Real-Time Noise-Robust Speech Detection

by

Kevin Y. Luu

B.S., Massachusetts Institute of Technology (2009)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 2010

Author .....
Department of Electrical Engineering and Computer Science
June 30, 2010

Certified by .....
James R. Glass
Principal Research Scientist
Thesis Supervisor

Certified by .....
David Scott Cyphers
Research Scientist
Thesis Supervisor

Accepted by .....
Dr. Christopher J. Terman
Chairman, Department Committee on Graduate Theses

Real-Time Noise-Robust Speech Detection

by

Kevin Y. Luu

Submitted to the Department of Electrical Engineering and Computer Science
on June 30, 2010, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

As part of the development of an autonomous forklift of the Agile Robotics Lab at MIT's Computer Science and Artificial Intelligence Lab (CSAIL), this thesis explores the effectiveness and application of various noise-robust techniques towards real-time speech detection in real environments. Dynamic noises in the environment (including motor noise, babble noise, and other noises in a warehouse setting) can dramatically alter the speech signal, making speech detection much more difficult. In addition to the noise environments, another issue is the urgent nature of the situation, leading to the production of shouted speech. Given these constraints, the forklift must be highly accurate in detecting speech at all times, since safety is a major concern in our application.

This thesis analyzes different speech properties that would be useful in distinguishing speech from noise in various noise environments. We look at various features in an effort to optimize the overall shout detection system. In addition to identifying speech features, this thesis also uses common signal processing techniques to enhance the speech signals in audio waveforms.

In addition to the optimal speech features and speech enhancement techniques, we present a shout detection algorithm that is optimized towards the application of the autonomous forklift. We measure the performance of the resulting system by comparing it to other baseline systems and show 38% improvement over a baseline task.

Thesis Supervisor: James R. Glass
Title: Principal Research Scientist

Thesis Supervisor: David Scott Cyphers
Title: Research Scientist

Acknowledgments

Before I begin my exhaustive list of acknowledgments, I would like to say how deeply appreciative I am of the past research efforts that have made this thesis possible. In working with the Spoken Language Systems (SLS) Group at the MIT Computer Science and Artificial Intelligence Lab (CSAIL), I have been exposed to a wide variety of technologies. I am truly thankful for the audio processing framework and the developed tools that have fully supported me while taking on the research challenges of this thesis.

I would like to thank my advisor, Jim Glass, for his patience and guidance. I am greatly honored to have had the opportunity to work in his research group, and without his instrumental support, this thesis would not have been possible. Secondly, I want to thank Scott Cyphers. Without his technical expertise and assistance on a multitude of aspects pertaining to this thesis, I would have been truly lost.

Many acknowledgments go towards Tara Sainath, a graduated Ph.D student of the group who introduced me to the field of automatic speech recognition. During my undergraduate studies, it was Tara who helped me to develop a foundation and passion for the field of this thesis, and trained me to think like a researcher. I cannot thank her enough for all that she has done for me, and I am forever grateful to have worked with her.

I would also like to thank Ekapol Chuangsuwanich, a current Ph.D student of the group, for his extensive support and contributions throughout the year. In addition to the countless advice and recommendations towards improving my thesis, Ekapol has provided much help in developing the evaluation framework. His help is greatly appreciated.

And, of course, I would like to acknowledge all the unnamed people in my life that contributed to the production of this thesis, however indirect their contribution may have been. I would like to thank my friends and family for their unwavering love and support. To all the friends I have made over the past 5 years, thank you for making my time at MIT truly rewarding and enjoyable.

Finally, I would like to thank my parents for being wonderful role models and for always being a source of inspiration in my life. Their constant love and support, and endless sacrifices have made so many wonderful opportunities in my life possible.

This work was sponsored by the Office of Secretary of Defense under Air Force Contract FA8721-05-C-0002.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Over the past few years, there has been considerable progress in the field of speech recognition, resulting in highly accurate performance for specific tasks in constrained environments. For example, a less than 1% word error rate was achieved on a speaker-independent word recognition system with a large vocabulary database of over 20,000 words [31]. However, those systems operate well only in quiet environments; their performance degrades rapidly once the environment becomes noisy. In the area of telephone network communications, the accuracy of a speech recognition system can decrease by over 30% in a noisy environment, as compared with a clean environment [29].

Various phenomena that occur under noisy conditions cause degradations in the performance of speech recognition systems [6]. For example, additive noise introduced in the speech signal alters the feature vectors used by speech recognizers to represent the signal. In the recording environment, unaccounted conditions, spontaneous speaking styles, and the quality of the recording microphone can also distort the speech signal. Changes in speech production due to adverse noise conditions (a phenomenon known as the Lombard effect) can also alter the speech signal and, ultimately, reduce the performance of the recognition system [20].

1.2 Noise-Robust Speech Recognition

Numerous techniques have been studied to improve the robustness of speech recognition systems in different noise environments. These methodologies can be categorized into three primary areas based on their objectives: noise-resistant features, speech enhancement techniques, and noise adaptation [37]. Many of them employ general pattern recognition and statistical learning algorithms to improve the noise robustness of the speech recognition system without requiring prior knowledge of the environment.

Noise-resistant features rely on identifying better speech recognition features, taking into consideration the existence of noise. Many of the methods make no assumptions or estimates about the characteristics of the noise. However, each technique has advantages as well as disadvantages, depending on the noise conditions for which it is being evaluated [37]. Ultimately, these approaches aim to identify better, more robust, speech recognition features in noisy environments. Among the noise-resistant features, the most common technique is cepstral mean normalization (CMN) [17], which is used to normalize the spectral variation across recording environments. Other techniques, such as relative spectral processing (RASTA) [14], also attempt to remove background noises that varies slowly compared to variations in the speech signal. Both techniques can achieve a considerable amount of environmental robustness at negligible cost [11].

Speech enhancement techniques attempt to limit the effect of noise on speech by extracting the clean speech from the corrupted signal. Many of the common techniques were originally developed to improve the perceived quality of speech for human listening. These techniques aim to reduce the acoustic noise in the speech signal. The reduction in noise can be achieved by improving the signal-to-noise ratio (SNR) of the input signal [18]. A common technique is spectral subtraction, in which the estimated magnitude spectrum of the noise is subtracted from that of the noisy speech to obtain the estimated magnitude spectrum of the clean speech [35]. While many algorithms have been shown to improve the quality of the speech signal, these enhancements may not always translate to improvements in the performance of the

speech recognition system [36]. In some noise conditions, speech enhancement can actually cause over-estimation of the noise statistics and lead to degradation of the speech quality [38].

As opposed to acquiring clean speech through speech enhancement techniques, *noise adaptation* attempts to adapt the recognition models to the given noise environment. These techniques change the model parameters of the speech recognizer to account for noisy speech. Additionally, some noise adaptation techniques add noise models to the recognizer itself. These approaches perform well in environments with high SNR, where there is a clear distinction between the speech and the background noise. However, the adjustments in the model parameters often result in large variations in the recognizer's performance [34].

This thesis focuses on one component of speech recognition that can improve all noise-robust techniques: voice activity detection (VAD). VAD is commonly used to detect the presence of speech in an input signal by marking the boundaries of speech and non-speech segments. Studies have shown that the performance of a speech recognition system can be drastically improved by integrating a VAD module into the system [10]. For example, a real-world evaluation of an isolated-word recognizer showed that more than 50% of the word error rate is due to errors in endpoint detection [7]. Moreover, an accurate VAD reduces the response time and computation cost of speech recognition systems, since only detected speech frames are passed to the recognition algorithm [21].

1.3 Voice Activity Detection (VAD)

VAD is a classification problem in which features of the audio signal are used to separate the input into speech and non-speech. The two main components of VAD are feature computation and a classification algorithm. Typically, the audio signal is split into fixed-length frames, and feature values are computed for each frame. These values are then passed into the classification algorithm [28]. The VAD process is presented in Figure 1-1.

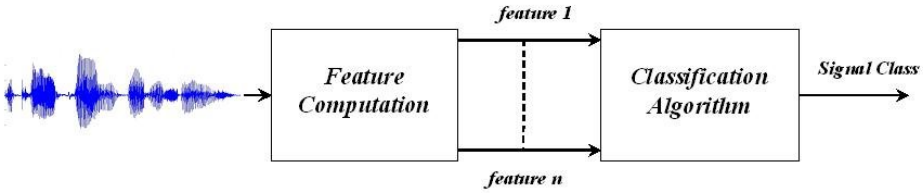

Figure 1-1: Block Diagram of Typical Voice Activity Detection (VAD) Systems [10].

Feature computation determines values for features that can effectively classify the audio into different classes. Dimension reduction is used to combine the raw feature values into a smaller number of more independent features. Quantization may also be applied at this time [10]. We will be exploring the details of this process of VAD systems later in this thesis.

The *classification algorithm* is a crucial step in VAD systems. Existing algorithms for classification are divided into rule-based algorithms and machine learning algorithms. For rule-based algorithms, the user decides how to compute the threshold values of the features in the classification decision. However, finding the optimal threshold values can be difficult, especially when the feature set is large and complex. Machine learning algorithms, on the other hand, use training data to construct a decision function, which classifies the input signal into one of the predefined classes without the need to define thresholds [10].

Selecting the appropriate features and classification algorithms is itself a difficult task, but the problem becomes increasingly challenging when the environment introduces noise into the speech signal. When this happens, it may be necessary for the VAD to be conservative in its decision function, indicating detection of speech when the decision is in doubt. Perhaps the most difficult environment for speech detection is the environment with low SNR, a condition where it is almost impossible to distinguish speech and noise using simple detection techniques [28].

In noisy recording environments, people often alter their speech to compensate for the noise. These changes in speech production by adverse conditions (known as the

Lombard effect) can have a profound effect on the signal, since the speech production is altered in an effort to communicate more effectively in a noisy environment [20].

1.4 Shouted Speech

Shouted speech is often referred to as the highest vocal mode of speech, and causes dramatic changes in vocal excitation [9]. The differences between normal and shouted speech lead to major degradation in speech recognition performance. In one study [20], the speaker-independent recognition accuracy is 90% when tested with normal speech, while the accuracy dropped to only 27% when tested on shouted speech. In another study [9], results from a speaker-ID system show that a mismatch in vocal mode can seriously impact the system's performance. When the system is trained on normal speech, but tested on shouted speech, the system's accuracy decreases from 97% in matched mode to 54% in mismatched modes. From these studies, it is clear that shouted speech is a major concern in the performance of speech systems. Unfortunately, little research has been carried out on improving the accuracy of recognition algorithms for shouted speech [13].

1.5 Agile Robotics Lab Autonomous Forklift

The research and progress in this thesis is focused on the speech component of the autonomous robotic forklift being developed by the Agile Robotics Lab at MIT, as shown in Figure 1-2 [4]. The project is intended for military warehouse applications, and it is designed to operate in unstructured environments such as outdoor-packed earth or gravel regions. Its goals include the ability to move pallet loads under voice command, and eventually, autonomously transport pallets of a warehouse to a new location [4].

Figure 1-2 showcases the autonomous forklift of the Agile Robotics Lab at MIT, the primary application of this thesis. Currently, four array microphones are installed on the front, left, rear, and right of the forklift's top rack. These microphones are

Figure 1-2: Image of the autonomous forklift of Agile Robotics Lab at MIT [4]. Photo courtesy of Jason Dorfman.

continuously recording the audio signals for the forklift's speech recognition system.

1.6 Thesis Goals

While many commercial speech recognition systems are very accurate in detecting speech, the challenges come from the recording settings and surrounding environments of the forklift's speech recognition system. First, speech commands directed at the forklift will be spoken from a distance of ten feet of more from the recording microphones. This is because an operating forklift is potentially dangerous and will stop if personnel are within ten feet. At the same time, background noises can dramatically alter the speech signal. People speak commands loud enough to ensure that they would be heard over engine noise, so most of these commands are shouted (see Section 1.4 for challenges faced with shouted speech). Finally, the microphones are continuously recording (as opposed to a push-to-talk model), requiring the VAD to identify speech endpoints.

The overall goal of this thesis is to develop an effective shout detection algorithm

that addresses these challenges. The algorithm would extract the necessary features suitable for shouted speech. More specifically, this thesis investigates different speech features and develops a robust detection method that leads to an improvement in shout detection accuracy. We hope that our method will be robust in many different noise conditions, overcoming limitations of many current noise robust techniques.

While our shout detection algorithm can be used in many different applications, this thesis focuses on integrating our system into the autonomous forklift to detect any commands in noise environments that are common for the forklift. In a setting where background noises are always present and occur at random times and places, there is a great need for a noise-robust speech recognition system. Additionally, many forklift operations can be dangerous if the system fails to operate correctly. Having a shout detection system on top of the existing speech recognition system provides an additional fail-safe mechanism.

1.7 Overview

The remainder of this thesis is organized as follows: Chapter 2 describes past research that pertains to the development of this thesis. It discusses various features that were developed for speech detection in various noise environments. Chapter 3 presents the methodology of the thesis. Particularly, Chapter 3 presents the different data sets that are used in the thesis, along with their data collection process. This chapter also describes the system evaluations that are used for this thesis. Chapter 4 presents the experiments conducted on the features of interest and showcases the design decisions behind the final shout detection algorithm. Finally, Chapter 5 concludes the work of this thesis and provides a few remarks about future work.

Chapter 2

Related Work

This chapter describes work related to the research presented in this thesis. As an example of how the various modes of speech production affect speech systems, we first look at the effectiveness of various modes of speech production on the accuracy of a speaker-ID system. This provides us with a better understanding of the effect of shouted speech in a speech recognition system trained under normal speech. We then cover various noise-robust features and speech enhancement techniques commonly used in speech detection. We also discuss the various systems that are essential to this thesis. We describe SUMMIT, the speech recognition system that is integrated into the autonomous forklift of the Agile Robotics Lab. Lastly, we explain the speech activity detection module that we use as a baseline system in evaluating the work of this thesis.

2.1 Speaker-ID System Based On Speech Modes

While speech recognition systems are designed to work well with normal speech (normally phonated speech), their performance degrades dramatically with the introduction of other modes of speech production, including shouted speech. Often, shouted speech introduces errors in the recognition and, ultimately, alters the performance of the overall system. The authors in [9] present an analysis of speech characteristics for five vocal modes: whispered speech, soft speech, neutral speech, loud speech, and

shouted speech (as described in Table 2.1). They identify distinctive features between the different modes to develop a classification system for vocalization. To demonstrate the effects of the different vocal modes, the authors in [9] measure the accuracy rate of the speaker-ID system for training and testing data under the different speech modes.

Table 2.1: Speech Modes in Automatic Speaker-ID (ASI) system [9]

Speech Mode	**Description**
Whispered Speech	Lowest vocal mode with limited vocal cord vibration
Soft Speech	Moderate mode between whispered and neutral speech
Neutral Speech	Normally phonated speech
Loud Speech	Mode between neutral and shouted speech mode
Shouted Speech	Highest vocal mode with most dramatic change in vocal excitation

The speech modes are said to be matched when the same speech mode is used in both training and testing data, and mismatched when the training and testing modes differ. To explore the impact of the vocal mode on the speaker-ID system, speech from the five vocal modes is employed in a closed in-set speaker recognition system. The experimental data used in [9] consists of 110 sentences collected from 12 subjects in the neutral mode. A round-robin technique is used to obtain the average performance of the speaker-ID system for each speech mode. Table 2.2 presents the findings of the research in [9].

The results in Table 2.2 suggest that matched speech modes perform well, but mismatched speech modes can seriously impact the system's performance. This phe-

Table 2.2: Accuracy Rate (%) of the Automatic Speaker Identification (ASI) System for training and test data under five speech modes [9]

		Test				
		Whispered	**Soft**	**Neutral**	**Loud**	**Shouted**
Train	**Whispered**	94.6	33.3	30.4	23.3	17.9
	Soft	57.9	97.5	86.3	61.7	41.7
	Neutral	46.7	86.7	98.8	86.3	56.3
	Loud	39.2	66.7	92.1	98.3	64.2
	Shouted	27.1	40.4	53.8	68.3	97.1

nomenon is evident through the drastic reduction in average accuracy rate, from 97% in matched mode to 54% in mismatched mode. Particularly, the accuracy rate of testing shouted speech on a speaker-ID system trained in normal speech is only 56%, suggesting the negative impact that shouted speech has on the speech system [9].

While most speech systems are trained from a large data set, the training data rarely accounts for variations in the articulation of the speech. Hence, the performance of the system degrades dramatically with the introduction of shouted speech. Research in [9] highlights this phenomenon, showing that recognizing shouted speech is an important challenge faced by many speech recognition and detection systems. To address this challenge and other difficulties, speech features are investigated and developed to be robust in different noise conditions.

2.2 Noise-Robust Features for Speech Detection

The key to having a high accuracy rate is selecting the appropriate features to optimize the performance of the speech recognition system. Thus, much research and effort is dedicated towards investigating the use of different features in speech detection. Some features are based on the autocorrelation of the signal, which takes advantage of the periodic characteristics of the speech signal. On the other hand, spectral-based features use the short-time power spectrum of the signal in making classification decisions [33].

However, in any recording situation, speech is produced by a human; some properties of speech do not occur in many types of background noises. Likewise, much research has been devoted toward improving the robustness of speech detection systems by finding ways to exploit properties of human speech signals. One such feature is the fundamental frequency and its harmonics in voiced speech, since voiced speech has a predictable harmonic structure [27]. Particularly, the authors in [27] present a new algorithm for generating noise-robust features based on the Harmonic+Noise Model (HNM). HNM decomposes a speech signal into its harmonic and random components, providing the capability to process the signal components independently.

This framework is used to improve the robustness of the speech recognition systems to additive noise. HNM can be a means of generating enhanced features for speech recognition [27].

Machine-learning approaches have also been applied towards selecting the best features for a given noise environment. For example, the authors in [8] use the Partially Observable Markov Decision Process (POMDP) framework for the Voice Activity Detection (VAD) process [23]. POMDP models a decision process, which assumes that observations cannot be made to estimate the current states. Decisions are based on the action with the maximum reward over a long period of time, where reward is defined as the improvement in classification rate based on a particular decision. By using this model, only the most suitable features for the current noise are considered. Not only does this reduce the complexity of computation, but POMDP also improves the overall performance of the VAD system [8].

Using POMDP, VAD systems can utilize different types of features according to the estimated state of the recording environment. As an example, energy has been frequently used as a feature for VAD in high SNR conditions, but its effectiveness diminishes drastically with low SNR. In low SNR environments, spectrum-based features tend to work better. POMDP VAD systems can use both features effectively by tracking the type of noise and SNR in the current environment, and selecting the most salient approach.

While there are a wide variety of noise-robust features, this thesis focuses on two particular characteristics of speech signals that are commonly used in speech detection: energy and entropy. The following sections discuss work pertaining to using these features in speech detection systems.

2.3 Energy-Based Speech Detections

Features derived from the energy of the speech signal are the most widely used features in locating the endpoints of speech utterances. In environments with low intensity stationary background noises (high SNR conditions), the energy of the signal is a

simple and effective feature [33]. With high SNR, the signal power is much higher than the noise power corrupting the signal. Hence, a high SNR means that the background noise is less obtrusive. Because of the distinction in energy between speech and noise signals in high SNR conditions, energy has been commonly used as a threshold in labeling a particular signal as either speech or non-speech.

While there are many energy-derived measures of a signal, a common method is the root mean square energy (RMSE), which is expressed formally by Equation 2.1 [12].

$$E_n = [\frac{1}{W}\sum_{i=1}^{W} s_n^2(i)]^{\frac{1}{2}} \tag{2.1}$$

A rectangular window of width W is used to segment the speech into frames, and $s_n^2(i)$ denotes the i^{th} windowed speech sample in the frame number n. Taking into account the mentioned parameters, E_n in Equation 2.1 presents the RMSE of the speech frame n.

Using Equation 2.1 as a means of computing the energy of a given frame, we are interested in seeing the effectiveness of using frame-based energy as a feature in distinguishing speech from non-speech in a given waveform. The test waveform is taken from the Simulated Shout Data Set, as described in Section 3.1.1. This particular waveform contains the forklift beeping noise as the dominant background noise. Its SNR is 10 dB, a level that is closest to the noise environment of the forklift application. Using this test waveform, we analyze the feature's responses.

Figure 2-1 presents the feature's responses to the test waveform. The first plot shows the normalized magnitude of the waveform, with the marked speech boundaries. The second plot shows the spectrogram of the same waveform. From these two plots, we can see that the forklift beeping noise is present in the signal. However, the speech is the dominant signal because of the SNR level in this waveform. Because of the waveform's SNR level, the graph of the feature (shown in the third plot of Figure 2-1) shows that there is a significant difference between the energy level of the speech segments and that of the forklift beeping noise. This difference allows us to set a hard

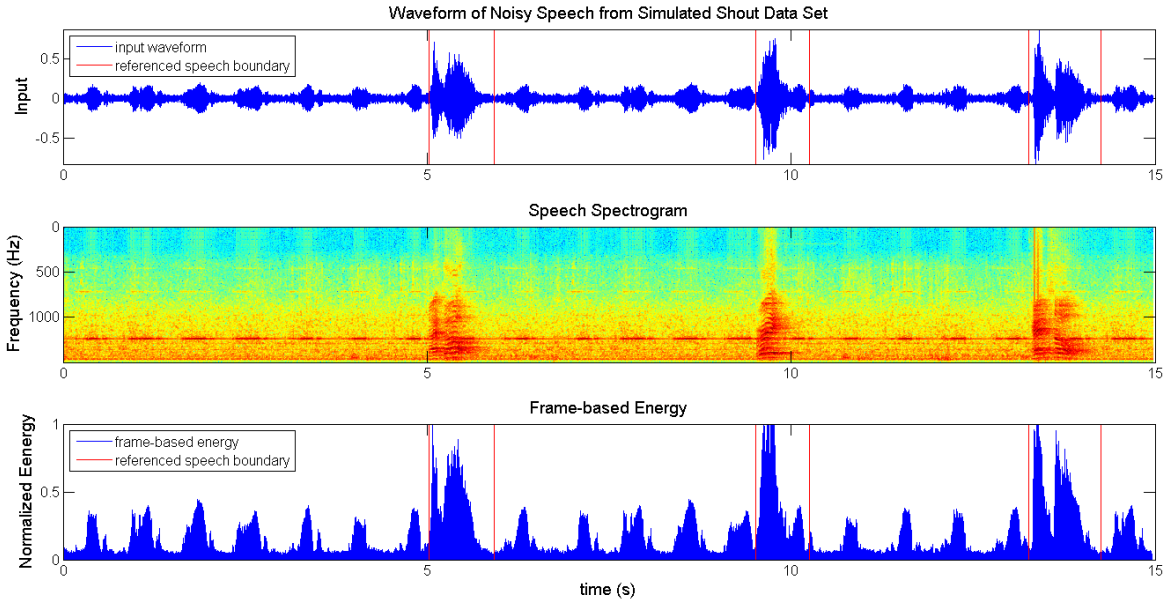

Figure 2-1: Responses of frame-based energy to three utterances of speech in a high-SNR environment (SNR of 10 dB) where the forklift beeping noise is the dominant background noise. The first plot shows the input waveform of noisy speech, with the marked referenced speech boundaries. The second plot shows the spectrogram of the same waveforms. Lastly, the third plot shows the feature values computed for this particular waveform.

threshold to effectively classify each frame as speech or non-speech.

While the feature response in Figure 2-1 shows a clear distinction between speech and non-speech segments, the distinction is primarily due to the high SNR level. If the SNR level is lower, it would be more difficult to distinguish speech from non-speech with this feature.

Figure 2-2 showcases the response of the frame-based energy in an input waveform of low SNR. It shows the feature's response to the same waveform used in Figure 2-1, but at an SNR level of 0 dB. In reducing the SNR from 10 dB to 0 dB, we can see that the differences between the speech and non-speech components diminish, making it much more challenging to identify the speech components based solely on the feature's response. Particularly, the peaks of the speech components are very similar to the peaks of the forklift beeping components of the signal, making it very difficult to use the peak levels to distinguish the speech from the non-speech segments. The

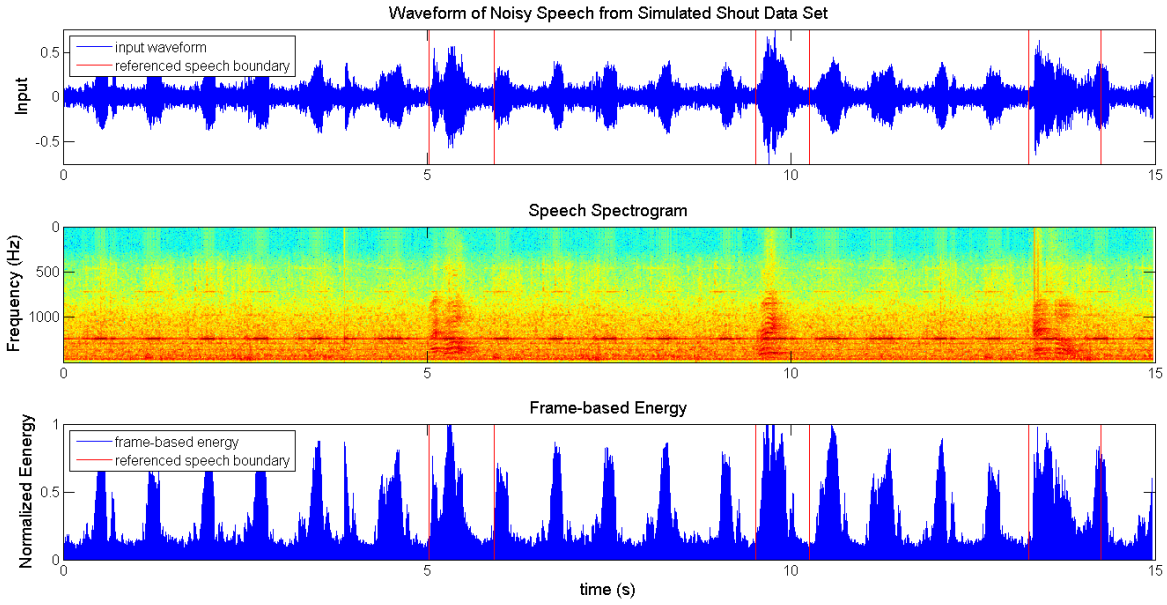

Figure 2-2: Responses of frame-based energy to three utterances of speech in a low-SNR environment (SNR of 0 dB) where the forklift beeping noise is the dominant background noise. By reducing the SNR level from 10 dB to 0 dB, the feature's response (the bottom plot) shows that it is much harder to distinguish speech from non-speech in the input waveform.

comparison between Figure 2-1 and 2-2 shows how dependent the performance of the feature is on the SNR level. The dependency on high SNR conditions marks one of the many limitations in the frame-based energy.

2.4 Entropy-Based Speech Detections

While energy-based features are often used for endpoint detection, they become less reliable in low SNR conditions and in the presence of non-stationary noise (i.e. when the mean or the variance of the noise varies) [2]. To address these limitations, various studies investigate entropy-based algorithms for accurate and robust speech detection in noisy environments. Entropy is defined as a measure of disorganization or uncertainty in a random variable. The application of the entropy concept is based on the assumption that the short-term spectrum is more organized during speech segments than during noise. Additionally, the spectral peaks of the spectrum are supposedly

more robust to noise. Therefore, a voiced region of speech would induce low entropy, since there are clear formants in the region [35]. Particularly, the entropy in the time-frequency domain (known as spectral entropy) has been investigated as a new approach to endpoint detection.

2.4.1 Spectral Entropy

The spectral entropy for the n^{th} frame is defined and measured in the following manner. Let the spectrum $X(n,k)$ of $x(i)$ be obtained by the K-point Discrete Fourier Transform (DFT), where $k = 0, ..., K/2$. First, a pseudo probability density function (pdf) $P(|X(n,k)|^2)$ for the spectrum $|X(n,k)|^2$ is computed by Equation 2.2.

$$P(|X(n,k)|^2) = \frac{|X(n,k)|^2}{\sum_{k=0}^{K/2} |X(n,k)|^2} \tag{2.2}$$

To improve the discriminability of the pdf between speech and non-speech signals, some heuristics are used by the authors in [2]. They are defined formally in Equations 2.3 and 2.4.

$$|X(n,k)|^2 = 0, \quad k < 250\ Hz \quad or \quad k > 3750\ Hz \tag{2.3}$$

$$P(|X(n,k)|^2) = 0 \quad if \quad P(|X(n,k)|^2) \geq 0.9 \tag{2.4}$$

The heuristic in Equation 2.3 is used to limit the frequency range of the signal to the most significant perceptual region of the speech spectrum (e.g. telephone bandwidth), which is in the frequency range from 250 Hz to 3750 Hz. Additionally, Equation 2.4 states that if the pdf of the signal is above 0.9, then it is disregarded in the calculation of the spectral entropy. This heuristic is to avoid strong tones.

With these heuristics, the spectral entropy $H(n)$ for the n^{th} frame is expressed formally by Equation 2.5 [2].

Table 2.3: Word error rates for endpoints obtained by hand-labeling, entropy-based, and energy-based algorithms for different types and levels of noise [19]

Noise	Method	**Word Error Rate (%)**		
		5 dB	10 dB	15 dB
white	energy	51	28	12
	entropy	43	24	11
	hand	36	20	11
pink	energy	40	14	0
	entropy	32	10	0
	hand	20	10	0
car	energy	7	0	0
	entropy	2	0	0
	hand	0	0	0
average	energy	32.7	14	4
	entropy	25.7	11.3	3.7
	hand	22	10	3.7

$$H(n) = -\sum_{k=0}^{K/2} P(|X(n,k)|^2) \cdot \log(P(|X(n,k)|^2)) \tag{2.5}$$

It has been found that the spectral entropy is very useful in distinguishing the speech segments in continuously recorded utterances from non-speech portions, especially in noisy environments. To showcase this trend, research described in [19] shows the effectiveness of entropy-based algorithms over energy-based algorithms in robust speech detection.

In the experiment conducted in [19], the recognition results (with the speech boundaries) obtained by hand-labeled, energy-based and entropy-based algorithms are compared in the presence of different levels and types of noise. The recognition accuracy is evaluated by the average of 10 testing speakers. Different types of noise are collected from the NOISEX-92 noise-in-speech database [5]. The word error rates with respect to different levels of white noise, pink noise, and car noise are shown in Table 2.3 [19].

Table 2.3 presents the findings of the research conducted in [19]. We can see that the error rates obtained using entropy-based endpoint detection are only slightly higher than (or even identical to) those obtained by hand-labeled algorithms in all

cases. However, for all types and levels of noise, the entropy-based algorithm has a lower error rate than the energy-based algorithm. This finding suggests that entropy-based algorithms performs better than energy-based algorithms [19].

2.4.2 Relative Spectral Entropy

To further improve the robustness of spectral entropy as a speech detection feature, relative spectral entropy (RSE) has been proposed. The RSE is very useful when the speech signal is contaminated by stationary noise. In our application, where the speech detection system is in an environment with many constant noise sources, this modification to the spectral entropy feature acts as a useful enhancement. The RSE is determined with respect to the mean spectrum, which is computed over neighboring frames. These frames form a shifting segment (with length of a few hundred milliseconds). The RSE is computed as follows [2]. Let the pseudo probability density function $P(|X(n,k)|^2)$ for the spectrum $|X(n,k)|^2$ be computed in the same manner as in Equation 2.2. The mean spectrum $|Y(n,k)|^2$ for the n^{th} frame is expressed in Equation 2.6.

$$|Y(n,k)|^2 = \frac{1}{M+1} \sum_{m=n-M/2}^{n+M/2} |X(m,k)|^2 \tag{2.6}$$

In Equation 2.6, M represents the number of frames in the shifting segment. We apply the heuristics as presented in Equation 2.3 and 2.4. With this, the RSE $H_r(n)$ is defined by Equation 2.7 [30].

$$H_r(n) = -\sum_{k=0}^{K/2} P(|X(n,k)|^2) \cdot \log(\frac{P(|X(n,k)|^2)}{P(|Y(n,k)|^2)}) \tag{2.7}$$

Note that the RSE is a Kullback-Leibler (KL) divergence between the current spectrum $|X(n,k)|^2$ and the mean spectrum $|Y(n,k)|^2$. In computing the KL divergence, we are measuring the change in the the spectral entropy. Hence, the new computation of spectral entropy takes into account noise that is stationary in a longer window[30].

To analyze the effectiveness of using RSE as a feature for detecting speech in different (clean and noisy) environments, we analyze the feature's response when given a test waveform. The RSE is computed using the process described above. Particularly, we use a frame length of 30 ms, frame shift of 10 ms, and a 1024-point FFT. Additionally, we set the number of frames in the shifting segment M to be 50. These parameters are suggested by the authors in [2] and tested in our application for optimality. Figure 2-3 presents the spectrogram of three speech segments in a clean environment and the corresponding RSE.

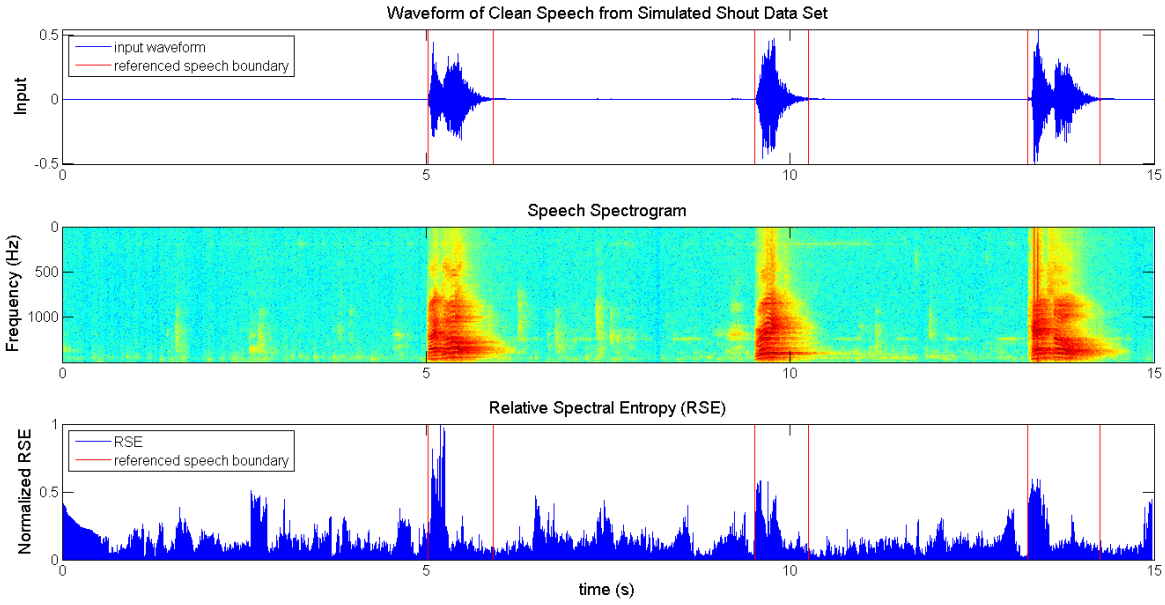

Figure 2-3: Relative Spectral Entropy (RSE) of the three utterances of speech in clean environment. The first plot shows the input waveform of clean speech, with the marked referenced speech boundaries. The second plot shows the spectrogram of the same waveforms. Lastly, the third plot shows the RSE values computed for that particular waveform.

From Figure 2-3, we can see that RSE is very sensitive to any background noise. While the input waveform shows three distinct speech signals, the corresponding spectrogram shows signs of minor noises in the non-speech regions (as indicated in yellow). The existence of noise in the input signal shows that the environment is not completely clean; any background noise is captured by the recording microphone. Because of the subtle background noise, there is much disturbance in the response

of RSE, making it difficult to determine the speech segments based on the values of the RSE. However, in comparing the referenced speech boundaries to the RSE, we can see that the RSE increases as soon as speech is introduced into the environment. This phenomenon suggests that we can use this property of RSE in distinguishing speech in a noisy environment.

To further investigate the usefulness of RSE in our application, we analyze the RSE of a speech signal in a noisy environment. Figure 2-4 showcases the RSE of a speech signal in an outdoor street environment at an SNR level of 10 dB. The audio signal is from the Simulated Shout Data Set (as described in Section 3.1.1). In this data set, the noise condition used is labeled Mass. Ave.

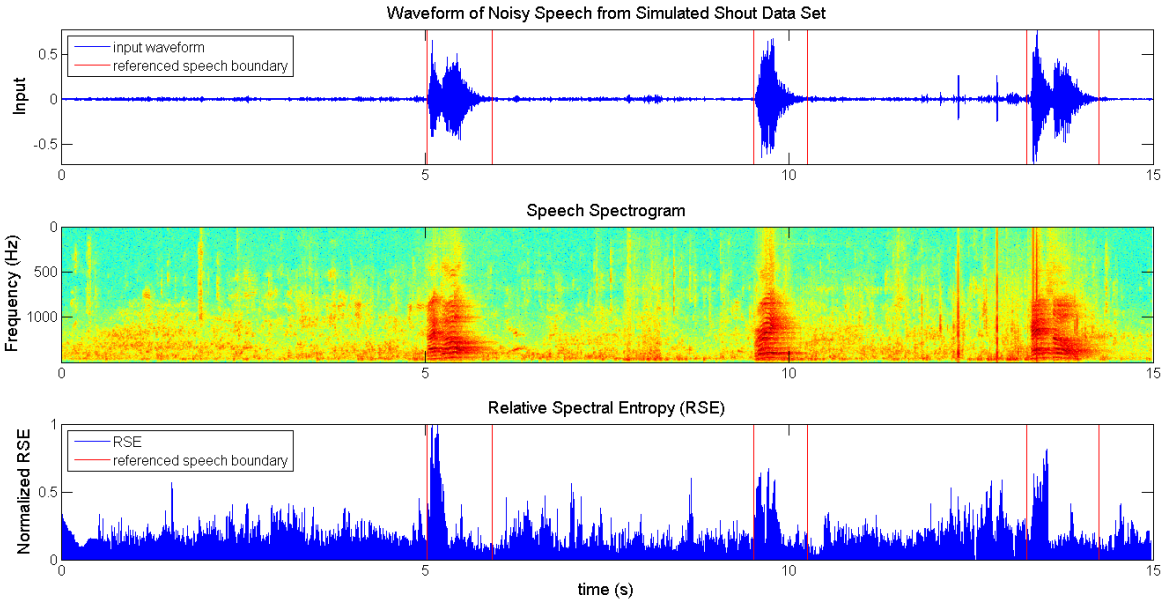

Figure 2-4: Relative Spectral Entropy (RSE) of the three utterances of speech in noisy environment with a SNR of 10 dB. The input signal is collected from an outdoor street environment.

In viewing Figure 2-4, we can see the same trend in the speech/non-speech transition. In both environments (clean and noisy), the value of RSE has a rapid increase, suggesting the introduction of a speech segment. Even with much more noise in the recording environment, the RSE is still able to distinguish the speech regions of the signal.

While the RSE shows minimal differences in responses between clean and noisy environments, the feature should not have the same level of challenges in distinguishing speech components in a clean environment. The feature's response in Figure 2-3 and 2-4 are similar, yet detecting speech in a clean environment should be much easier than detecting speech in a noisy environment. On the other hand, the frame-based energy is able to distinguish the speech components well in environments of high SNR, but Figure 2-2 shows that the feature does not perform well in low SNR conditions. In other words, both features (frame-based energy and RSE) have their limitations in distinguishing speech from non-speech segments.

2.5 Energy-Entropy-based Speech Detections

While energy and entropy are common features used for speech detection, they both have limitations in noisy environments. Entropy is useful in differentiating voiced from unvoiced speech, but it fails in the presence of babble noises. Energy, on the other hand, performs well in these conditions because of its additive property: energy of the sum of speech and noise is almost always greater than the energy of noise. However, energy-based algorithms face difficulty in differentiating speech from unexpected background noises (both stationary and non-stationary noises). From this, the authors in [24] propose a method to combine the advantages of the two features to form a feature known as Energy-Entropy (EE).

The EE feature is computed as follows. First, the energy and entropy are calculated for each frame. For each frame i, the energy $E(i)$ is obtained by Equation 2.1. The spectral entropy $H(i)$ is calculated in the same way as described in Section 2.4.1.

With the energy $E(i)$ and entropy $H(i)$ of each frame i computed in parallel, the next step is to adjust both parameters by shifting their respective baselines. This is achieved by subtracting the average amount of the first 10 frames accordingly. We denote the average energy and entropy of the first 10 frames as C_E and C_H, respectively. In this model, we are assuming that the first 10 frames are non-speech. Given this setup, the Energy-Entropy feature $EE(i)$ is expressed formally by Equation 2.8.

$$EE(i) = \sqrt{1 + |(E(i) - C_E) \cdot (H(i) - C_H)|} \tag{2.8}$$

To further investigate the usefulness of the EE feature towards our forklift application, we use the presented method of computing the EE feature to analyze the feature's response when given a test waveform. We measure the feature's response to the same input waveform used in Figure 2-3. Figure 2-5 presents the spectrogram of three speech segments in a noisy environment and the corresponding EE values.

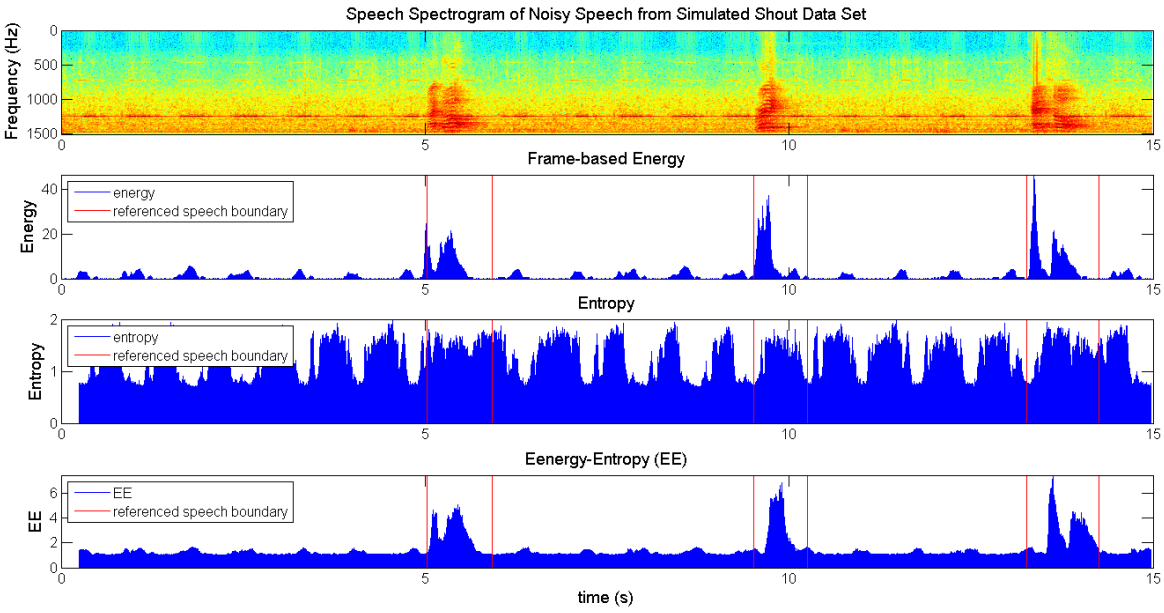

Figure 2-5: Response of the Energy-Entropy (EE) feature to three utterances of speech in noisy environment. This sample waveform has the forklift beeping noise as its background noise, and its SNR is 10 dB. The first plot is the spectrogram plot, while the subsequent plots represent the computed features (Frame-based Energy, Entropy, and EE). The red vertical lines on the feature plots represent the referenced speech boundaries. Note that combining the energy and entropy into one feature emphasizes the speech components in the signal while reducing the effect of the background noise.

In Figure 2-5, we see the responses of the energy feature, the entropy feature, and the EE feature to the test waveform (three utterances collected from the Forklift Beep noise condition of the Simulated Shout Data Set). In the frame-based energy plot (the second figure), we can see that the energy feature is able to locate the

speech segments of the waveform fairly well, but the forklift beeping noise causes small disturbances in the energy calculation. However, including the entropy in the feature calculation improves the overall feature by locating the speech portion and ignoring the noise components of the signal. The EE plot shows that the speech segments are emphasized while the noise segments are minimized.

Through this result, we can see that the EE feature utilizes the advantages of both individual features; each feature is able to compensate for the limitation of the other feature. Given our application, we believe that the method of combining the two features could be beneficial in improving the accuracy and performance of the shout detection system [24].

2.6 SUMMIT Speech Recognition Framework

SUMMIT is a landmark-based speech recognition system developed by the Spoken Language Systems Group in MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) [16]. It uses Finite State Transducer (FST) networks to transform audio that matches a set of low level acoustic models into recognized utterances; these utterances are defined to obey a series of context-free grammar rules or n-gram language model [15]. The speech recognizer is a composition of four stages: $C \circ P \circ L \circ G$. C and P encode context-dependent phone labels and phonological rules. L is the mapping from phonemes to words, and G is the language model. The acoustic model of the SUMMIT recognizer is created by using the audio recordings of speech and their transcription to create statistical representation of the word units.

2.7 Speech Activity Detection Element for GStreamer (gstSAD)

As a baseline system, this thesis uses the Speech Activity Detection Element for GStreamer (gstSAD), an energy-based speech activity detection system developed by Dr. Michael Mason, School of Engineering Systems of Queensland University

of Technology [26]. The gstSAD system is a three-stage process that analyzes the incoming audio signal as a series of non-overlapping frames. The first stage of the process, the feature extractor, observes salient frame-based features. The classifier in the second stage uses these features to decide on the presence of speech in the frame. In the third stage, the frame-based decisions are smoothed in a simple state machine to avoid very short sections of activity, and to capture onset and roll-off sections of activity. For gstSAD, frame energy is the only feature being used to classify the frames. For classification, a hard threshold (default at -26 dB) is being used for decision making.

Chapter 3

Experimental Methodology

This chapter discusses the methodology for collecting the data and evaluating the features in this thesis. There are two primary components of the experimental methodology: data collection and system evaluation. Section 3.1 describes the data sets used for the autonomous forklift application, along with the process of collecting the data sets. Section 3.2 discusses how we evaluate the system's performance. Specifically, the section describes the evaluation metrics that are used and the process of generating the Detection Error Tradeoff (DET) curves.

3.1 Data Collection

This section focuses on the data sets that are collected for the experiments described in Chapter 4. The data used in this thesis consists of the following sets of audio samples: Simulated Shout Data Set, Forklift Data Set, and Outdoor Data Set.

3.1.1 Simulated Shout Data Set

The Simulated Shout Data Set is designed and collected specifically towards the application of the autonomous forklift. This data set focuses on different physical settings, which translates to very diverse noise environments. The data consists of two components: clean shouts and noise. The two components are collected separately,

then merged together to simulate a shout data set in noise.

The clean shout data is recorded in a quiet environment, sampled at 16 kHz and quantized to 16 bits per sample. The microphone used to collect the audio sample in this data set is a voice-tracking array microphone, as shown in Figure 3-1.

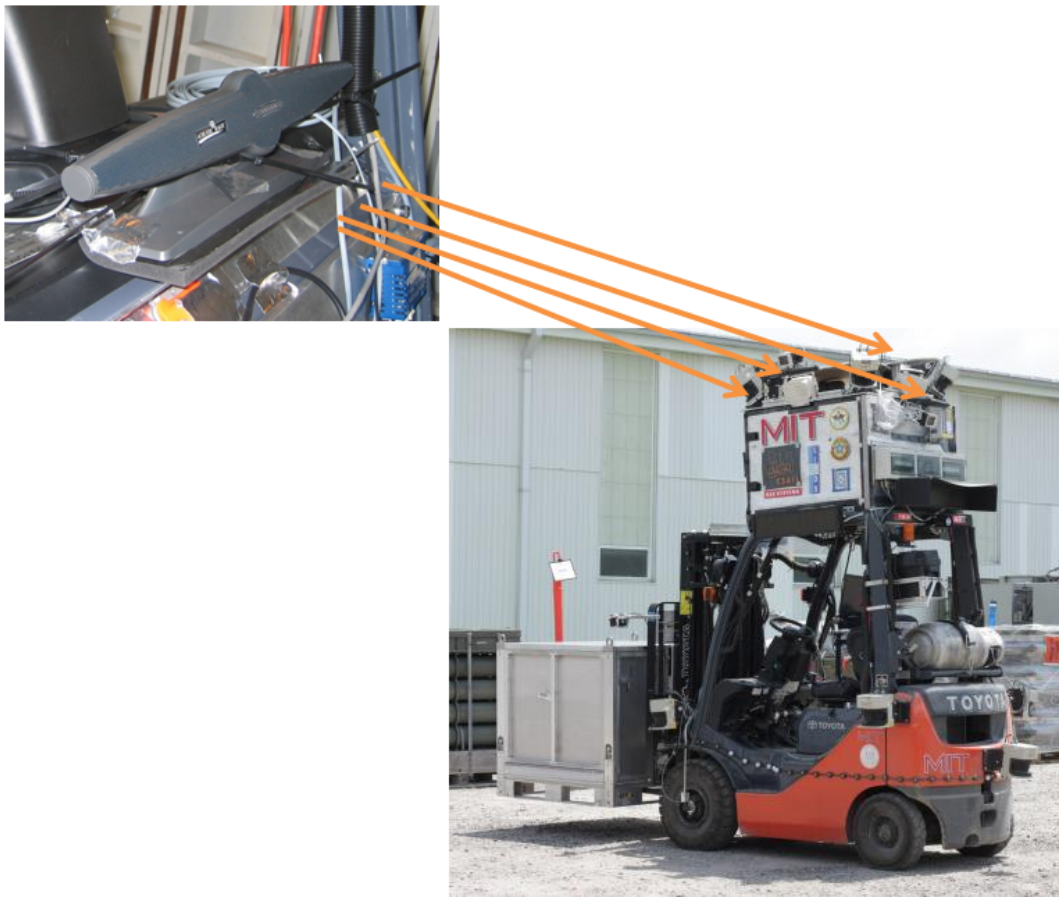

Figure 3-1: The array microphone used on the autonomous forklift. A total of four microphones are installed, one on each side of the forklift's top rack. Photo courtesy of Jason Dorfman.

The microphone used for this data collection is the Acoustic Magic Voice Tracker Array Microphone. It is designed to form an active beam, making it a good fit for the requirements of the autonomous forklift [3].

For this data set, we use one of these microphones to record shouted utterances in an enclosed environment. The shout utterances are drawn from 23 subjects: 13 males and 10 females. A total of 25 speech phrases are shouted by each subject, as presented in Table 3.1.

Table 3.1: Speech Phrases in Simulated Shout Data Set

Back up	Brake	Cancel	Don't move	Forklift
Freeze	Get out of the way	Go back	Halt	Hey
Ho	Hold it	Hold on	Hold up	No
Pause	Quit it	Stop	Stop Forklift	Stop it
Stop moving	Wait	Watch it	Watch out	Whoa

While the shout data collection is not collected in the active forklift environment, we aim to simulate the actual environment by emulating the various noise conditions. During the data collection, the subject is exposed to a recording of forklift noises (beeping, motor, wind, babble, etc.) at a loud volume through headphones. The goal is to trigger Lombard effect that would occur in the forklift environment.

The noise data is recorded in various noisy locations using the same array microphone, sampled at 16 kHz, and quantized to 16 bits per sample. The data set consists of seven noise conditions, specified by their physical locations. The noise data is recorded in real environments, to emulate possible noise conditions in which the autonomous forklift might be. The noise conditions are described in Table 3.2.

Table 3.2: Noise Conditions in Simulated Shout Data Set

Name	**Description**
Forbe's Cafe	Babble noise in cafe environment
Forklift Beep	Beeping noise from operating forklift
Forklift Motor	Motor noise from operating forklift
Loading Dock	Recordings from loading dock during normal operation hours
Mass Ave	Street noise recorded at 84 Massachusetts Ave
Vassar St	Street noise recorded at Vassar St
Wind Tunnel	Stationary noise recorded next to MIT Wind Tunnel (Building 17)

Each set of audio data (clean shout and noise) is collected separately, and mixed together at different SNRs. For this particular experiment, we are interested in SNRs of 0 dB, 5 dB, and 10 dB. To produce the desired SNR condition, we adjust the gain of the noise accordingly while mixing it with the clean shout audio data (i.e. the resulting waveform is a sum of the adjusted clean shouts and the noise). The

resulting waveform simulates the subject shouting in the various noise conditions. With 23 speakers, 7 noise conditions, and 3 SNRs, the shout data set consists of 483 waveforms.

A reference transcription is also created by manually transcribing the corresponding audio sample. The reference transcription is composed of the start and end time stamps of each speech and non-speech segment. It is used in the evaluation component of the experiments described in Chapter 4. An example of the reference transcription can be found in Appendix A.

3.1.2 Forklift Data Set

The Forklift Data Set is designed to collect realistic audio data that best represents the actual noise environment in which the autonomous forklift operates. This data set is created from a collection of live recordings with the operating forklift. This enables us to design a more robust shout detection algorithm that is optimized for the actual noise and speech conditions in which the forklift is operating.

This data set is collected with the autonomous forklift operating in an enclosed warehouse setting. The audio waveforms are sampled at 16 kHz, and quantized to 16 bits per sample. The audio samples are collected using the four array microphones that are installed on the forklift.

The Forklift Data Set consists of 18 subjects: 15 males and 3 females. Each subject was given a list of stop commands to speak on each side of the forklift (left, right, front, back). There are a total of 5 distinctive stop commands, as presented in Table 3.3.

Table 3.3: Speech Phrases in Forklift Data Set

Robot stop
Forklift stop
Robot stop moving
Forklift stop moving
Stop right now

For each stop command, the subject spoke in two vocal modes: normal speech and

shouted speech. Additionally, each subject shouted a random utterance, which corresponds to an Out Of Vocabulary (OOV) utterance. In total, there are 11 utterances (collected from each subject) on each side of the forklift. Ultimately, each subject contributed 44 utterances. With 18 subjects and 4 noise conditions, the forklift data set consists of 72 waveforms, with over 3,000 utterances.

For this data set, the forklift operates in different states, which corresponds to the different noise conditions. The four noise conditions in this data set are babble noise, forklift beep noise, forklift motor noise, and quiet noise. These noise conditions are described in Table 3.4.

Table 3.4: Noise Conditions in Forklift Data Set. Note that in all noise conditions, all array microphones on the forklift are continuously recording during data collection.

Name	**Description**
Babble Noise	Babble noise around the operating forklift
Forklift Beep Noise	Beeping noise from operating forklift
Forklift Motor Noise	Motor noise from operating forklift
Quiet Noise	Quiet environment, with non-operating forklift

For each waveform, a reference transcription is created in the same manner as the transcriptions in the Simulated Shout Data Set. Using the reference transcriptions, we can evaluate the performance of a particular configuration of the shout detection system.

3.1.3 Outdoor Data Set

Unlike the Simulated Shout Data Set and the Forklift Data Set, the Outdoor Data Set provides an even more realistic setting by operating the forklift in an outdoor environment while collecting audio samples of subjects shouting at the forklift as the forklift is performing pallet operations. The objective of this data set is to provide a set of data that is as close as possible to the operating environment of the forklift. The noises that we collect in an outdoor environment are spontaneous; we do not have control of street noises and construction that occur during data collection. However,

we want to create a data set that acts as a testing platform to fully evaluate the final shout detection system for our application.

The Outdoor Data Set is collected with the autonomous forklift operating in a gravel lot. The audio waveforms are sampled at 16 kHz, and quantized to 16 bits per sample. The audio samples are collected using four array microphones (as described in Section 3.1.1) that are installed on the forklift. During the data collection, the forklift is performing pallet operations (such as picking up pallets and driving forwards and backwards).

For this data collection, the subjects are positioned on different sides of the forklift. Each subject stands at least five feet away from the forklift while shouting the given list of commands. When the forklift is moving, the subjects are instructed to follow the forklift while maintaining the same distance away from the forklift.

During the data collection for the Outdoor Data Set, we encounter various types of noises, which include: wind, occasional gravel noises from other vehicles entering and leaving the lot, helicopters, trains, sirens, and traffic.

In addition to the change in environmental setting, the collected waveforms in the Outdoor Data Set are long continuous recordings of the entire data collection session. The reason for this is to collect any random utterances that could be occurring in real situations, and force the shout detection system to respond appropriately. Providing the random utterances will enable us to design an even more robust shout detection system.

Each subject is given the set of possible commands to shout at the forklift, as shown in Table 3.5. The subject shouts a subset of this list of commands, which are chosen randomly.

For each waveform, a reference transcription is created in the same manner as the transcriptions in the other data sets. Using the reference transcriptions, we can evaluate the performance of a particular configuration of the shout detection system, as described in Section 3.2.

Table 3.5: Speech Phrases in Outdoor Data Set. In this data set, we are beginning to explore the feasibility of commanding the forklift.

Forklift stop	Forklift resume
Forklift take this pallet to reception	Forklift go to reception
Forklift take this pallet to queueing	Forklift go to queueing
Forklift take this pallet to bulkyard	Forklift go to bulkyard
Forklift take this pallet to depot	Forklift go to depot
Forklift take this pallet to issue	Forklift go to issue
Forklift take this pallet to storage alpha charlie	Forklift go to storage alpha charlie
Bot go to bravo bravo	Bot go to issue
Bot go to receiving	Bot go to queueing
Bot go to reception	Bot go to the issue area
Bot move pallet to alpha alpha	Bot move pallet to issue
Bot move pallet to storage alpha charlie	Bot move this pallet to issue
Bot put pallet in bravo charlie	Bot put pallet in the receiving area

3.2 System Evaluations

This section discusses the process of evaluating the performance of the shout detection system. First, we review the evaluation metrics that are commonly used in the field of speech recognition, and the modifications we have made to accommodate the application of interest. In evaluating the performance of different speech features, we look at two common techniques: Receiver Operating Characteristics (ROC) and Detection Error Tradeoff (DET). We discuss the method of generating the DET curves for the system, a means of comparing performance curves among different configurations of the shout detection system.

3.2.1 Receiver Operating Characteristic (ROC)

In evaluating the performance of different speech features, we compare their respective Receiver Operating Characteristic (ROC) curves. An ROC curve is a method for visualizing, organizing, and selecting an optimal speech detection algorithm based on its performance. It is used to find a suitable tradeoff between detection rates and false alarm rates of different systems [32].

The performance of a given speech system model can be characterized through the following process. We have two classes: a true class and a hypothesized class. The true class corresponds to the label of each frame of the speech signal; each frame would be classified as either p (for positive speech), or n (for negative speech/non-speech). The hypothesized class presents the classification of each frame, as determined by a speech detection system. To distinguish between the true class and the hypothesized class, we use the labels Y, N for the hypothesized predictions by the speech detection system. For our case, Y corresponds to hypothesized speech, while N corresponds to hypothesized non-speech [32]. With this, there are four possible outcomes.

- **True Positive**: if hypothesized prediction is speech, and there is actually speech
- **False Negative**: if hypothesized prediction is non-speech, and there is actually speech (a missed detection)
- **True Negative**: if hypothesized prediction is non-speech, and there is actually non-speech
- **False Positive**: if hypothesized prediction is speech, and there is actually non-speech (a false alarm)

From this, a two-by-two confusion matrix can be constructed to represent all possible scenarios, as shown in Figure 3-2.

In Figure 3-2, the correct decisions are represented by True Positives and True Negatives, while the errors are represented by False Positives and False Negatives. From this, the detection rate (true positive rate) can be expressed by Equation 3.1.

$$detection\ rate = \frac{True\ Positives}{\#\ speech\ frames} \tag{3.1}$$

The false alarm rate (false positive rate) can be formally expressed by Equation 3.2.

$$false\ alarm\ rate = \frac{False\ Positives}{\#\ non\text{-}speech\ frames} \tag{3.2}$$

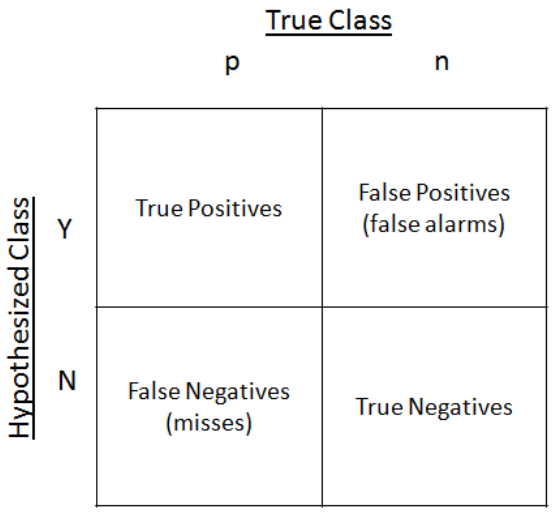

Figure 3-2: General Confusion Matrix [32]

Note that one pair of (false alarm rate, detection rate) corresponds to a given speech detection system configuration, with a set of defined parameters. As we vary a particular parameter value, we can accumulate a set of (false alarm rate, detection rate) pairs and ultimately, generate the ROC curve for the speech system [32].

ROC curves are two-dimensional graphs in which the false alarm rate is plotted on the X-axis, and the detection rate is plotted on the Y-axis. A ROC curve depicts relative tradeoffs between benefits (true positive) and costs (false positives). Figure 3-3 presents a set of basic ROC curves. The red ROC curve represents the strategy of random guessing, meaning that the prediction of the hypothesized class is random. The blue ROC curve presents a speech detection system that performs much better than the strategy of random guessing; for all false alarm rates, the detection rate of the blue ROC curve is equal to or higher than that of the red ROC curve. This trend shows that a speech detection system that produces the blue ROC curve performs better than the random strategy.

While ROC curves are a common metric used to visually evaluate the performance of a speech detection system, comparison of performance between different systems can be difficult. A technique which better distinguishes well-performing systems is Detection Error Tradeoff (DET) [1].

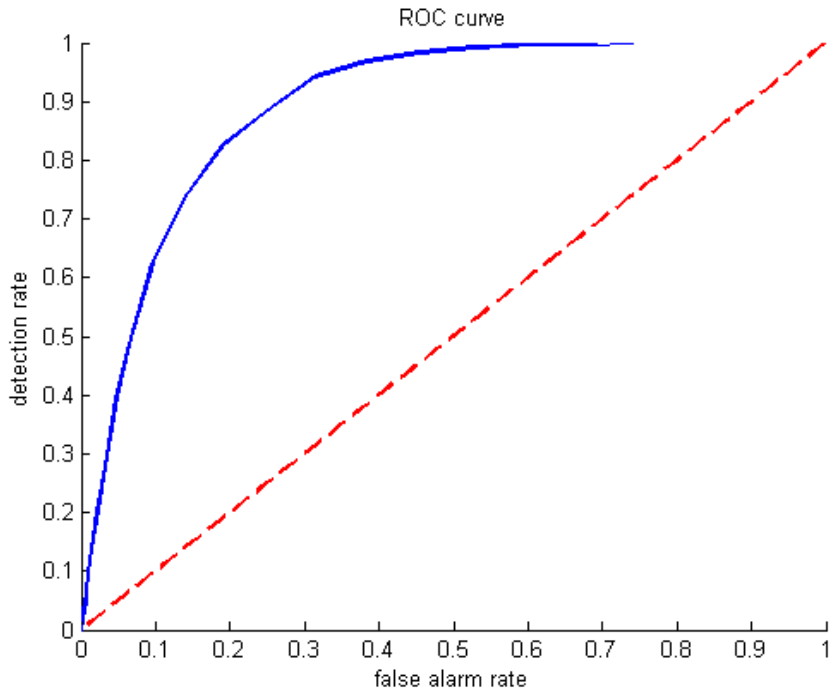

Figure 3-3: Basic ROC curve [32]. The red curve corresponds to a system of random strategy, while the blue curve corresponds to a much better algorithm with higher detection accuracy.

3.2.2 Detection Error Tradeoff (DET)

The ROC curve traditionally has been used to represent the performance and capability of a system. While decision tasks can be viewed as a tradeoff between miss detections and false alarms, the ROC curve generally plots the false alarm rate and corresponding detection rate. A variant of the ROC curve that actually plots the error rates on both axes, giving uniform treatment to both type of errors, is the Detection Error Tradeoff (DET) curve [1].

Calculations of parameters to construct the DET curve are consistent with the parameters of the ROC curve. The DET curve depends on two error types: false alarm rate and miss detection rate (also known as miss rate). The false alarm rate is computed in the same manner as Equation 3.2, while the miss rate can be expressed formally by Equation 3.3 [1].

$$miss\ rate = 1 - detection\ rate = \frac{False\ Negatives}{\#\ speech\ frames} \tag{3.3}$$

Using the false alarm and miss rates, we can generate the DET curves by evaluating each frame (speech/non-speech) of each waveform in a given data set. Figure 3-4 presents an example of the DET curves.

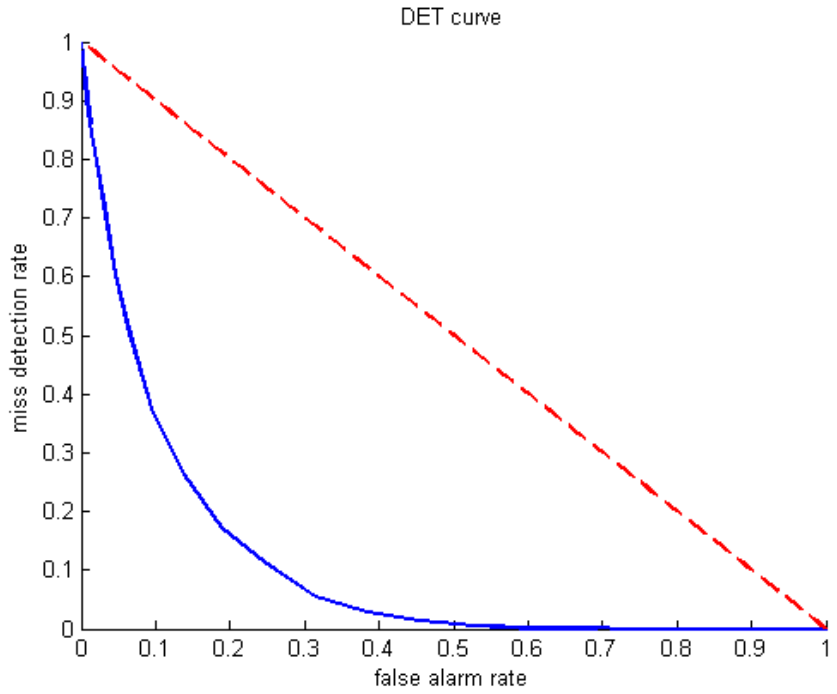

Figure 3-4: Basic DET curve [1]. The red curve is based on a random strategy, while the blue curve is based on a more optimal detection algorithm.

Figure 3-4 presents two DET curves, using the same data as the corresponding ROC curves in Figure 3-3. We can see that the DET curve is simply a vertical flip of the ROC curve (since the Y axis of the DET curve is the complement of the Y axis in the ROC curve). The red DET curve corresponds to the strategy of random guessing of the hypothesized class, while the blue DET curve utilizes the information in the data to generate a better curve. The optimal point is the lower left corner of the plot, so curves that are closer to this region correspond to systems that have a higher classification rate.

While the ROC and DET curves are based on the same data, we feel that the DET

curve better represents the performance on detection tasks by presenting the tradeoff between the two error types (false alarms and miss rate) [1]. For this reason, we will use the DET curves as our primary evaluation metric in evaluating the performance of speech detection systems.

3.2.3 Evaluation Metrics

The evaluation of the system's performance is based on the statistics used to generate the DET curve. In our evaluation metrics, we define the true boundary to be considered detected if the hypothesized boundary lies within one second of the true boundary. *P(miss detection)* is a measure of accuracy of the speech recognition system. While many speech recognition systems prefer to have a high accuracy rate, it is a necessity in our application. Because our application involves a forklift performing operations that can potentially put lives in danger if the system fails to perform correctly, the system must detect all true speech words. Hence, the main goal is to minimize the number of undetected speech words *P(miss detection)*.

false alarms per minute is a measure of the system's performance. In our application of the noise-robust speech recognition system for the forklift, lowering the count of false-alarm speech detections is essential to the forklift's performance. If the recognition system detects everything as speech (both true speech and noise), its performance severely degrades. The system will detect all true speech, but it would interpret all noises as speech and, as a result, the system would serve no purpose. Because of this, it is important that we minimize the *false alarms per minute*.

The two statistics are expressed by Equations 3.4 and 3.5, both in terms of the parameters for the shout detector.

$$P(\textit{miss detection}) = \frac{\textit{undetected true boundaries}}{\textit{true boundaries}} \tag{3.4}$$

$$\textit{false alarms per minute} = \frac{\textit{hypothesized boundaries - detected true boundaries}}{\textit{waveform duration (in minutes)}} \tag{3.5}$$

With these two statistics, we run experiments on the recognizer to measure its performance. The experiments are modelled after Equation 3.6.

$$\log P(W|O) = \log P(O|W) + \log P(W) + WordWeight_W \tag{3.6}$$

The term $WordWeight_W$ is the word weight score added to the log scale of the probability of the word, given the acoustic observation. While varying the word weights, we generate the hypothesized speech segments. We compare the hypothesized speech segments to the reference speech utterances, and determine the corresponding *P(miss detection)* and *false alarms per minute* for the given word weights. By varying the word weight, we can generate a set of points corresponding to the pairs of *P(miss detection)* and *false alarms per minute*. With this, we can plot the curve of *P(miss detection)* vs. *false alarms per minute*, also referred to as the DET curve (as described in Section 3.2.2). In general, a DET curve can be defined for any decision rule that causes the *P(miss detection)* to be uniquely fixed, once the *false alarms per minute* is specified. We can use the DET curve to identify the optimal word weight for our system. More specifically, we can use the DET curve to determine whether modifying the word weight to allow a slightly higher *false alarms per minute* will result in a significantly higher *P(miss detection)*. Ultimately, we aim to have a low *P(miss detection)* with a low *false alarms per minute*. In other words, the optimal curve would be towards the lower left of the plot, as shown in Figure 3-4.

3.2.4 Generating DET Curves

The evaluation of the performance of each feature set consists of two steps: feature computation and classification algorithm (as described in Section 1.3). In running these two processes on the set of waveforms, along with the reference transcription files, we can construct the DET curve to analyze the performance of the system. Since the shout detection system is evaluated on each noise condition, we generate a DET curve for each noise condition. The noise conditions are described in Section 3.1.

The *feature computation* process is a form of dimensionality reduction. Since the

inputs to the speech detection system are sets of waveforms, the input data needs to be transformed into a reduced representation set of features known as the feature vector. For our application, the feature computation process constructs the feature vector of each waveform in the training data set. With a set of waveforms, this process focuses on computing features that correspond to particular characteristics of the signals; examples of these characteristics include energy, entropy, etc. Additionally, new features can be created through modifications of existing features or combination of various features. Lastly, the feature vectors are normalized with respect to the data set for consistency purposes.

After computing the feature vectors, the next step in generating the DET curves is the *classification* process. We use a score threshold to decide whether a given frame is speech or non-speech. This decision produces what is known as the hypothesized transcription. As we compare the hypothesized transcription to the reference transcription, we compute the evaluation statistics, including True Positive, True Negative, False Positive (i.e. false alarm), and False Negative (i.e. missed detection). These statistics are described in Section 3.2.1. The score threshold is a hard threshold that corresponds to the normalized value of the feature. As we vary this threshold, the counts of hypothesized speech and non-speech change accordingly. As the threshold becomes less lenient, we expect the number of speech segments to increase (and hence, the number of non-speech segments to decrease). Since the speech and non-speech segment counts correspond to a particular evaluation, we essentially generate the DET curve with the set of computed statistics.

3.3 Summary

In this chapter, we describe the data sets that are used in this thesis, along with their collection processes. We also formulate a means of evaluating our system through computing the *P(miss detection)* and *false alarms per minute*, generating a DET curve for each configurations of the system. In the next chapter, we investigate different features to develop the optimal shout detection algorithm for our application. Using

the frame-based energy as a baseline feature, we measure the performances of other features through comparisons of DET curves.

Chapter 4

Experiments

This chapter focuses on extracting different features from recordings of shout utterances in various noise environments. We use an energy-based speech activity detection [26] as our baseline system. Since the goal of the experiments is to determine the set of features that best detect shouted speech, we evaluate the quality of each feature through speech detection experiments. By comparing the performance of the developed features to the frame-based energy feature, we aim to determine the effectiveness of each feature as a shout indicator.

4.1 Frame-based Energy

Energy has been one of the most commonly used features for speech detection [12], and an energy-based detector is used as the baseline for performance comparisons. To measure the performance of using energy as a shout feature, we look at how accurately the system is able to locate the frames of shouted speech.

4.1.1 Methodology

For each audio sample in the data set, we evaluate the system by comparing the hypothesized transcription to the reference transcription. We use the gstSAD [26] to produce the start and end time stamps of the detected speech segments, using

frame-based energy as its feature. From the output of the gstSAD, we generate the hypothesized transcription for a particular audio sample. We repeat the same procedure for the entire Simulated Shout Data Set.

The gstSAD is implemented as an independent module, where the input is a stream of audio and an energy threshold. Given the energy threshold, the gstSAD classifies each frame as either speech or non-speech. The decision smoothing stage processes the initial decisions to remove indications of activity which are too short to be speech. It also appends short periods of signal (duration of 250 ms or less) prior to and following active segments to ensure that the boundaries of the speech are not incorrectly eliminated. After the decision smoothing process, the gstSAD module outputs the start and end time stamps of detected speech segments. The segments of each individual waveform are compiled into the hypothesized transcription.

The gstSAD uses the energy in each frame to classify the frame as either speech or non-speech. The classification is based on a strict energy threshold setting. If the frame energy passes the energy threshold, the frame is considered a potential speech segment. Likewise, the frame is considered non-speech if its frame energy is below the set threshold. With this classification, the variation of the energy threshold can ultimately change the detection rate and the false alarm rate. With a lower energy threshold, there would be a higher detection rate but more false alarms. On the other hand, a higher energy threshold translates to fewer frames passing such threshold and, as a result, there would be fewer false alarms and more missed detections. We can see that there is a tradeoff between detection and false alarms as the energy threshold varies.

For this experiment, we vary the energy threshold with a 5 dB increment in the -90 dB to -5 dB range, and an increment of 1 dB in the -5 dB to -35 dB range. Having a finer increment in the energy threshold in this range allows us to generate a smoother DET curve. In total, we have 43 energy thresholds.

4.1.2 Results and Discussion

We run the Simulated Shout Data Set through the gstSAD module as we vary the energy threshold, and generate the DET curve for each of the seven noise conditions (as shown in Figure 4-1). With the DET curves, we can completely describe the detection error rates of the system with energy as its baseline feature.

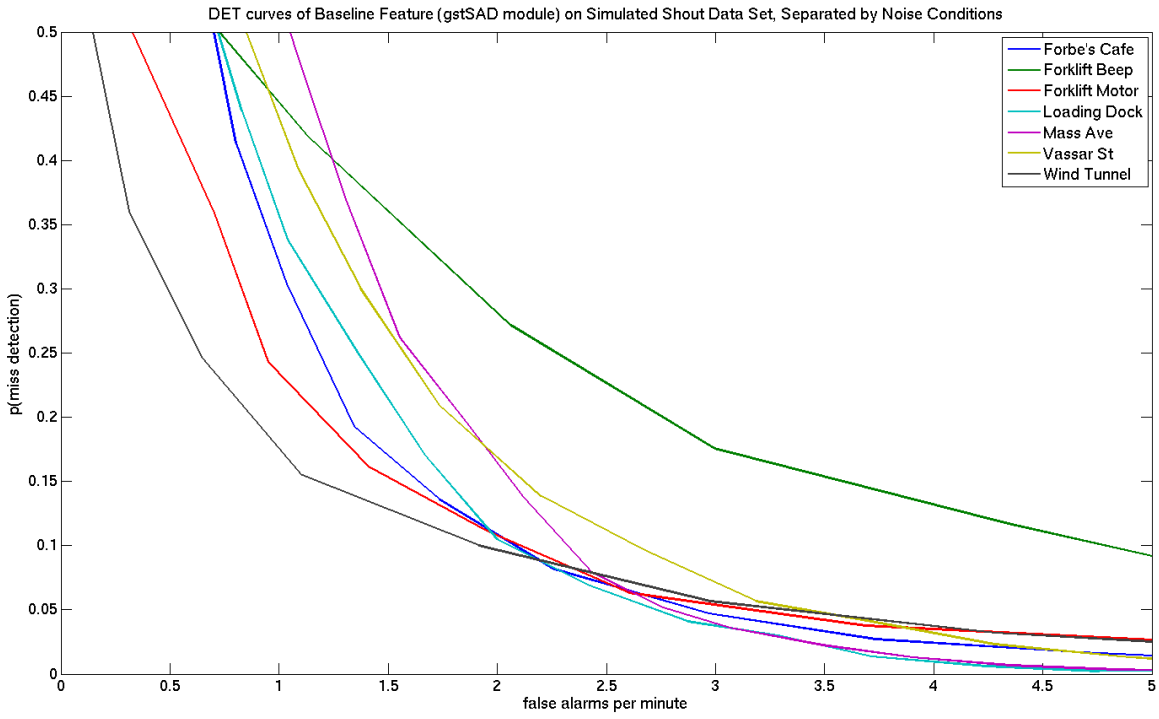

Figure 4-1: DET curves of Energy-based Speech Activity Detection (gstSAD) on the Simulated Shout Data Set (separated by noise condition). Note that the x-axis is the false alarms per minutes, ranging from 0 to 5. The y-axis is the probability of miss detection *P(miss detection)* from 0 to 0.5.

Figure 4-1 presents the DET curves of the gstSAD module to highlight the performance of the energy-based detection system. In comparing the different noise conditions, the forklift beeping noise condition performs the worst. One explanation for this inconsistency is the similarity in energy level between forklift beeping and shouted speech, making it very challenging to classify frames based on a hard energy threshold. On the other hand, the noise energy in the other noise environments are lower than the energy from the shouted speech.

4.1.3 Summary

For this thesis, we use the frame-based energy feature (the gstSAD module) as the baseline feature. We measure the performance of the other features based on how they compare to the baseline feature. In generating the DET curves from the Simulated Shout Data Set, we set the baseline performance for each noise condition. The results show that the system performs the worst in the Forklift Beep condition. To solve the issue of the forklift beeping noise, we use signal processing to reduce the effect of the noise. In the next sections, we investigate various techniques to reduce the effects of some types of noise.

4.2 Notch Filter

In analyzing the different noise conditions, we notice that the forklift beeping noise is very consistent; it always occurs at a particular frequency. As a speech enhancement technique, we implement a Butterworth notch filter to eliminate that frequency with the smallest bandwidth possible. We hope to eliminate only the forklift beeping noise from the audio data set with minimal disruption to the speech component of the signal.

4.2.1 Methodology

As a test implementation of the notch filter, we determine the frequency and its corresponding bandwidth of the forklift beeping noise through spectral analysis. We analyze various audio samples in the forklift beeping; these samples are taken from the Forklift Data Set. We determine that the beeping noise signal is indeed very consistent across the different samples collected, and its frequency is approximately 1380 Hz with a bandwidth of 200 Hz. Using this information, we implement a Butterworth notch filter.

The Butterworth notch filter is a band-stop filter that passes most frequencies unaltered, but attenuates those in a specific range to very low levels. Figure 4-2

presents the magnitude response of the notch filter.

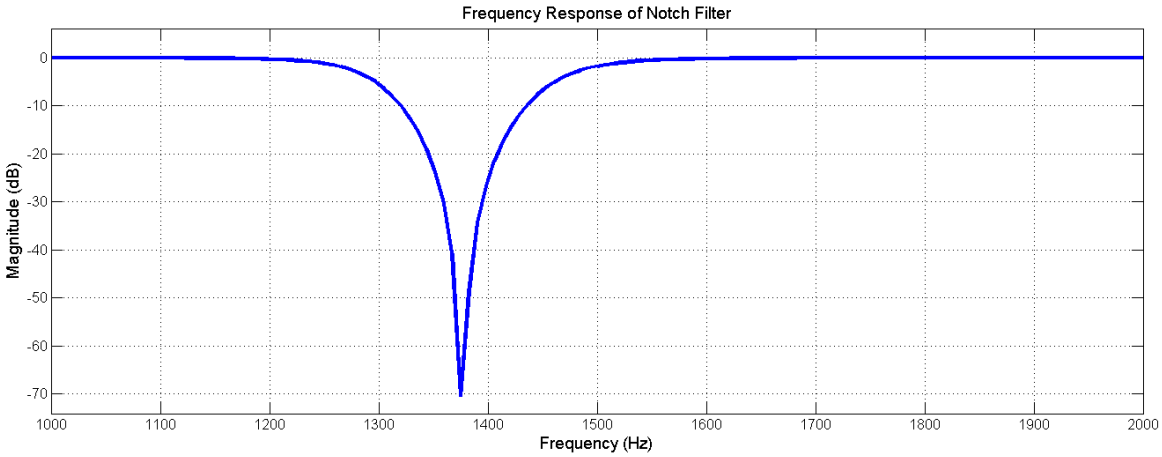

Figure 4-2: Magnitude response of the notch filter. The filter has a narrow band to eliminate the frequency around 1380 Hz, which is the frequency of the forklift beeping noise.

Given the presented design, we apply the notch filter to an audio sample to measure the effectiveness of the filter in reducing the forklift beeping noise. Figure 4-3 shows a spectral analysis of an audio sample from the Forklift Data Set (from the Forklift Beep noise condition) before and after processing it through the implemented notch filter.

From Figure 4-3, we can see that the frequency of 1380 Hz represents the forklift beeping noise. Through the implemented notch filter, we are able to completely eliminate that frequency, resulting in the bottom spectrogram. In this spectrogram, the blue horizontal line shows that the beeping frequency is no longer present in the signal. Although the beeping noise is clearly eliminated, one of our main concerns is how the speech component is affected. Since we eliminate a particular range of frequencies across the entire audio sample, we are concerned whether recognition of the speech utterance is compromised.

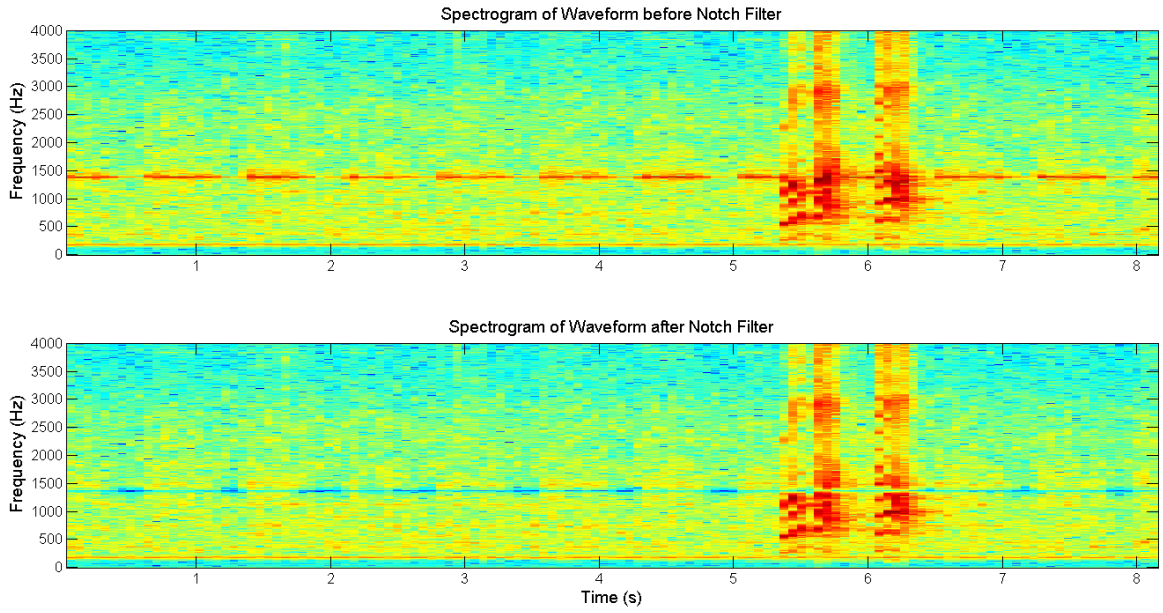

Figure 4-3: Audio sample (from the Forklift Data Set) before and after notch filter. The top spectrogram represents the waveform before the notch filter, while the bottom spectrogram represents the waveform after processing it through the notch filter. We can see that the notch filter completely removes the forklift beeping noise from the audio sample. Note that the horizontal band in the spectrogram changes from red (on the top spectrogram) to blue (on the bottom spectrogram).

4.2.2 Results and Discussion

We repeat the process as described in Section 4.1.1 with a Butterworth notch filter as a pre-processing unit for our shout detection system. Using the same methodology as mentioned in Section 4.1.2, we generate the DET curves of the baseline system with the notch filter for each noise condition.

Figure 4-4 displays the DET curves of the two configurations of the baseline shout detection system, showcasing the effect of the notch filter in different noise environments. In most noise conditions, there is very little change in the DET curves from using the notch filter. However, the notch filter results in a drastic improvement for the Forklift Beep condition (DET curve (b)). Eliminating the specified frequency bands results in removing the dominating noise in the entire data waveform. Since the Forklift Beep condition contains consistent forklift beeping noise as its background noise, applying the notch filter removes the beeping without interfering with

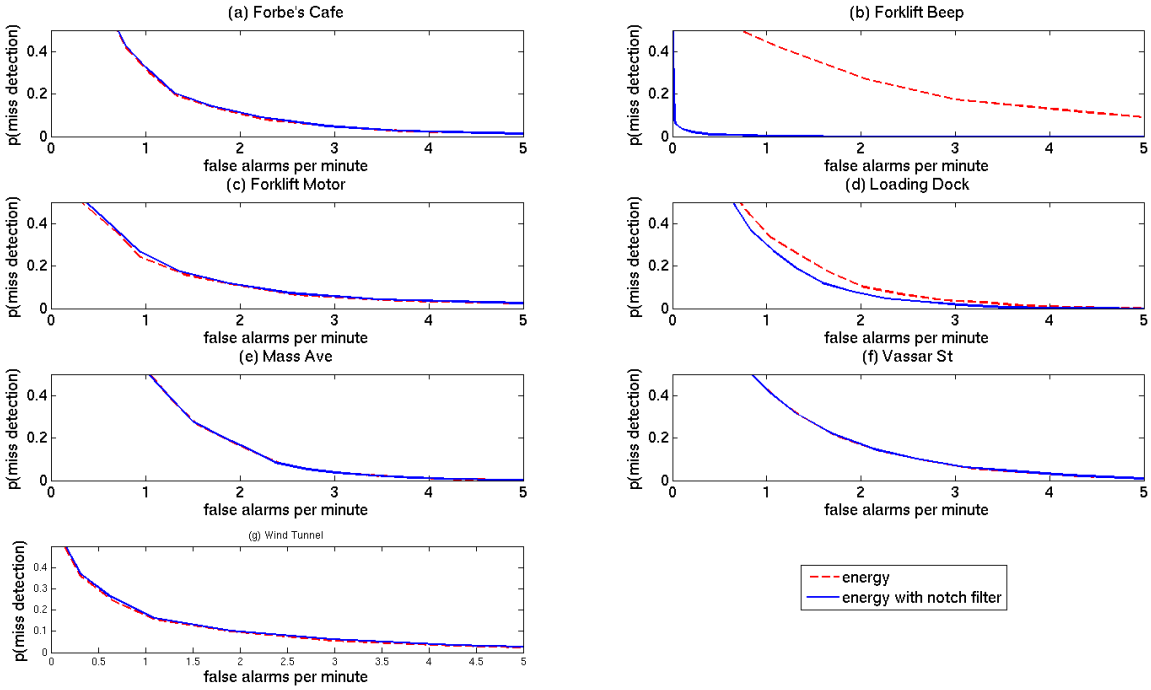

Figure 4-4: DET curves of the baseline shout detection system on the Simulated Shout Data Set to show the effect of notch filter on each noise condition. The notch filter drastically improves the Forklift Beep noise condition. At the same time, the new shout detection system maintains a consistent level of performance for the other noise conditions.

the speech signal in the waveform. Additionally, the Loading Dock noise condition contains various instances of beeping noises, so applying the notch filter improves the detection accuracy of the system in this noise condition. Using this notch filter, we can see improvements in the system's performance towards DET curves (b) and (d) while maintaining a consistent level of performance in the other noise conditions.

4.2.3 Summary

In an attempt to address the noisy environments in which the autonomous forklift operates, we implement a notch filter as a speech enhancement to eliminate forklift beeping noises. Through spectral analysis, we identify the frequency range of the forklift beeping noise, and implement a notch filter to eliminate that particular fre-

quency band. Spectral analysis shows the effectiveness of the notch filter, and the DET curve shows that the notch filter translates to significant improvements in the performance of the shout detection system. In viewing Figure 4-4, we can see that the notch filter offers improvements in conditions with forklift beeping noise (e.g. Forklift Beep and Loading Dock noise conditions).

While the forklift beeping noise is very consistent, other devices or vehicles with beeping noises have different frequencies, so this notch filter will not eliminate all beeping noises. To address this concern, we look at other speech enhancement techniques in an effort to find ones that will offer significant contributions to the other noise conditions.

4.3 Spectral Subtraction

A common signal processing technique used to reduce added background noise is spectral subtraction. The general spectral subtraction algorithm is based on the assumption that the power spectrum of a signal corrupted by uncorrelated noise is approximately the sum of the signal spectrum and the noise spectrum. More formally, this assumption can be expressed as follows: let the input signal be represented as $x = s + n$, where s is the speech signal and n is the noise. In the frequency domain, $X(k) = S(k) + N(k)$, and $|X(k)| \approx |S(k)| + |N(k)|$. The goal of spectral subtraction is to find $Y(k)$ such that $|Y(k)| \approx |X(k)| - |N(k)|$, so that $Y(k) \approx S(k)$.

We incorporate the proposed spectral-subtraction-based algorithm as a speech enhancement technique into our shout detection system. We evaluate the system on all noise conditions of the Simulated Shout Data Set. We use the implementation provided by the VOICEBOX Speech Processing Toolbox for MATLAB [25]. This tool takes in waveforms as inputs, as well as the sampling frequency of the signal. A waveform is produced after processing it through the spectral subtraction algorithm.

4.3.1 Methodology

To evaluate the effectiveness of spectral subtraction in enhancing the speech signal, we analyze a sample waveform before and after spectral subtraction. Figure 4-5 shows the input waveform and its corresponding spectrogramof three speech utterances in an outdoor street environment (noise condition is Mass. Ave. of Simulated Shout Data Set). The top set of plots represent a waveform without any noise, while the other two sets of figures represent the waveform before and after spectral subtraction. We compare the top set of plots to the bottom set to visually see how effective spectral subtraction is in eliminating the background noise.

In viewing Figure 4-5, we can evaluate the effectiveness of spectral subtraction in reducing the background noise in the sample waveform. In viewing the second plot (the spectrogram of the original waveform), we can see that the speech signal has been contaminated by the background noise, resulting in difficulty in distinguishing the speech components of the signal. In the spectrogram, the background noise is represented in red and yellow, while the speech is shown in red. We can see that the speech and background noise are very integrated, making it difficult to locate the speech boundaries. From the third and fourth plots, we can see that spectral subtraction is able to eliminate the majority of the background noise from the waveform. Particularly, the reduction of the background noise results in a clearer distinction in the speech boundaries.

After processing the original waveforms using spectral subtraction, we generate the corresponding DET curves for each noise condition. Like the notch filter, we use the shout detection system with the baseline feature. By generating the DET on the processed waveform, we can measure the performance of spectral subtraction on our application.

4.3.2 Results and Discussion

We measure the performance of the shout detection system with spectral subtraction by comparing it to (1) the shout detection system without any noise enhancements

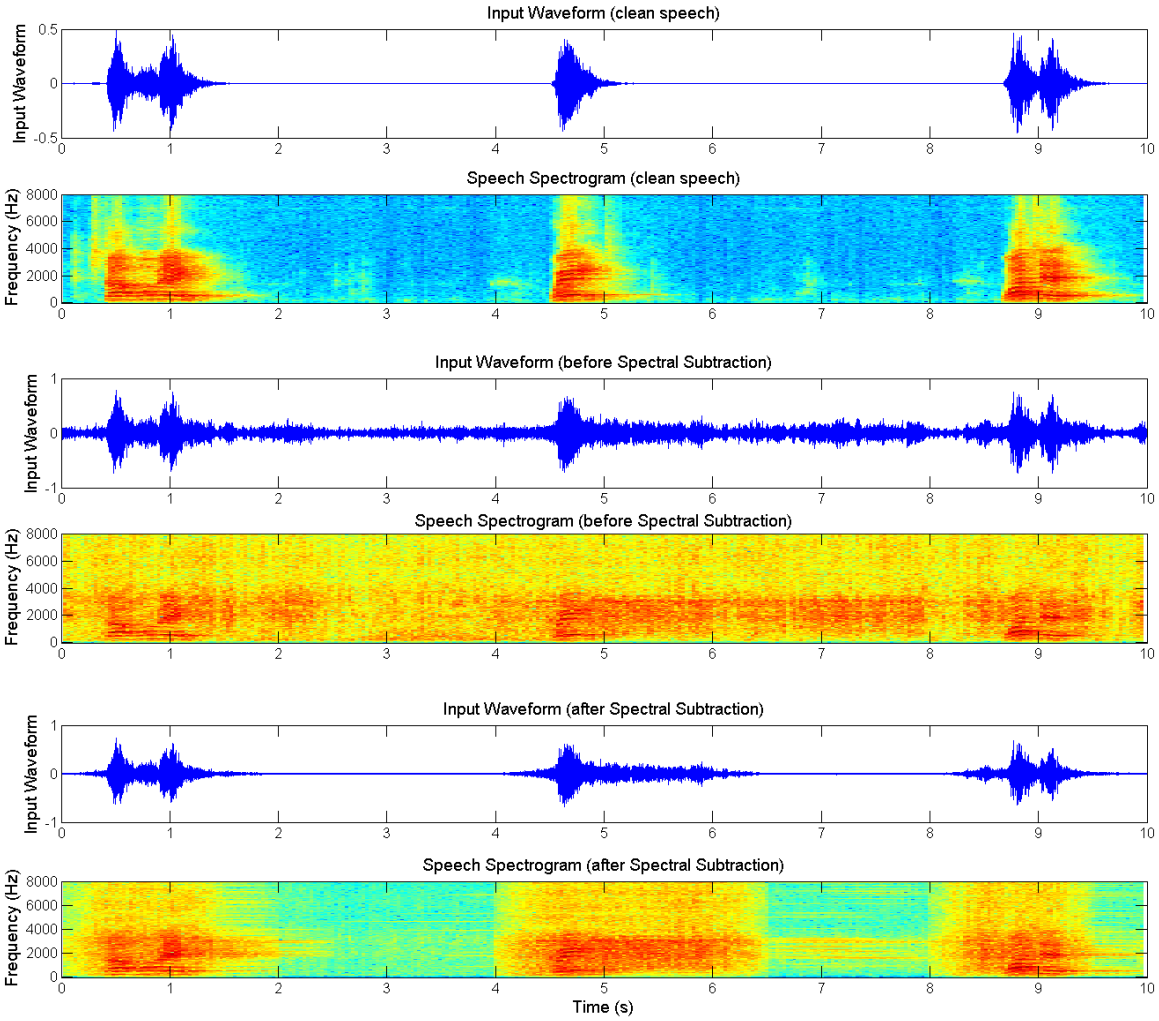

Figure 4-5: Comparison of a sample waveform before and after spectral subtraction, along with their respective spectral analysis. The sample waveform collected from an outdoor environment with an SNR level of 10 dB. In comparing the top set of plots (only clean speech) to the bottom set of plots (waveform after spectral subtraction), we can see that the speech enhancement technique is able to reduce much of the background noise, providing a clear distinction of the speech components.

and (2) the shout detection system with notch filter. First, we process the waveforms from the Simulated Shout Data Set through spectral subtraction. Using the procedure as described in Section 3.2.4, we generate the DET curves for each noise condition. The DET curves are shown in Figure 4-6, along with the DET curves of the other two discussed modifications of the shout detection system.

Figure 4-6 presents the DET curves of the three different configurations of the shout detection system. For this experiment, the baseline system is the frame-based

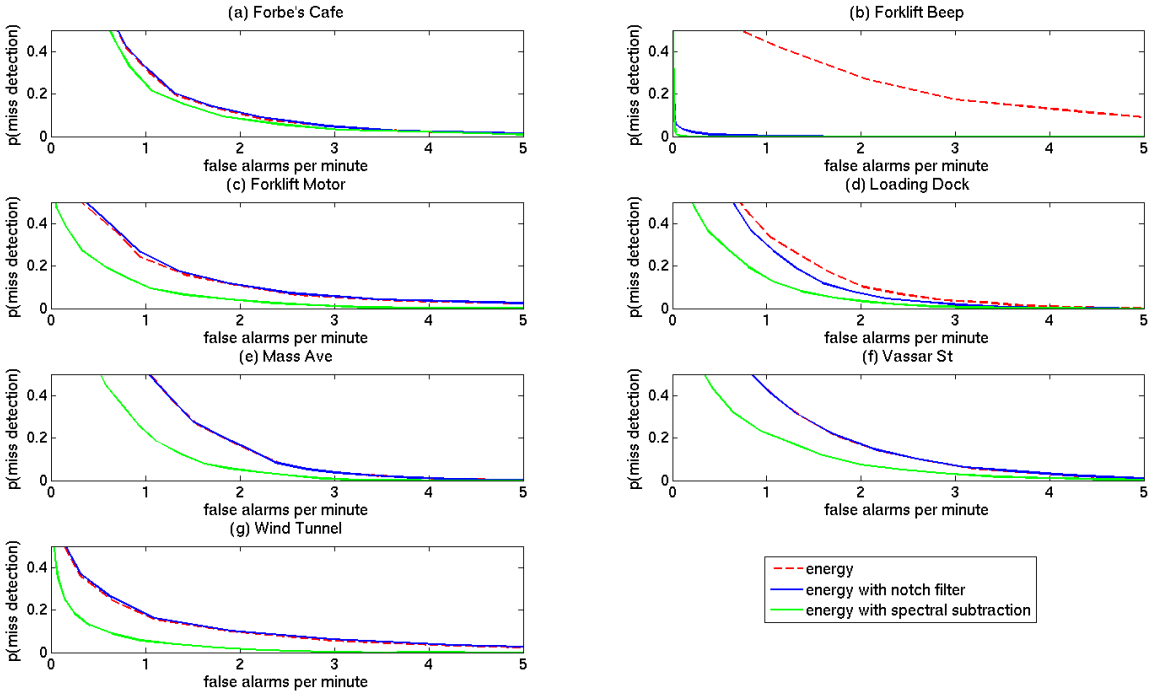

Figure 4-6: DET curves of the baseline shout detection system on the Simulated Shout Data Set to compare the various speech enhancement techniques. In all noise conditions, the system with spectral subtraction (in green) outperforms the system with notch filter (in red) and the baseline system with no speech enhancements (in blue).

energy feature without any speech enhancement. We compare the DET curve of the shout detection system with spectral subtraction to this baseline to measure the performance improvement that the speech enhancement has provided. At the same time, we plot the DET curves of the shout detection system with notch filter to compare its effectiveness against the spectral subtraction.

The presented figure shows the significant improvements that spectral subtraction offers. The DET curve of the shout detection system using spectral subtraction outperforms the other two approaches across all noise conditions. Particularly, we see major improvements in the following noise conditions: (c) Forklift Motor, (d) Loading Dock, (e) Mass. Ave., and (f) Vassar St.I n analyzing the common traits among these noise conditions, we notice that there are many instances of constant

background noise, and the spectral subtraction algorithm is designed to be effective for this type of noise. Through the DET curves, the spectral subtraction is able to reduce the effect of the non-stationary noise on the speech signal and, ultimately, improve the performance of the shout detection system.

4.3.3 Summary

Spectral subtraction is a common speech enhancement technique because it has proven to be very effective in minimizing the effects of background noises in our speech signals. In evaluating the spectral subtraction on our current shout detection system, we see many improvements when we compare the DET curves to the baseline system (no speech enhancement) and the system with the notch filter. The system with spectral subtraction outperforms the other two configurations across all noise conditions from the Simulated Shout Data Set. The effectiveness of spectral subtraction in this experiment shows that it greatly contributes to the accuracy of the shout detection system.

However, as we investigate other features, we will refrain from including the spectral subtraction as a speech enhancement technique. This way, we can measure the performance of each feature by itself.

4.4 Relative Spectral Entropy (RSE)

While energy is a common feature for speech detection, it has many limitations, particularly in environments with unexpected background noises [24]. Spectral entropy, a measure of the expected information, can address these limitations. The authors in [19] experiment with spectral entropy on endpoint detection, and find that voiced spectral entropy is significantly lower than non-voiced entropy. Because of this distinction, entropy-based detection algorithms are reliable for endpoint detection in many noisy environments (particularly in the presence of stationary noises). Additionally, relative spectral entropy (RSE), which is a KL divergence between the current spectrum and the mean spectrum, acts to enhance robustness in environments with a

constant voicing source [30]. We look at how RSE performs on the Simulated Shout Data Set (described in Section 3.1.1) to measure the performance of using RSE as a shout feature.

4.4.1 Methodology

For this experiment, the RSE is computed using the process described in Section 2.4.2. Particularly, we use a frame length of 30 ms, frame shift length of 10 ms, and FFT-points of 1024. Additionally, we set the number of frames in the shifting segment M to be 50 frames. These parameters are suggested by the authors in [2] and tested in our application for optimality.

With the mentioned parameters, we generate the DET curves to measure the performance of RSE in the Simulated Shout Data Set. The RSE feature is trained and tested under each noise condition, and generates a corresponding DET curve. Each DET curve reflects the performance of the feature over all subjects and SNR levels.

4.4.2 Results and Discussion

With the generated DET curves, we compare the performances of RSE to our baseline feature, as shown in Figure 4-7.

Figure 4-7 showcases the DET curves of RSE and the baseline in different noise conditions. As mentioned in Section 3.2.4, the optimal location for the DET curve is in the lower left corner, where the *P(miss detection)* and *false alarm per minute* are low. With this in mind, we can see that the RSE feature performs better than the baseline feature only in certain noise conditions; those noise conditions include (b) Forklift Beep, (c) Forklift Motor, and (g) Wind Tunnel. In (a) Forbe's Cafe, the baseline feature outperforms the RSE. In the remaining noise conditions, the performance of the two features varies, depending on the operating point of the system on the DET curve.

The performance trend shown in Figure 4-7 is consistent with our understanding of

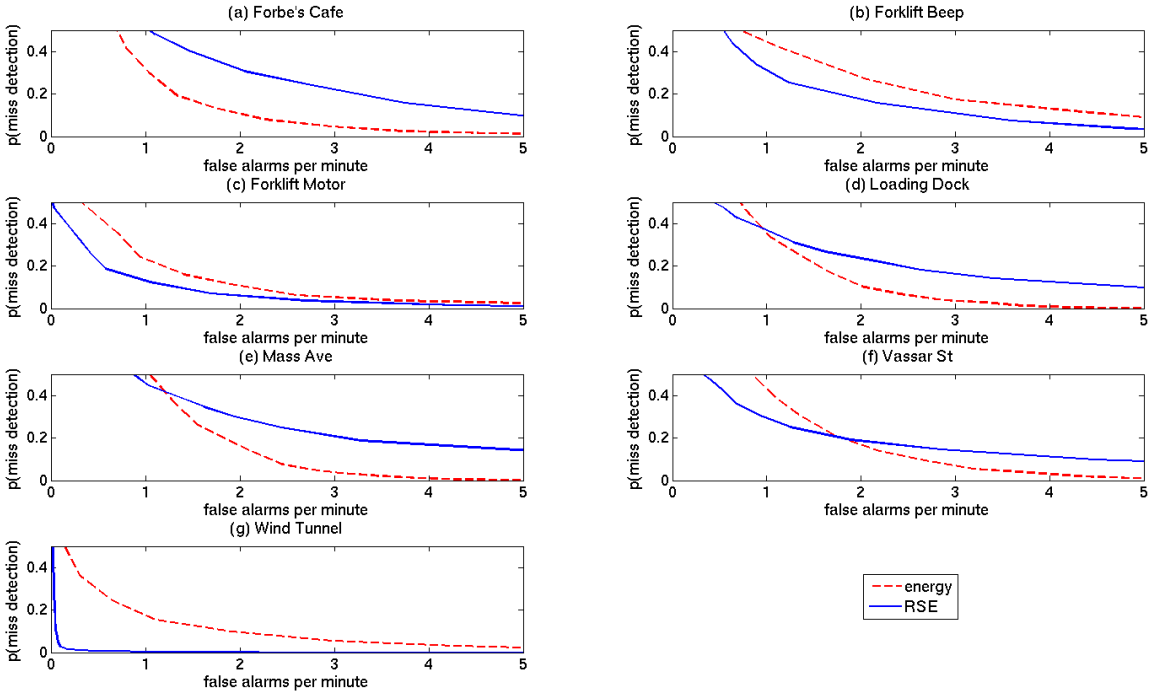

Figure 4-7: DET curves of baseline feature (Frame-based Energy) and relative spectral entropy (RSE) for each noise condition. In plots (b), (c), (g), we can see that RSE outperforms the baseline feature. However, the baseline feature exceeds the performance of RSE in plot (a). For the other noise condition, the performance of the two features depends on the operating point on the DET curves.

RSE as a shout detection feature. RSE is able to detect speech well in stationary noise environments. Since speech regions in general have lower entropy than non-speech segments, having a constant background noise should not influence the ability of RSE to discriminate speech from the stationary noise. This attribute is one advantage that RSE has over the baseline feature, and hence, we see that RSE performs better than the baseline features in noise environments with consistent noise. However, the RSE fails to perform well in non-stationary noise environments. We can see (from plot (a) of Figure 4-7) that the baseline feature performs better than the RSE. Lastly, we can see that the two features have similar performance in the remaining noise conditions. Because the DET curves of the two features cross each other, the relative performance of the two features depends on where the operating point is located. Therefore, for

the noise conditions (d) Loading Dock, (e) Mass. Ave., and (f) Vassar St., the RSE feature performs better than the baseline feature only on certain operating points of the DET curve.

4.4.3 Summary

In addition to energy, entropy is another common feature used in speech detection. For this thesis, we look at a variation of entropy: the relative spectral entropy (RSE). By computing the mean spectrum from neighboring frames, RSE has shown to be very useful in situations where the speech signal is contaminated with a constant noise source [2]. In comparing the DET curves of the RSE feature to the baseline feature (see Figure 4-7), we can see that the RSE feature performs better only in certain noise conditions. In the next section, we present a method of combining the two features in an effort to utilize the benefits of each feature.

4.5 Energy-Entropy: Fusion of the Two Features

The Energy-Entropy (EE) feature combines the features of energy and entropy into a new form of speech detection feature, as described in Section 2.5. It utilizes the advantages of each individual feature while compensating for their limitations. The EE feature is more reliable than pure energy-based and entropy-based methods, making it more tolerant to a wide variety of noises. The process of computing the EE feature emphasizes the speech region of the signal while attenuating the non-speech segments, making the overall approach well suited for endpoint detection tasks [22]. With this approach, we are interested to see how the EE feature performs on the Simulated Shout Data Set and ultimately, in the forklift application.

4.5.1 Methodology

In measuring the performance of the EE feature as a speech detection feature, we generate the DET curve of the feature on the Simulated Shout Data Set and compare

it to our baseline feature. The EE feature is computed using the method as described in Section 2.5.

To further improve the effectiveness of the overall feature, we look into the way the entropy is computed. In viewing Figure 2-5, we see that the entropy is sensitive to the forklift beeping noise; the feature detects the forklift beeping noise as speech. We extend the current design of the EE feature by replacing the spectral entropy with RSE. We refer to the new feature as Energy-Relative Spectral Entropy (ERSE).

The calculation of the ERSE feature is consistent with that of the EE feature; the difference lies in the calculation of entropy. For the ERSE feature, the entropy component is calculated as described in Section 4.4. The effectiveness of the new feature is shown in Figure 4-8.

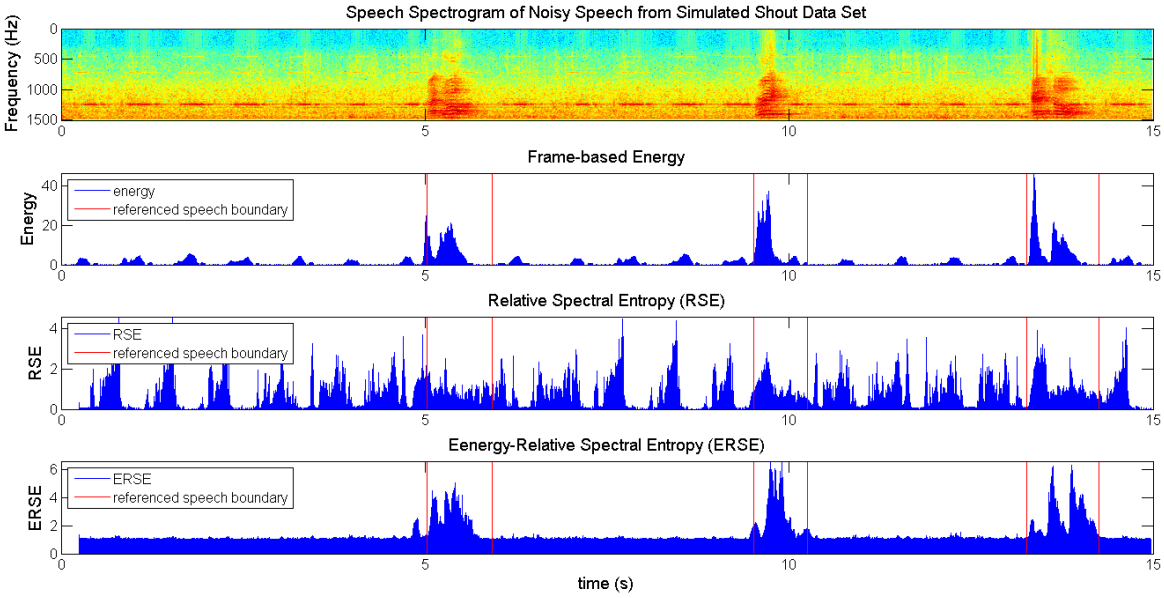

Figure 4-8: Energy-Relative Spectral Entropy (ERSE) of the three utterances of speech in a noisy environment. The noise condition in this signal is Forklift Beep. By using the RSE feature, a more effective mean of calculating the entropy of the signal, background noises in the signal has been dramatically reduced, showing the effectiveness of the new feature.

Figure 4-8 uses the same waveform as Figure 2-5 to show the effectiveness of the ERSE feature on three utterances in a forklift beeping environment. We can see (from the third plot of Figure 4-8) that the new method of computing entropy has

minimized the effect of noise on the feature. As a result, the ERSE feature shows very minimal signs of noise, and is able to clearly detect the speech boundaries in the waveform.

From generating the spectrogram plots and the responses of the developed features, we can see that the ERSE feature is able to identify the speech segments of the signal by having very distinctive values for speech regions. The next step is to generate the DET curve of the feature to compare it to the baseline feature. We generate the DET curves to measure the performance of ERSE on the Simulated Shout Data Set. The process of generating the DET curves, as described in Section 3.2.4, uses ERSE as the only feature. The feature is trained and tested under each noise condition, and a corresponding DET curve is generated. Each DET curve reflects the performance of the feature over all subjects and SNR levels.

4.5.2 Results and Discussion

Using EE and ERSE as shout features, we generate the respective DET curves of the shout detection system. In comparing these DET curves to the curves of RSE and the baseline feature, we can measure the effectiveness of the EE and ERSE as shout features. Figure 4-9 presents the DET curves of the four features in different noise conditions.

Figure 4-9 presents the DET curves for the four features in different noise conditions. We can see that in all noise conditions, the feature with the best performance (in terms of most optimal DET curves) is either ERSE or RSE. For conditions such as Forklift Beep, the ERSE feature outperforms the other features. Likewise, the RSE feature performs the best under the Forklift Motor noise condition. For the other noise conditions, choosing the best feature depends on the operating point. For the other noise conditions, we can see that the DET curves cross paths, suggesting that no single feature dominates in that particular noise condition.

In this experiment, we expect the ERSE feature to perform the best in the various noise conditions, since the feature is designed to utilize the strong points of the energy and RSE features. Hence, the ERSE feature should outperform the other features

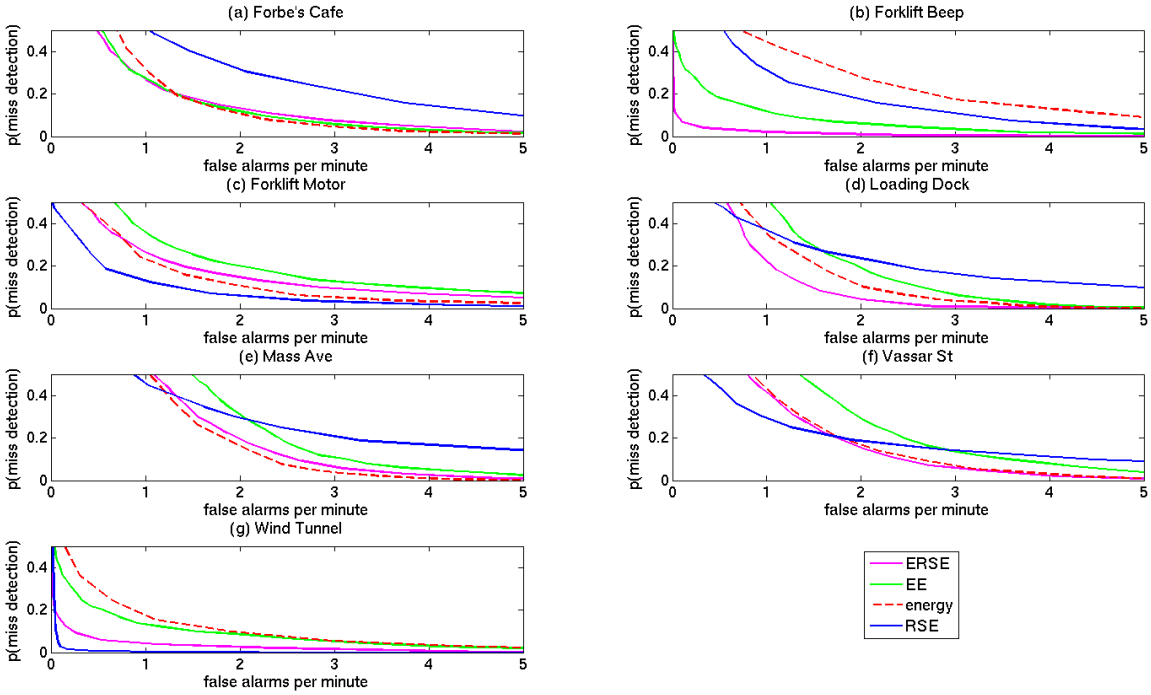

Figure 4-9: DET curves of following features for each noise condition: energy-relative spectral entropy feature (ERSE), energy-entropy feature (EE), baseline feature (frame-based energy), and relative spectral entropy feature (RSE).

by having the most optimal DET curve. At the very least, the ERSE feature should maintain the same level of performance as the other features. However, as shown in Figure 4-9, the ERSE feature is not the most optimal feature in all noise conditions.

4.5.3 Summary

While energy and entropy perform well in certain noise conditions, their abilities to locate speech in other environments are limited. By combining the two features to form the Energy-Entropy (EE) feature, we utilize the advantage of each individual feature while compensating for their limitations. Furthermore, we improve the performance of the feature by replacing the conventional entropy with the RSE feature, forming ERSE. This replacement provides a more robust feature towards environments with non-stationary noises. In comparing the DET curves of the ERSE feature

to the other features (see Figure 4-9), we can see that the ERSE feature performs better in most of the noise conditions from the Simulated Shout Data Set. In the next section, we aim to further improve the performance of the ERSE feature by incorporating spectral subtraction and enhancing the speech signals of the waveforms before computing the feature.

4.6 Robust Shout Detection Algorithm

With the results and knowledge gathered about the investigated features and speech enhancement techniques, this section presents the Robust Shout Detection Algorithm (RSDA). We aim to develop the most optimal algorithm for the autonomous forklift application. The final algorithm is compared to the baseline system and the other configurations that we have presented in the previous chapters. We measure the performance of the algorithm through analyzing the generated DET curves.

4.6.1 Methodology

The algorithm consists of two components: speech enhancement and feature extraction, as described in Figure 4-10.

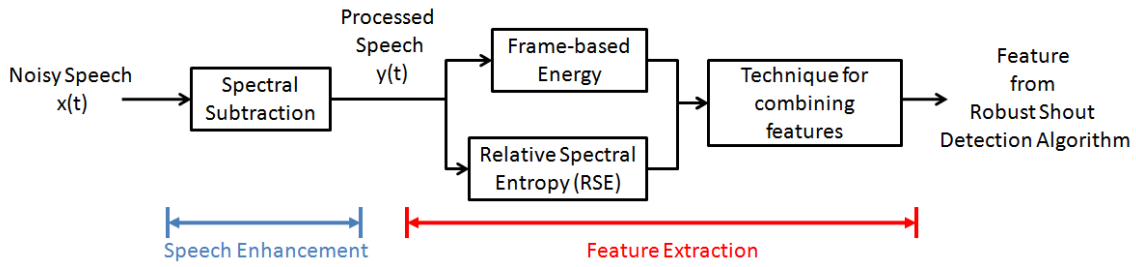

Figure 4-10: Block Diagram of the Robust Shout Detection Algorithm (RSDA). The algorithm consists of two components: speech enhancement and feature extraction. Each component is derived from previous chapters, where we empirically created the optimal component.

Figure 4-10 showcases the RSDA. First, the input waveforms are processed through spectral subtraction (as described in Section 4.3). Afterwards, we extract the features

from the processed waveforms. The feature we extract is the ERSE, a fusion of frame-based energy and relative spectral entropy (RSE). The methodology and parameters of the ERSE are consistent with those described in Section 2.5. After extracting the final feature of the system, the speech/non-speech discrimination can be made through a hard threshold. In other words, a classification for a particular frame of the input waveform is made based on whether the final output of the RSDA is above or below a certain threshold. This result marks the final output of the shout detection system.

To measure the effectiveness of the RSDA, we visually analyze the response of each step of the algorithm to a sample waveform. Figure 4-11 showcases the output of each step of the RSDA: the processed waveform after the spectral subtraction, the feature outputs of the frame-based energy, the RSE, and the ERSE (the final output of the algorithm).

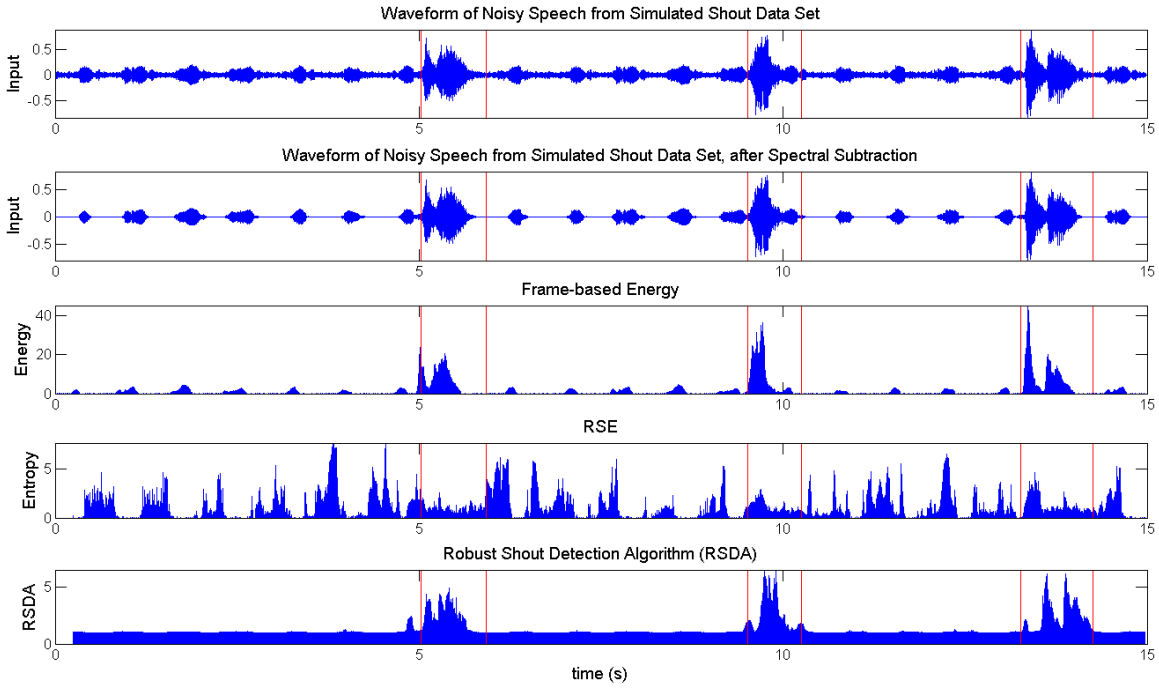

Figure 4-11: Plots of each step's response in the Robust Shout Detection Algorithm (RSDA). For these plots, the test waveform, with SNR of 10 dB, is based on a forklift operating environment, with frequent forklift beeping noise. In the final output (the bottom plot), we can see that the feature is able to easily distinguish the speech components from the background noises in the original signal.

In Figure 4-11, the top plot presents the test waveform, which consists of three utterances in a forklift operating environment (noise condition is Loading Dock). For this waveform, the SNR level is 10 dB, so speech is the dominant signal. The second plot shows the waveform after spectral subtraction. We can see that this reduces much of the background noise, except the forklift beeping noise. The third plot presents the response of the frame-based energy feature, while the fourth plot shows the response of the RSE feature. Finally, the bottom plot shows the response of combining the two features together, forming the output of the RSDA.

In analyzing the output of each process of RSDA, we can see that incorporating the spectral subtraction into the process of feature computation improves the feature's ability to locate the speech components in the test waveform. While the ERSE feature is able to utilize the strong points of both features, the spectral subtraction eliminates a large portion of the background noise from the test waveform, reducing the chances of the feature to interpret any background noise as speech.

Given this setup of the RSDA, we compare the performance of the algorithm to the other mentioned features by generating the DET curves for each noise condition in the Simulated Shout Data Set.

4.6.2 Results and Discussion

To measure the performance of the algorithm in our shout detection system, we generate the DET curves of the RSDA for all noise conditions. We compare the DET curves to that of the baseline system (frame-based energy), along with the optimal feature without speech enhancement (ERSE). In generating the DET curves, we use the Simulated Shout Data Set. More information on the data collection process of this data set can be found in Section 3.1.1. Figure 4-12 presents the DET curves of the various configurations of the shout detection system.

Figure 4-12 presents the DET curves of the RSDA, in comparisons to the baseline system and the most optimal feature without speech enhancements. In comparing the RSDA to the baseline system, we see much improvement across all noise conditions. Particularly, the RSDA shows the greatest improvements for Forklift Beep

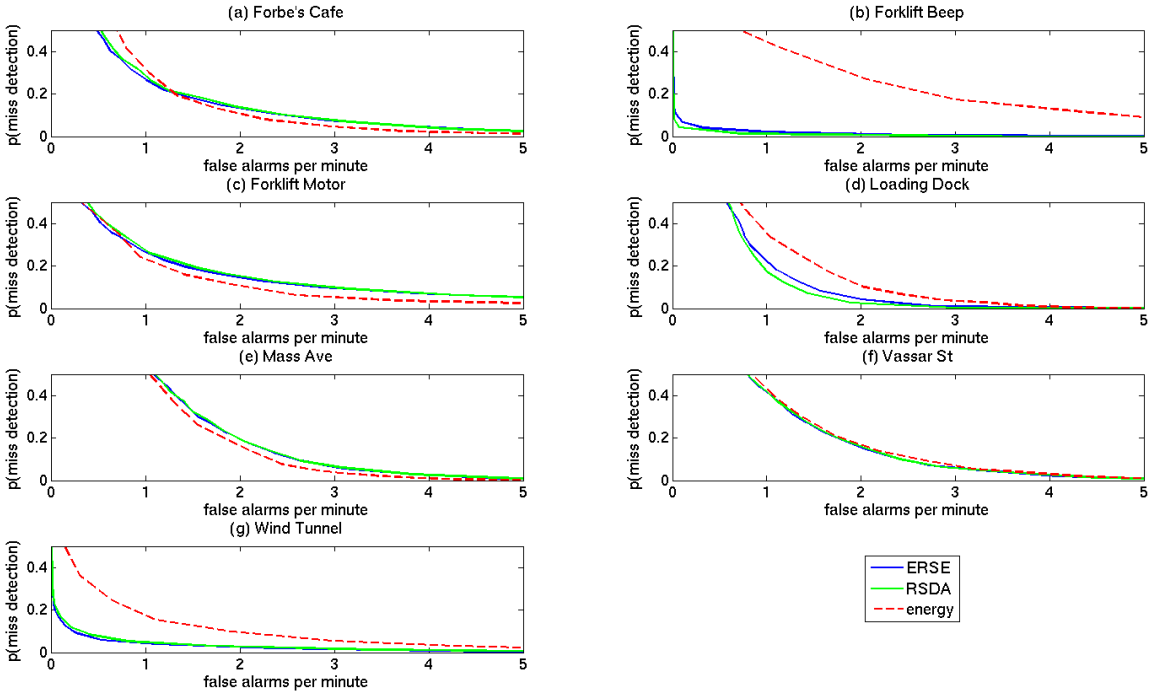

Figure 4-12: DET curves of the Robust Shout Detection Algorithm (RSDA) for each noise condition, along with the baseline system (frame-based energy) and most optimal feature without speech enhancement (ERSE). Across all noise conditions, the RSDA outperforms all other configurations of the system, suggesting that it is the most optimal algorithm for our application.

noise, achieving near-perfect speech detection. Many of the improvements in the Forklift Beep noise conditions (and other stationary noise conditions) are accredited to spectral subtraction, which is able to completely eliminate the background noise, resulting in only the speech components in the input waveform.

To provide a quantitative analysis on the improvements of the system's performance, we compute the Equal Error Rate (EER) for the baseline system and RSDA for each of the seven noise configurations. The EER is defined as the point on the conventional DET curve (e.g. Figure 3-4) where *P(miss detection) = P(false alarm)*. The EERs for the baseline system and RSDA are presented in Table 4.1. In comparing the EER of the RSDA to the baseline, we see a 38% reduction in the average EER, suggesting the overall improvements in the system's performance through RSDA.

Table 4.1: Equal Error Rate (EER) for baseline system and Robust Shout Detection Algorithm (RSDA) approach in Simulated Shout Data Set. The EER under each noise condition is an average across three SNRs: 0 dB, 5 dB, 10 dB. In comparing the EER between the baseline feature and the RSDA, we see a 38% overall EER reduction in the system's performance.

		EER (%)	
		baseline	**RSDA**
Noise Condition	**Forbe's Cafe**	10.9	10.8
	Forklift Beep	14.7	0.6
	Forklift Motor	8.8	7.5
	Loading Dock	8.8	8.2
	Mass Ave	10.8	8.2
	Vassar St	11.8	7.6
	Wind Tunnel	7.9	3.1
	AVERAGE	10.5	6.6

4.6.3 Summary

With the knowledge that we gain from experimenting with different features and speech enhancement techniques, we develop a robust shout detection algorithm (RSDA) that optimizes the performance of the shout detection system. The RSDA is tested in the seven noise conditions from the Simulated Shout Data Set, which comprises of the possible noise environments in which the forklift operates. In comparing the DET curves of the RSDA to the baseline system and the best feature thus far (see Figure 4-12), we can see that the RSDA outperforms the other features across all noise conditions, suggesting that it is the optimal algorithm among the set of features that are tested.

Chapter 5

Conclusion

5.1 Summary

In this thesis, we explored the effectiveness and application of various noise-robust techniques towards real-time speech detection in real environments. As part of a collaborative effort in the development of an autonomous forklift for the Agile Robotics Lab at MIT's CSAIL, our goals focused on investigating various approaches to speech detections by analyzing different features and speech enhancement techniques. Ultimately, our objective is to develop an optimal algorithm to detect shouted speech in the noise environments in which the autonomous forklift will be operating.

Through the results of our experiments, we developed the Robust Shout Detection Algorithm (RSDA), which utilized the speech enhancement and feature extraction techniques that we have analyzed. The algorithm is designed to perform well in the noise conditions that are similar to the environment of the operating forklift. The RSDA utilized the frame-based energy and relative spectral entropy (RSE) features, along with spectral subtraction in enhancing the speech signals, to accurately identify speech components in a given waveform. At the same time, the RSDA aimed for a low count of false alarms by disregarding stationary and non-stationary noise as background noise. We measured the performance of the algorithm by comparing its Design Error Tradeoff (DET) curves with that of the baseline system (frame-based energy), and other features. From the DET curves, we saw that the RSDA

outperformed the other features across all noise conditions in the Simulated Shout Data Set. From the DET curves, we computed the Equal Error Rate (EER) for the baseline feature and the RSDA (see Table 4.1) and concluded that the new shout detection approach provided 38% reduction in the EER over the baseline task.

5.2 Future Directions

We would like to expand this work in a number of areas in the future. One particular area is the SNR level in which we choose to operate. Currently, many of the results from the experiments are derived from data collected in various SNR levels, from 0 dB to 10 dB. However, in the application of the autonomous forklift, we can expect subjects to be shouting over a certain volume. Thus, we can expect that the SNR level should not drop to 0 dB. Future work should involve refining the range of SNR levels to provide a better model to the forklift application.

To further our interest in the SNR levels, another area to investigate is how the features perform in different SNRs. This thesis only analyzed the performance of the features in the set of noise conditions, keeping the SNR constant. While it is expected that the performance of the features should decrease with the decrease in the SNR levels, it would be interesting to see how they change. If a particular feature is not sensitive to a change in the SNR level, it means that the feature is robust to changes in the SNR level.

Furthermore, the experiments from this thesis assume that the SNR is fixed, even though the SNR in the real environment is constantly changing. Any introduction of new artifacts or external noises will influence the recording environment and change the SNR. Because of the application of the forklift, having an adaptive algorithm is very important. Therefore, it would be interesting to investigate means of adaptively adjusting the RSDA to take into account any changes in the noise conditions.

Finally, we would like to expand our research into other types of features that would be useful for our application. Aside from conducting more research on the effect of SNR levels, another area to expand upon is investigating on other types of

features that would be robust enough to perform well in the environment that the autonomous forklift will be operating in. Essentially, we aim to exploit more speech properties in an effort to develop an algorithm that is optimal in detecting shouted speech in various noise environments.

Appendix A

Reference Transcription

Table A.1: Example of part of a reference transcription from the Simulated Shout Data Set. These transcriptions are created by manually listening to the waveform, and recording the starting and ending time stamps of each utterances.

Start Time (s)	**End Time (s)**	**Utterance**
0.0000	5.0225	NONSPEECH
5.0225	5.9225	Back up
5.9225	9.5175	NONSPEECH
9.5175	10.2475	Brake
10.2475	13.2700	NONSPEECH
13.2700	14.2575	Cancel
14.2575	17.1100	NONSPEECH
17.1100	18.0425	Don't Move
18.0425	21.1025	NONSPEECH
21.1025	22.0250	Forklift
22.0250	24.8725	NONSPEECH
24.8725	25.5750	Freeze
25.5750	28.9075	NONSPEECH
28.9075	30.2600	Get Out Of The Way
30.2600	33.3750	NONSPEECH
33.3750	34.3225	Go back
34.3225	37.4450	NONSPEECH
37.4450	38.2625	Halt
38.2625	40.8050	NONSPEECH
40.8050	41.5975	Hey
41.5975	44.1775	NONSPEECH
44.1775	45.0675	Ho
45.0675	48.2125	NONSPEECH
48.2125	49.0500	Hold It
49.0500	52.3700	NONSPEECH
52.3700	53.3450	Hold On
53.3450	56.6000	NONSPEECH
56.6000	57.5375	Hold Up
57.5375	60.7700	NONSPEECH
60.7700	61.6800	No

Bibliography

[1] A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Pryzbocki. The DET Curve in Assessment of Detection Task Performance. In *Proc. Eurospeech*, 1997.

[2] A. Ouzounov. Robust Features for Speech Detection - Comparative Study. In *Proc. ICASSP*, 2005.

[3] Acoustic Magic, Inc. Voice Tracker Array Microphone, 2007.

[4] Agile Robotics, Autonomous Forklift. Massachusetts Institute of Technology, 2009.

[5] A.P. Varga, H.J.M. Steeneken, M. Tomlinson, D. Jones. The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition. Technical report, DRA Speech Research Unit, 1992.

[6] B. H. Juang. Speech Recognition in Adverse Environments. *Computer Speech and Language*, 5:275–294, 1991.

[7] B. Mak, J. Junqua, B. Reaves. A Robust Speech/Non-speech Detection Algorithm Using Time and Frequency-Based Features. In *Proc. ICASSP*, 1992.

[8] C. Park, N. Kim, J. Cho. Voice Activity Detection Using Partially Observable Markov Decision Process. In *Proc. Interspeech*, 2009.

[9] C. Zhang, J.H. Hansen. Analysis and Classification of Speech Mode: Whispered through Shouted. In *Proc. Interspeech*, 2007.

[10] D. Singh and F. Boland. Voice activity detection. *Crossroads: The ACM Student Magazine*, Xrds:13–14, 2009.

[11] F. Liu, R.M. Stern, X. Huang, A. Acero. Efficient Ceptral Normalization for Robust Speech Recognition. In *Proc. Human Language Technology Workshop*, 1993.

[12] G.S. Ying, C.D. Mitchell, L.H. Jamieson. Endpoint Detection of Isolated Utterances Based On A Modified Teager Energy Measurement. In *Proc. ICASSP*, 1992.

[13] H. Blatchford, P. Foulkes. Indentification of Voices in Shouting. *The International Journal of Speech, Language, and the Law*, 13.2:241–254, 2006.

[14] H. Hermansky and N. Morgan. RASTA Processing of Speech. *IEEE Transactions on Speech and Audio Processing*, 2:578–589, 1994.

[15] I. Hetherington. PocketSUMMIT: Small-Footprint Continuous Speech Recognition. In *Proc. Interspeech*, 2007.

[16] J. Glass. A Probabilistic Framework for Segment-Based Speech Recognition. *Computer Speech and Language*, 17:137–152, 2003.

[17] J. Odell, D. Kershaw, D. Ollason, V. Valtchev, D. Whitehouse. *The HAPI Book: A Description of the HTK Application Programming Interface*. Entropic Ltd, 1999.

[18] J. Ortega-Garcia and J. Gonzalez-Rodriguez. Overview of Speech Enhancement Techniques for Automatic Speech Recognition. In *Proc. ICSLP*, 1996.

[19] J. Shen, J. Hung, L. Lee. Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments. In *Proc. ICSLP*, 1998.

[20] J.H. Hansen, S.E. Bou-Ghazale. Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database. In *Proc. Eurospeech*, 1997.

[21] K. Yamamoto, F. Jabloun, K. Reinhard, A. Kawamura. Robust Endpoint Detection For Speech Recognition Based On Discriminative Feature Extraction. In *Proc. ICASSP*, 2006.

[22] L. Gu, S. Zahorian. A New Robust Algorithm for Isolated Word Endpoint Detection. In *Proc. ICASSP*, 2002.

[23] L.P. Kaelbling, M.L. Littman, A.R. Cassandra. *Planning and Acting in Partially Observable Stochastic Domains*. Number 101. Artificial Intelligence, 1998.

[24] L.S. Huang, C.H. Yang. A Novel Approach to Robust Speech Endpoint Detection in Car Environments. In *Proc. ICASSP*, 2000.

[25] M. Brookes. VOICEBOX: Speech Processing Toolbox for MATLAB, 2009.

[26] M. Mason. Speech Activity Detection for gStreamer (gstSAD), 2009. Vers. 0.2.8.2.

[27] M.L. Seltzer, J. Droppo, A. Acero. A Harmonic-Model-Based Front End for Robust Speech Recognition. In *Proc. Eurospeech*, 2003.

[28] M.Y. Appiah, M. Sasikath, R. Makrickaite, M. Gusaite. *Robust Voice Activity Detection and Noise Reduction Mechanism Using Higher-Order Statistics*. Institute of Electronics Systems, Aalborg University, 2005.

[29] P.J. Moreno and R.M. Stern. Sources of Degradation of Speech Recognition in the Telephone Network. In *Proc. ICASSP*, 1994.

[30] S. Basu. A Linked-HMM Model for Robust Voicing and Speech Detection. In *Proc. ICASSP*, 2003.

[31] S. Das, R. Bakis, A. Nadas, D. Nahamoo, M. Picheny. Influence of Background Noise and Microphone on the Performance of the IBM Tangora Speech Recognition System. In *Proc. ICASSP*, 1993.

[32] T. Fawcett. An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

[33] T. Kristjansson, S. Deligne, P. Olsen. Voicing Features for Robust Speech Detection. In *Proc. Eurospeech*, 2005.

[34] T. Sainath. *Applications of Broad Class Knowledge for Noise Robust Speech Recognition*. PhD thesis, MIT, 2009.

[35] W.S. Ching, P.S. Toh. Enhancement of Speech Signal Corrupted by High Acoustic Noise. In *Proc. ICASSP*, 1979.

[36] Y. Gao and J-P. Haton. Noise Reduction and Speech Recognition in Noise Condition Tested on LPNN-based Continuous Speech Recognition System. In *Proc. Eurospeech*, 1993.

[37] Y. Gong. Speech Recognition in Noise Environments: A Survey. *Speech Communication*, 16.3:261–291, 1995.

[38] Y.D. Cho, K. Al-Naimi, A. Kondoz. Mixed Decision-Based Noise Adaptation For Speech Enhancement. *Electronics Letters*, 37:540–542, 2001.