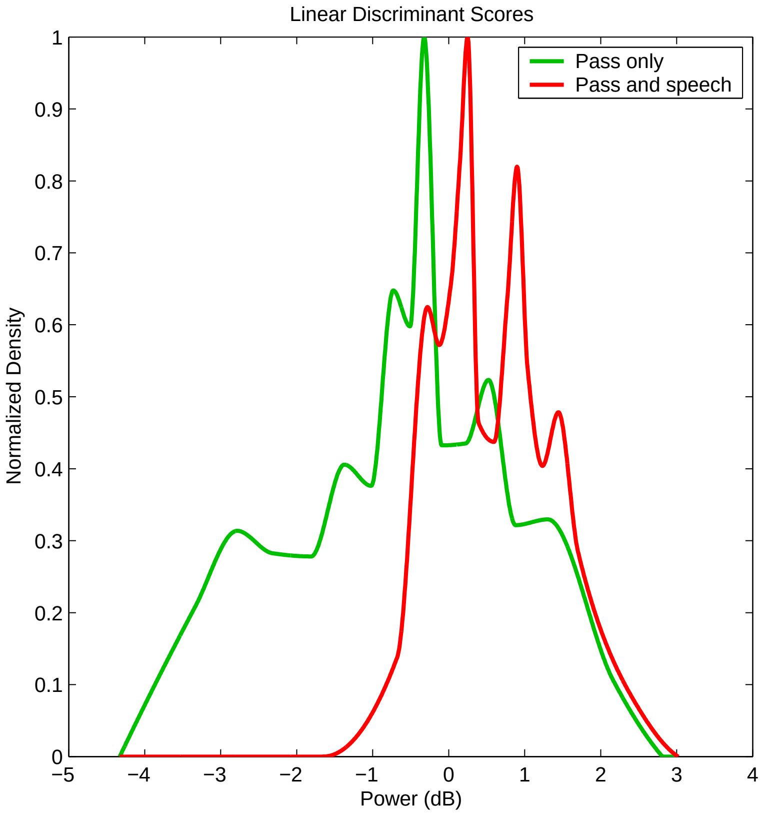

Figure 5.14: Conditional distribution of linear discriminant score.

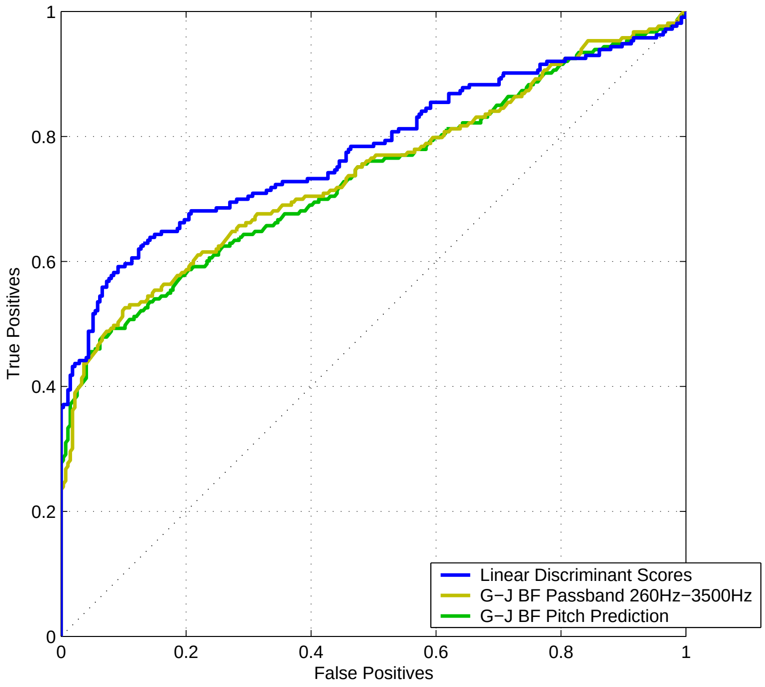

Figure 5.15: ROC of linear discriminant score.

features using the linear discriminant instead of using only the better of the two.

5.3.2 Quadratic Discriminant Analysis

The quadratic discriminant is concerned with finding the optimum quadratic combination of the elements of the feature vector. Like the linear discriminant it also finds an optimum assuming normal distributions but, given the quadratic form, allows for circular or hyperbolic decision regions. In the 2-D case the coefficients of each term of an equation of the form $y^2 + xy + x^2 + x + y = c$ are determined based on an LRT of two normals.

Like the linear discriminant, the quadratic discriminant must be optimized on a different sample than it is applied to in order to generate meaningful results. The same alternate sample was used to generate the quadratic combining coefficients. The original test sample shown in Figure 4.1 was then evaluated using these coefficients.

A scatter plot of the feature is again shown but with the quadratic score contours overlayed in Figure 5.16. The corresponding conditional distributions of score are shown in

Figure 5.17 and an ROC curve based on score is shown in Figure 5.18. We observe that the quadratic discriminant is close to uniformly better than the linear discriminant. The gains it provides are, however, rather small.

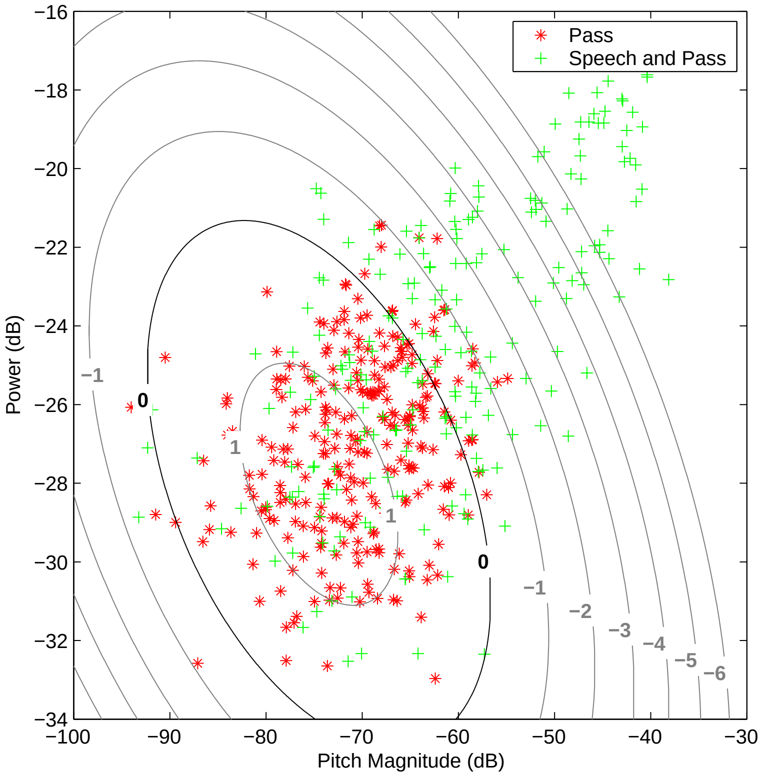

Figure 5.16: Scatter plot of feature data with quadratic discriminant score contours.

5.4 Comparison

In the case of a voice activity detector, previous well-known algorithms exist for comparison of accuracy. G.729B (an ITU standard) is a popular voice activity detector used in telephony. A comparison with this off-the-shelf algorithm demonstrates how much an automotive application stands to benefit from this custom algorithm.

A reference implementation of G.729B, including the VAD, is available from the ITU in C. This C code was instrumented to allow of the extraction of the VAD state for each frame. Both the original signal and the output of the G-J BF were then run through the

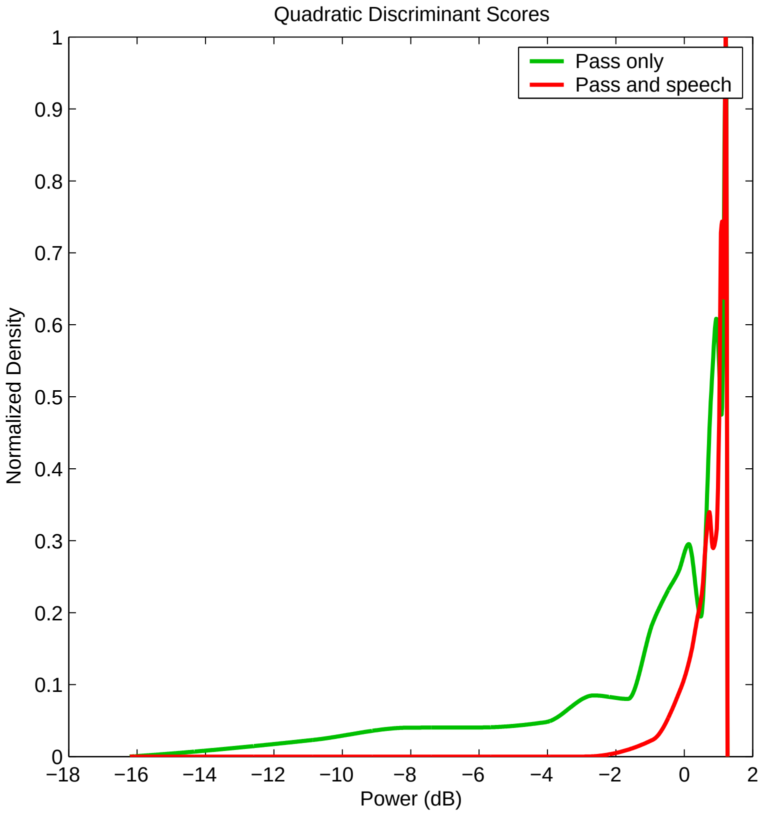

Figure 5.17: Conditional distribution of quadratic discriminant score.

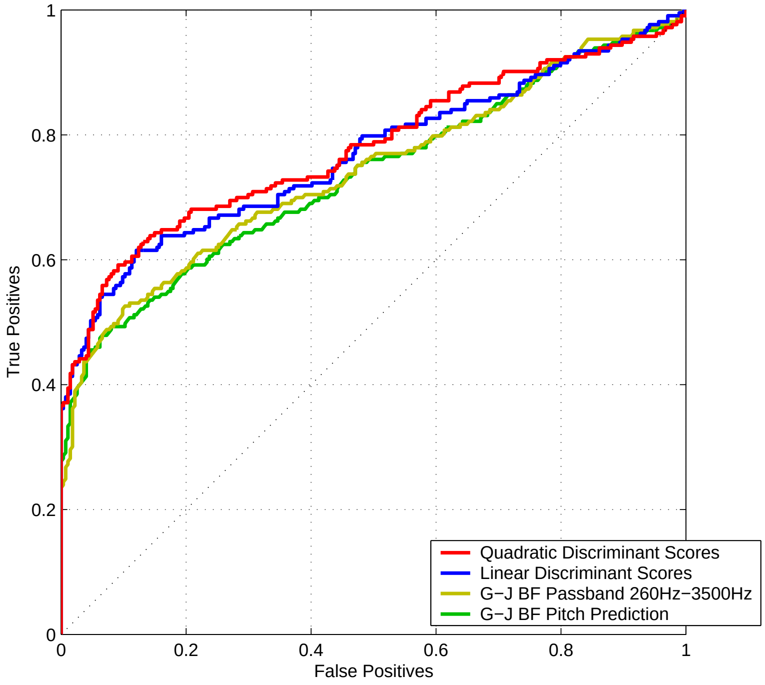

Figure 5.18: ROC of quadratic discriminant score.

algorithm. The two accuracy measurements generated from applying the G.729B VAD to the test sample appear in Figure 5.19 against a plot of ROC curves shown previously. We can see from this plot that the performance of the off the shelf algorithm is roughly equivalent when applied to the test sample but much lower levels of false positives are made available by this custom algorithm.

The final detector offers much greater performance for low false positive rates than conventional techniques, at least as suggested by its application to this test sample. While there is still room for improvement, applications that depend on voice activity detection can benefit from this gain. Such applications include automatic speech recognition and discontinuous transmission (DTX) of conversational audio. This result can also be combined with the passing vehicle noise detector to make better decisions about mitigation, based not only on the presence of the passing vehicle noise but also on whether speech is present simultaneously.

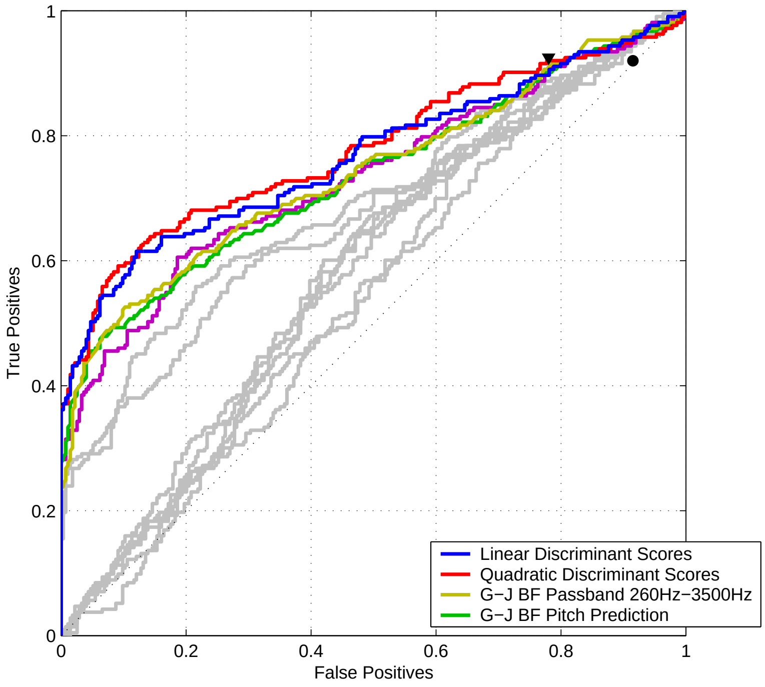

Figure 5.19: ROC curves of all feature data and operating points for the G.729B VAD.

Chapter 6

Conclusions and Future Research

Four major outcomes are provided by this thesis. First, a library of multichannel recordings was produced. Second, a characterization of passing vehicle noise was developed resulting in a model and simulation technique. Third, a highly effective detector for passing vehicle noise was developed. Fourth, a voice activity detector was developed that substantially out performs off-the-shelf algorithms when speech is corrupted by passing vehicle events.

A new set of recordings were made both out of a need for data for this project and the desire to provide a generally useful dataset for future work. This data comes from four different tests: driving in traffic, accelerating and coasting in isolation, parked in traffic, and speech in isolation. Each of these tests were performed four times to cover all combinations of two vehicles and windows up and down. An accompanying video was created for one vehicle to assist in identification of noise sources.

The characterization and detection efforts in this thesis produced three important outcomes. First, a basic characterization of passes from experimental data was performed and verified through listening tests of derived simulations. Second, an extremely effective detection scheme was developed and then refined in terms of processing delay. Third, a voice activity detector for use during passing vehicle events was developed. While the accuracy of this detector leaves plenty of room for improvement, it represents a significant improvement over the existing G.729B VAD, at least when applied to the chosen test sample.

Together, these two detectors allow for several improvements in hands-free speech acquisition in the automobile if the performance seen using the test sample generalizes well. Noise mitigation can be applied more appropriately by knowing when passing vehicle noise is present and whether speech is also present. Automatic speech recognition can incorporate appropriate confidence adjustments when passing vehicle noise is present with speech as well as benefit from fewer false voice activity detections when only passing vehicle noise is present. Lastly, the voice activity detection used in the mobile phone can be improved so that bandwidth reduction from discontinuous transmission can be better realized in this noisy environment.

While the results in Chapters 4 and 5 demonstrate the potential for impressive performance, it is important to remember that these results were obtained, using a single test sample. To confirm performance claims about either detector, future work should involve further testing. Cross-validation of any optimal parameter would provide a more general choice of value and a better measure of performance. Introduction of samples from other vehicles and other speakers would also serve to further generalize the performance results.

While these detectors represent a substantial improvement in performance over existing techniques, two specific further improvements could also be realized. Future work might focus on either of these two reasonably attainable targets, both of which fit into the framework defined in this thesis. First using classification to define a four state output, consisting of:

- no passing vehicle noise or speech
- speech but no passing vehicle noise
- passing vehicle noise but no speech
- and both passing vehicle noise and speech

could provide substantially better performance than cascading the two detectors by avoiding the compounding of error. Currently detecting whether speech is present in a pass depends on correctly detecting whether a pass is present. Therefore the accurate detection of speech during a pass is subject to the error of detecting the pass correctly as well as the error of determining whether speech is present.

Second, a promising classification technique is identified in [BFOS84] whose implementation was outside the scope of this project. The classification tree procedure allows for virtually any type of feature and possesses various desirable properties not least of which is guaranteeing equal or better classification when new features are added. This technique also provides insight into the data by producing human readable decision trees with associated statistics as well as more advanced measures such as stability via cross validation. Either of these techniques could potentially improve the results obtained by the work in this thesis.

Bibliography

[AK05] Sungjoo Ahn and Hanseok Ko. Background noise reduction via dual-channel scheme for speech recognition in vehicular environment. *IEEE Transactions on Consumer Elect*, 51(1):22–27, February 2005.

[BFOS84] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees.* Wadsworth International Group, Belmont, CA, 1984.

[BP88] Julius S. Bendat and Allan G. Piersol. *Measurement and Analysis of Random Data.* John Wiley and Sons, Inc., New York, London, Sydney, 1988.

[Bro04] Donald R. Brown. Removing passing car noise from an automotive hands-free phone. Final Project Report, 2004.

[CCC$^+$05] T. Chau, D. Chau, M. Casas, G. Berall, and D.J. Kenny. Investigating the stationarity of paediatric aspiration signals. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 13(1):99–105, March 2005.

[Coh04] Israel Cohen. Multichannel post-filter in nonstationary noise environments. *IEEE Transactions on Signal Processing*, 52(5):1149–1160, May 2004.

[Fan02] Jeffery Faneuff. Spatial, spectral, and perceptual nonlinear noise reduction for hands-free microphones in a car. Master's thesis, Worcester Polytechnic Institute, 2002.

[FGW94] S. Fisher, J. Goodall, and L. Wark. *Effects of New Technology in Cars on Driver Attitudes and Behavior.* University of Strathclyde, Glasgow, UK, 1994.

[GB$^+$97] M. Goodman, F. D. Bents, et al. *An Investigation of the Safety Implications of Wireless Communications in Vehicles.* National Highway Traffic Safety Administration, 1997.

[Han81] D. J. Hand. *Discrimination and Classification.* John Wiley and Sons, Chichester, 1981.

[Hay01] Simon Haykin. *Adaptive Filter Theory.* Prentice Hall, Upper Saddle River, NJ, 4 edition, 2001.

[Hop01] J.R. Hopgood. *Nonstationary Signal Processing with Application to Reverberation Cancellation in Acoustic Environments.* PhD thesis, Cambridge, UK, 2001.

[Jam85] Mike James. *Classification Algorithms.* John Wiley & Sons, New York, NY, 1985.

[JD93] Don H. Johnson and Dan E. Dudgeon. *Array Signal Processing: Concepts and Techniques.* Prentice Hall, Englewood Cliffs, NJ, 1993.

[Kee04] Kevin Keenaghan. A novel non-acoustic voiced speech sensor: Experimental results and characterizaton. Master's thesis, Worcester Polytechnic Institute, 2004.

[Lee02] S. J. Leese. Microphone arrays. In G. M. Davis, editor, *Noise Reduction in Speech Applications.* CRC Press, Boca Raton, FL, 2002.

[MPC04] Matthew Musiak, Michael Porcaro, and Jonathan Casey. Analysis of acoustic automobile noise, 2004.

[OVP92] Stephen Oh, Vishu Viswanathan, and Panos Papamichalis. Hands-free voice communication in an automobile with a microphone array. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1:281–284, March 1992.

[RZB97] D. C. Read, A. M. Zoubir, and B. Boashash. Aircraft flight parameter estimation based on passive acoustic techniques using the polynomial wignerville distribution. *Journal of the Acoustical Society of America*, 102, July 1997.

[Vas00] S. V. Vaseghi. *Advanced Digital Signal Processing and Noise Reduction*. John Wiley & Sons, New York, NY, 2 edition, 2000.

[WB95] David A. Whitney and Bruce Broder. Multi-scale signal feature processing for automatic, objective vehicle noise and vibration quality analysis. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 5:2959–2962, May 1995.

[WG92] R. B. Wallace and R. A. Goubran. Improved tracking adaptive noise canceler for nonstationary environments. *TSP*, 40(3):700–703, March 1992.

[WT95] W. W. Wierwille and L. Tijerina. An analysis of driving accident narratives as a means of determining problems caused by in-vehicle visual allocation and visual workload. In *Vision in Vehicles*. 1995.

Appendix A

Acquisition System Used by Punit Prakash

This appendix describes the acquisition system designed by Punit Prakash used to obtain passing vehicle recordings in a set of two automobiles. The system consisted of a laptop running Gold Wave, an eMagic 6|2m acquisition unit, a bias box, and six AU1000 microphones as shown in Figure 3.1. Four of the microphones were arranged in a broadside linear array attached to the driver's visor, one was attached to the rear view mirror and one was affixed to a boom on a head set which the driver wore. These attached to the bias box, which was constructed according to Figure A.2 and powered via the cigarette lighter socket in the vehicle. The bias circuit provides the required direct current *to* the microphones and allows only the alternating current provided *by* the microphones to be present at the output terminals. With the exception of the potentiometer this is the circuit recommended by the microphone manufacturer, Shure Communications. The six instances of the bias circuit in addition to the power supply circuit comprise the bias box. The bias box then attached to the eMagic unit which was connected to the laptop via USB.

This acquisition system exhibited three highly undesirable conditions:

- Dropped samples.

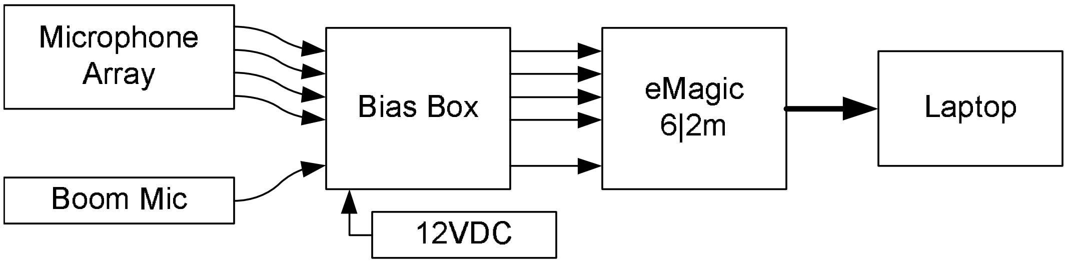

Figure A.1: Acquisition system.

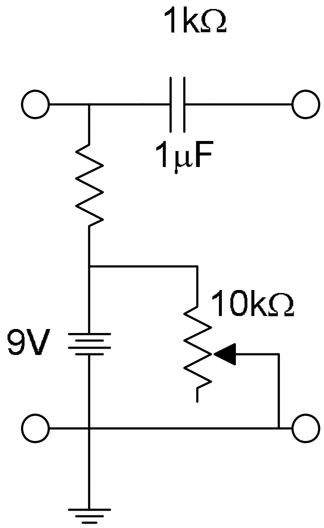

Figure A.2: Original bias circuit.

- Microphone saturation.

- Aliasing.

The first condition is that samples are dropped occasionally. Dropped samples made the recording system far from ideal but the nature of this condition did preserve its usefulness. Specifically it would seem that samples were dropped in groups large enough to be easily noticed and not very frequently. This observation would also be consistent with the operation of the laptop however this was never proven and may not be an entirely safe assumption.

The second undesirable condition was microphone buffeting because of the lack of a windscreen. This condition resulted from the mechanical saturation of the microphones as they were struck by puffs of air caused by the turbulence of having the window open at high speeds. This condition cannot be effectively compensated for and, unlike the dropped samples, is too frequent to be avoided.

The third undesirable condition was aliasing resulting from downsampling without an antialiasing filter. Unfortunately the original data was not available, only the improperly downsampled data. These three conditions made the data from this acquisition effort unusable but the overall system design in Figure A.1 served as the inspiration for the acquisition system used in this thesis which is described in Chapter 3.