# Characterization of Emotional Speech in Human-Computer Dialogues

by

Alicia D. Boozer

B.S., Spelman College, Atlanta, Georgia (2000)
B.S., North Carolina A&T State University, Greensboro, NC (2000)

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

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2003



Author ............ ..................
Department of Electrical Engineering and Computer Science
September 5, 2003

Certified by ...............
Stephanie Seneff
Principal Research Associate
Thesis Supervisor

Accepted by .... ......
Arthur Smith
Chairman, Departmental Committee on Graduate Students

# Characterization of Emotional Speech in Human-Computer Dialogues

by

Alicia D. Boozer

Submitted to the Department of Electrical Engineering
and Computer Science
on September 5, 2003, in partial fulfillment of the
requirements for the degree of
Master of Science

## Abstract

This thesis seeks to improve our understanding of the effects of prosodic and temporal variations on a speech recognition system by acquiring a tagged corpus of natural emotional speech elicited in human-computer interaction and developing a robust set of features to quantify acoustic correlates of emotion. A focus will be given to frustrated emotional speech since the accuracy using just this emotional state has been found to do a better job at discriminating between emotional and non-emotional speech [1], and frustration also leads to an unpleasant experience for the user.

Three corpora are used to investigate properties of emotional speech and identify features to be used during classification and recognition tasks. Two of the corpora, MERCURY and Lockheed, consist of utterances that occur naturally during a routine task. The Linguistic Data Consortium corpora of Prosody and Emotional Speech, representing the third corpus, is comprised of utterances spoken by actors. Both elicited and non-elicited speech are used to provide a comparative study of emotion.

Using parallel utterances as the basis of our investigation, an analysis of speech variations caused by emotion was performed using spectrograms and prosody-based tools. Six prosodic features were proposed to be good discriminators of emotional speech. Using data collected from a real-life scenario, the classification experiment determined that average pitch was the best at discriminating between emotional states, while maximum pitch was best for discrimination success using acted data.

Human-listener studies are presented to assess the labeling method used for tagging the emotional utterances, and to make comparisons between human and computer emotion recognition. Finally, a comparison of results from both the real and acted data supports our hypothesis that results reported on acted data may be misleading since non-elicited emotional speech is more subtle and harder to detect.

Thesis Supervisor: Stephanie Seneff
Title: Principal Research Associate

# Acknowledgments

First, I must thank my Heavenly Father for bringing me to this point in my life. It is because of you that I am and that I have been blessed to grow and mature during my experience at MIT. You deserve all praise and honor. Thank you for using the following people as instruments in helping me navigate this process:

My thesis advisor, **Stephanie Seneff**—I would like to express my heartfelt thanks to you for being such a wonderful advisor and mentor. Throughout this research, and a number of obstacles that surfaced, you still exuded patience and compassion.

My thesis advisor, **Michelle Spina**—I had the fortunate opportunity to have two advisors to give me guidance and suggestions on how to approach this research problem. Michelle, I would like to thank you for your mentorship, stressing the importance of documentation, and helping me hunt down bugs.

My office mates, Ernie, Han, and Laura—I am most appreciative of all of the laughs and office outings. Ernie, thank you for your discussions on diet and nutrition. Thank you Han for answering my incessant questions and being a great resource for getting things to work. Thank goodness for Laura who always came through with an appropriate social outing after too many days of just work.

The SLS group—Everyone in the SLS group deserves thanks. Specifically, I would like to thank Chao Wang for her invaluable assistance on organizing my thesis and performing experiments, and her expertise in the area of prosody. Also, to TJ Hazen for his help with the emotion classification and recognition experiments and for ideas on reporting the results. To Joe Polifroni, for his help with data collection and transcription. Jon Yi, Karen, and Min for being repositories of information on everything from understanding the GALAXY system to finding the best restaurants. I would also like to thank all of the present and several past members of SLS for making this group a fun, learning environment and for your encouragement in finishing this thesis.

To the Graduate Education for Minorities (GEM) Consortium—Words cannot express how truly grateful I am to have been financed by you during my first two years at MIT. Your financial and emotional support allowed me the opportunity to develop the skills and confidence needed to get through my program. Special thanks to Forestine Blake, who believed in my goals and magically found money when there was none.

To University friends—I will always have a special place in my heart for the staff in the Graduate Students Office. Thank you Dean Colbert, Dean Staton, Dean Charles, Brima, Ed, George and Heather for the words of encouragement, advice, and occasional chastising so that I could finish up this degree. I would also like to thank the ACME group for your help with troubleshooting problems and for providing a structure for completing goals. In addition, I would like to thank Eric B. for his love, support, and for sharing his testimony so that I could believe.

To my church family—Thank you to all of the saints, especially at Blue Hill, who prayed for me and encouraged me during this time. After going through many storms, I now recognize that my battles are spiritual ones. I will forever be indebted to everyone who prayed for me when I wasn't in a position to pray for myself.

To my biological family—I would like to thank my mom for sacrificing so much so that I could have more. Even with your quiet disposition, I see so much of you in me. I am most grateful for the qualities of diligence and resilience that you instilled in me through your example. I am also grateful to my aunts, granny, dad, and sisters who constantly encouraged me during this research.

To friends—Lisha, Jami, Shanda, Fred, and the members of BGSA, I thank God for His timing in allowing our paths to cross. Each of you has definitely added to my life.

I conclude by giving thanks to my Lord and Savior, Jesus Christ. He is truly the author and finisher of my faith. I thank you for loving me so much (despite how many times I fall short) that you gave your life for me. Throughout this experience, I could feel your presence. I pray that my experience will be a testament to others of your awesome power.

This research was supported by DARPA under contract number NBCH1020002, monitored through the Department of the Interior, National Business Center, Acquisition Services Division, Fort Huachuca, AZ.

I took the one less traveled by

and that has made all the difference.

*- Robert Frost*

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

The success of a speech recognizer depends mainly on two factors—accuracy and robustness. That is, in order to perform an intended task, speech recognition systems must be able to achieve low error rates and then maintain them under varying conditions. However, variations among speakers often pose problems for speech recognizers and make it difficult to maintain low error rates.

Variations in an individual's speaking style and emotional state may also degrade recognition performance. Exaggerated pronunciation of specific phones or syllables, also known as hyperarticulation, has been a topic of study as it relates to emotion [1]. Often users who are not understood by a dialogue system tend to hyperarticulate misrecognized words or phrases, usually leading to further misrecognition and frustration.

In order to build dialogue systems that are robust to speaking style variations, research on speech and emotion becomes an important topic of study. With the information gleaned from research on emotion, dialogue systems could assist frustrated users by offering a response that guides the user to speak neutrally so that the desired task can be achieved, suggesting an alternative mode for entering data, or transferring the user to a human operator. The latter would be extremely advantageous within call centers where a system's response could mean the difference between maintaining and losing a customer.

This thesis will identify characteristics of emotional speech and propose a set of features for maximum discriminability, with the goal of incorporating these features into a dialogue system. This objective will be pursued by visually inspecting emotional speech using spectrograms and prosody-based tools, computing measurements for observed physical features, and finally running classification experiments utilizing these features.

Emotion can be communicated through verbal and non-verbal cues. Using verbal cues, certain word choices such as "bored" or "frustrating" may be used to indicate a speaker's emotional state. Non-verbal communication may include facial expressions, body gestures, and changes in prosody within an utterance. For example, acoustic features such as a sigh or an increase in the total energy of an utterance may also suggest that a speaker is bored or frustrated, without the explicit semantics being used.

There are several challenges involved in developing systems to recognize emotion in speech. One challenge that often arises is finding an accurate description and annotation for the expressed emotion. For example, one individual may describe a perceived emotional state as frustration while another may choose to describe the same emotion as irritation. While both labels may indicate displeasure of the user, these differences make quantifying emotion very difficult. Across interdisciplinary domains, how emotion is characterized varies greatly. Another challenge is the collection of emotional speech. Real-life scenarios, which are ideal for collecting emotional data, are often difficult to monitor, and one is faced with the task of screening large quantities of collected data for those few utterances that are emotionally charged. Finally, after an expressed emotion is decoded within an utterance, the dialogue system must be able to integrate information about the user's emotional state and adjust accordingly.

In the remainder of this chapter, an overview of emotion research and data collection techniques will be presented. Section 1.1 covers prosody and gives some statistics on the recognition performance of emotional speech in the MERCURY corpus. Section 1.2 presents previous work that has been done on emotion and describes techniques that are employed to collect emotional data. Section 1.3 details the thesis objective followed by an outline of the chapters of this thesis in Section 1.4.

## 1.1 Background

Prosody is generally thought of as the organizational structure of an utterance. In human-human communication, from a listener's point of view, prosody is associated with the perception of a speaker's intention based on acoustic aspects of speech, i.e., duration, pauses, fundamental frequency ($F_0$), and amplitude. The speaking style and emotional state of an individual can also affect the prosody of an utterance. In the context of automatic speech recognition, the emotional state of a user significantly affects the prosody which in turn also affects the recognition performance. In fact, for a frustrated emotional state, the effect is often performance degradation.

This research is confined to investigating the information within the speech signal that corresponds to emotion. A study of the acoustic correlates and changes in the prosodic parameters is presented as a means of characterizing emotional speech. It is well known that an individual exhibiting frustration alters his or her speaking style in a way that can have adverse effects on speech recognition, subsequently leading to further frustration [29, 30, 47]. A vicious cycle ensues, and neither the user nor the system can carry out the desired task.

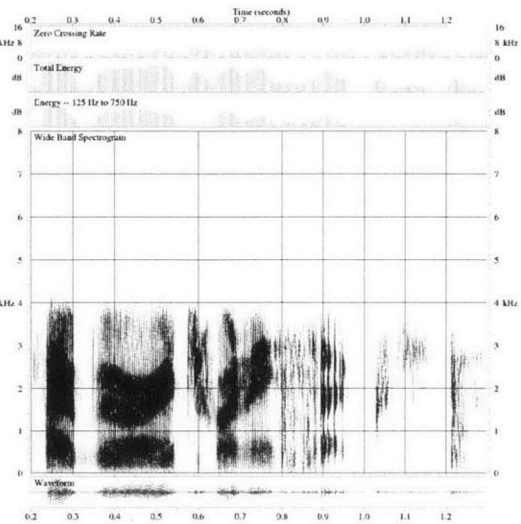
For example, preliminary experiments on MERCURY suggest that the recognition performance degrades from 15% to almost 33% for neutral and frustrated emotional states respectively. That is almost a 120% increase in word error rate when the emotional state of the user changes from neutral to frustrated. Similarly, there is a 67% increase when comparing the emotional state of neutral and happy speech. The degradation of the recognition performance for emotional speech, whether as a result of a pleasant or unpleasant experience by the user, causes serious problems for the speech recognizer. This thesis, then, seeks to do the following: (1) explore those properties, primarily prosodic, that may be useful in detecting emotion, and (2) investigate acoustic correlates of emotion that influence speech degradation. Ideally, developing a system that can detect the emotional state of a user and adjust its behavior accordingly will indirectly lead to better recognition by changing the user's behavior and reducing frustration.

Figure 1-1 displays the spectrograms of three utterances with the same linguistic content but different emotional states spoken consecutively by a female user in the MERCURY domain. The utterance spoken is "february twenty sixth". The spectrograms show an increase in frustration, with (a) being the non-emotional utterance and (c) representing the most frustrated utterance. The spectrogram for the non-emotional utterance shows that the overall energy is much less than that of the highly frustrated utterance. There also appears to be a lengthening of the vowels as the emotional state changes. As shown in the last spectrogram, there is also a noticeable increase in the pause duration for the emotional utterance. It is hoped that these visually observable changes within repetitive utterances with different emotional states will aid in characterizing and eventually detecting emotion within speech signals.

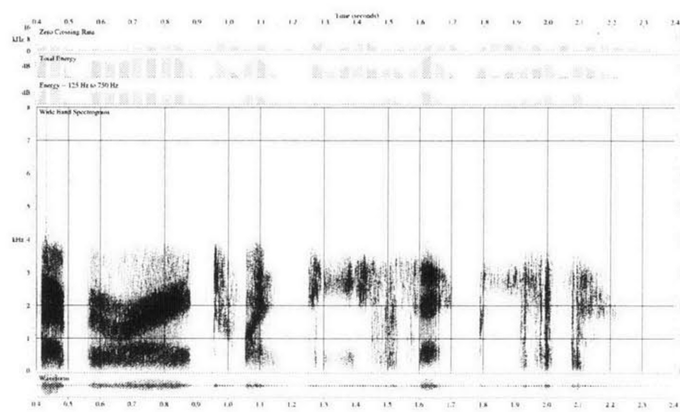
In other literature, the recognition accuracy of emotional data is noted to be about 70% on average [33, 34] while an accuracy of about 60% is achieved for automatic classification based on prosodic models [40]. While much of the literature accepts the recognition accuracy to be between 60-70% for emotional speech, there are no standard corpora, making it difficult to compare results across experiments.

## 1.2 Previous Work

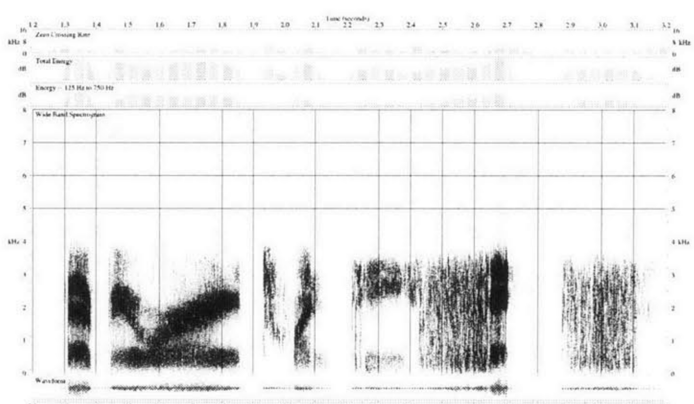
Much of the research that has been done on emotion recognition has been in the fields of psychology and linguistics. Research on the integration of prosody and emotion into more sophisticated human-computer systems has only occurred within the past few years. While much of the current research has acknowledged the importance that prosodics play in detecting emotion, there does not appear to be any general consensus on which features may be good predictors of emotional state, and whether other features not typically associated with prosody may also play a role.




(a)



(b)



(c)

Figure 1-1: Spectrograms of three utterances with the same linguistic content but different emotional states for the phrase "february twenty sixth" spoken by a female speaker. The differences in the neutral utterance (a) and the highly frustrated utterance (c) supply a great deal of information. The frustrated utterance illustrates an increase in energy and the duration of vowels and internal pauses.

### 1.2.1 Research Approaches

Previous research into emotional speech focuses on either recognizing emotion or expressing emotion and has been undertaken with differing motives. Some researchers take the social science approach whereby basic human needs are examined and then systems are developed to help meet those needs [2, 9]. Others investigate emotion to educate users of a system by providing feedback or mirroring their present emotion and providing devices to help users classify surrounding emotions[5, 21].

Researchers at the MIT Media Lab have explored several systems that recognize and respond to emotion and help the user increase self-awareness of his emotional state [35, 37]. The HAL 9000 computer was created as an affective computer and has abilities that influence an individual's emotions. Other systems that were developed were designed to help children and are used for educational and entertainment purposes.

Holzapfel and his colleagues are doing work that focuses on the cognitive and physical aspects of recognizing emotion [18]. To investigate the physical aspects of emotion, the observed emotion is obtained by recording speech, facial expressions, and through sensors in contact with a person. The Bayes Point machine [41] is another example of evaluating a user's emotion using sensors. This research is based on the hypothesis that users interact via a mouse to apply excess pressure when difficult events are encountered. Eight pressure sensors are mounted on a computer mouse and measurements are taken from one of two distinct regions, with or without a usability bug. This research is encouraging and emphasizes that human emotions impact even simple common tasks.

Different research approaches for recognizing emotion are also accompanied by different classification models. A classification model that uses specific features from the speech signal to recognize the user's emotion is needed if any meaningful characterization is to be obtained. For this project, only two target emotional states, frustrated and non-emotional (neutral), were used for the purpose of emotional characterization. These two states were chosen since they are of greatest relevance to our intended application.

### 1.2.2 Data Collection Techniques

Recognition rates are heavily dependent on the techniques used in data collection. Previous work has been done primarily using elicited emotions since data collection of real-life scenarios is often very difficult to monitor and acquire. As a result, more creative techniques for simulating and collecting emotional data have been employed.

#### Experienced Actors

The most common technique for collecting data for emotional studies is using a controlled group of experienced actors to simulate a specific emotional state. The same sentence and test conditions are generally used to minimize non-emotional variations. Using experienced actors often produces the greatest amount of emotional data since the only limitation involves the actor's ability to mimic a particular emotional state. This technique, although ideal for areas of research in speech synthesis, is not optimal for speech recognition, where testing conditions usually include natural emotions that are exhibited differently by different speakers and that are likely to be much more subtle.

#### Wizard-of-Oz

The Wizard-of-Oz (WOZ) scenario appears to be a good compromise between the availability of data and a real-life setting. In this data collection technique, users are presumably naive; that is, they are unaware that they are not communicating with a real computer system. Typically, WOZ experiments frequently reject or misrecognize user utterances in order to evoke a particular emotional state. Consequently, WOZ experiments can be a valuable tool for evaluating peculiarities and designs in human-computer dialogue systems [12].

### Real-life Settings

The goal of the preceding techniques is ultimately to model *normal* speech of a human being in a real-life setting. As stated earlier, there are difficulties in acquiring large amounts of emotional data in real-life scenarios. By switching to different applications, the emotional state as well as the targeted users may change. For example, monitoring a user's emotional state during the middle of rush hour would conceivably be much different than monitoring a subject's emotional state during a birthday party.

## 1.2.3 Discussion

Although researchers have begun to study the effects that emotion has on recognition performance, much of the published results have been obtained using elicited speech. While the results give some idea of the work that still needs to be done in this field, it is also misleading since non-elicited emotional speech is more subtle and has its own challenges. For example, an actor portraying a frustrated emotional state will usually exaggerate the fundamental frequency and energy of his speech so that the emotion present is obvious, but in a real-life scenario a user may become so frustrated that he lowers his fundamental frequency as if he is bored or even laughs. Now the emotional state that is present becomes harder to detect because the expressed emotion is similar to the emotional states of boredom or elation.

Techniques also vary making it difficult to compare classification and recognition results. As yet, there are no standard corpora used for emotion research. Researchers have used corpora that vary from movie segments to non-professional actors, such as students, to perform classification and recognition experiments. Because corpora use different vocabularies and systems use different language models, results are not as easily comparable.

Also, much more attention has been given to developing human-computer interfaces that express emotion as opposed to developing systems that detect emotion. Research on synthesis of emotional speech is another topic of interest for similar reasons to that of emotion recognition. The integration of prosodic information on the front-end and back-end of the spoken language system will hopefully lead to conversational systems that are more natural.

These differences in data collection, techniques, corpora and recognition systems have made understanding and comparing emotion classification and recognition results difficult. For these reasons, there is still a great deal to be accomplished within emotion and prosody research.

## 1.3 Objective

The objective of this thesis is to acquire a tagged corpus of natural emotional speech elicited in human-computer interaction, develop a robust set of features to quantify acoustic correlates of emotion, mainly frustration, and to improve our understanding of the effects of prosodic and temporal variations on the speech recognition system. To accomplish this goal, the thesis presents a comparative study of statistical results and human listener studies.

The fundamental frequency has been recognized as an important feature in disambiguating between emotional states. Pitch detection is essential for the completion of the thesis goals. There are several challenges associated with extracting pitch information using pitch detection algorithms. To address these challenges, this thesis uses a continuous pitch detection algorithm that was developed within the Spoken Language Systems group [49]. The main contributions of this thesis are a direct result of using the pitch detection algorithm to create vowel-level and utterance-level features for investigation. However, since an in-depth discussion on pitch detection is another research topic, this thesis will only present relevant details of the algorithm that are necessary for understanding the methodology used for this project.

## 1.4 Outline

The remainder of the thesis is organized into five chapters and an appendix. Chapter 2 covers the experimental methodology. The methodology is designed to minimize any variations within the speech signal caused by differences in speakers, gender, and language pronunciation not associated directly with the expressed emotion. The tasks to be performed are utterance-level emotion classification and recognition. The corpora, pitch tracking tools, and recognition system are also presented in this chapter.

Chapter 3 covers the preliminary data analysis. We identified a significant number of situations where users repeated an utterance multiple times and sequentially because of persistent recognition error. We decided to exploit these repeated utterances to serve as the basis for understanding acoustic correlates of emotion because they form an ideal condition of parallel utterances in terms of speaker and linguistic content, differing only in the presumed emotional state provoked by repeated recognition failure. An overview of the acoustic features used during the experiments will also be presented. Emotional characteristics of the speech signal are examined using spectrograms, graphical tools, and ROC curves. Training and classification techniques are also presented. These experiments are analyzed to better understand which parameters are most correlated with emotion.

Chapter 4 presents the human-listener studies. The design and interface of the human listener study will be presented and analyzed. Results are presented for comparison with the statistical results for the acted and real data. These experiments are analyzed to better understand the challenges associated with detecting emotion in spoken language systems.

Chapter 5 summarizes the thesis and suggestions for future work on this project are discussed. The appendix includes sample output from the corpora.

# Chapter 2

# Experimental Methodology

Within human-human communications, recognizing emotions is a critical skill for reasoning through life-threatening situations, integrating into social environments, and successfully communicating a desired goal that should be achieved. Overall, humans are very good at detecting the presence of emotion, although classifying the specific emotion proves more difficult. While linguistic evidence may influence a person's decision for selecting a particular emotional state, humans are able to perceive elicited emotion fairly accurately based merely on the prosodic cues present in a speech signal [32]. Using this knowledge as a foundation for our research, this thesis identifies prosodic features and investigates how variations in these parameters affect the performance of a speech recognition system.

The goal of this thesis can only be achieved if there are some identifiable and reliable acoustic correlates of emotion embedded within the signal characteristics. This thesis designs an experimental method that uses parallel utterances to analyze and identify acoustic and prosodic cues of emotion. These parallel utterances consist of $n$-tuple repetitions that are structured essentially the same linguistically but differ in their emotional state. This way, an optimal comparison can be made between the parallel utterances by isolating the expressed emotional state.

Parallel utterances within a particular $n$-tuple group are spoken by the same speaker. Speaker dependency eliminates any variability in prosodic properties due to speaker variations. Using the results found during the data analysis, emotion classification and recognition tasks can be performed. The emotion classification and recognition tasks use the SUMMIT ("Speech Understanding by Machine at MIT") [51] recognition system developed within the Spoken Language Systems (SLS) group.

The remainder of this chapter will cover the corpora and general methods used in this research. A brief discussion of the pitch detection algorithm will also be presented. Section 2.1 presents the corpora and data collection techniques. The emotional corpora are primarily configured from the Lockheed *Sea Shadow*[1] data although the MERCURY[2] data was used to verify statistical findings and investigate linguistic features of emotion, and the Linguistic Data Consortium (LDC) corpora of Prosody and Emotional Speech were used as a comparison between unelicited and elicited speech.

Because our research uses repeated utterances to analyze emotional speech, an understanding of the effects of prosodic variations on speech will be useful in determining reliable features. Section 2.2 presents a general overview of prosody and a discussion of the most recognized prosodic features is given. The pitch detection algorithm is also presented.

Section 2.3 covers the methods used to effectively analyze the emotional characteristics. A discussion of the analysis, classification experiments, and human listener studies will be given. Section 2.4 presents the system used during the classification and recognition tasks, and section 2.5 covers the tasks to be performed during classification. Finally, section 2.6 gives a summary of the chapter.

## 2.1 Corpora

The collection of emotional speech for building the corpora was a substantial part of this thesis. Our study relied on two corpora, chosen primarily on the basis that the emotion under investigation was not intentionally elicited. The corpora were collected

[1] A corpus provided by Lockheed Martin. Details about the ship and the program surrounding its creation can be found at several websites, including http://www.globalsecurity.org/military/systems/ship/sea_shadow.htm

[2] A corpus of human-computer dialogues in a flight reservation domain developed in the SLS group.

from real-life scenarios that naturally elicited emotion from the user. Consequently, the corpora are invaluable in emotion research. During the latter part of this thesis, the LDC corpora of Prosody and Emotional Speech were used to compare results between unelicited and acted speech. The amount of data and emotional utterances used during this research is summarized in Table 2.1.

Data collection, in general, is no easy task. In the area of emotion research, data collection is even more laborious because this topic of study is relatively new, often leaving the data collection the sole responsibility of the researcher, especially if a study of emotion produced in real-life situations is desired. The labeling of a sentence as emotional is a subjective process, and is not consistent among different listeners, as will become evident in the chapter on listener studies.

This thesis focuses primarily on differentiating between frustrated and neutral emotional states because understanding these two states would arguably lead to designing better spoken dialogue systems. Frustration is an emotional state that is also of interest to commercial companies because of the injurious threat that losing a customer poses when misrecognition occurs repeatedly. Focusing on characterizing the acoustic and prosodic properties of frustrated speech will then hopefully lead to insights about the difficulties and issues that should be addressed when dealing with emotional speech.

### 2.1.1 Mercury Data

The first corpus was constructed from MERCURY, an application based on the GALAXY architecture [45], developed within the SLS group, that allows users to arrange airline flights and price itineraries. In many cases, the users were actually trying to arrange a real flight itinerary; however, there were occasions where users were not making real travel plans. Consequently, the frequency of frustrated utterances in the data was lower than it would have been in an actual flight reservation scenario.

Data collection was done over the telephone and all data were sampled at 8 kHz. Approximately 22,000 utterances spanning a time-frame of about one year were listened to carefully for identification of one of three emotional states: neutral, frustrated, and happy. 1,674 utterances were labeled as frustrated, 38 as happy, and the remaining utterances labeled as neutral. 120 sessions were identified as having at least one utterance that was frustrated, 7 sessions as having at least one utterance that expresses satisfaction, and 3 sessions were identified as a mixed session, meaning there is at least one frustrated utterance and one happy utterance.

A number of interesting utterances that were identified in the MERCURY corpus provided valuable insight into events surrounding user frustration. For example, some users choose to express their emotion predominantly using linguistic cues, by using swear words, negative constructors, and termination commands, while others modify the prosodic features. We also observed that recognition error prompted repetition of information. In a few cases, the repeat utterance was spoken nearly the same; however, many users emphasized misrecognized words or altered the structure of the utterance. By making use of this information, we were able to qualitatively assess the effect that emotion has on recognition.

### 2.1.2 Lockheed Data

The second corpus was constructed from Lockheed data within a laboratory simulating being aboard the *Sea Shadow*. The *Sea Shadow* is a limited mobility platform used to research and test advanced technologies for surface ships in the areas of propulsion, automation, sea-keeping abilities, and reduced radar signature. The test craft was developed under a combined program by the Advanced Research Projects Agency (ARPA), the Navy, and Lockheed Martin Missiles and Space Company. The *Sea Shadow* program began in the 1980's. As part of its purpose to explore strategies for reduced manning automation, a task scenario was developed to monitor multiple events from within several domains and hopefully encourage multi-tasking. Originally, a corpus was constructed from actual speech aboard the *Sea Shadow* ship. However, because the speech was recorded on the ship's intercom microphones, many of the waveforms were noisy, therefore making it very difficult to obtain an accurate pitch extraction. As a result, a corpus was formed using the Lockheed lab data since the goal of this task was the same as the data collected aboard the *Sea Shadow* ship.

The Lockheed corpus was collected from 12 speakers, eleven male and one female, covering one American English dialect. Each speaker was informed to carry out a task scenario similar to what he or she may experience aboard the *Sea Shadow* ship. Sample tasks included reporting the delivery time and quantity of supplies, reporting the status of equipment and routine maintenance, and verifying user information. Each speaker spoke about 200 utterances for a total of almost 2,400 utterances. Of these 2,392 utterances, 268 were labeled as repeat utterances. Repeat utterances, spoken consecutively by an individual speaker, are labeled with an appropriate emotional state of either frustrated or neutral, and grouped by their linguistic content. Of the repeat utterances, 124 were labeled as frustrated.

An example scenario typically carried out on the CSSOC (pronounced /sɪs´ ɑk/) system is shown in Figure 2-1. The CSSOC recognizer was configured from the SUMMIT system.

### 2.1.3 LDC Prosody Corpus

The final corpus was compiled by the LDC and consisted of speech read by a group of actors in various emotional states. Unlike the other two corpora, this corpus was already tagged with an appropriate emotion and transcription, making the data collection relatively quick and painless.

The LDC corpus was collected from 7 speakers with 14 differing emotional states. Utterances were semantically neutral in that they consisted of only months and dates, as in `august thirteenth`. The emotional states of the LDC corpus included panic, anxiety, boredom, disgust, elated, and neutral speech. Since utterances for the Lockheed corpus were separated only into two emotional states, the emotional states for the LDC corpus were also separated into two distinct emotions by collapsing similar emotional states and eliminating others. For example, the emotions `disgust`, `hot anger`, `cold anger` and `dominant` were collapsed into the general emotional state `frustrated`. Neutral utterances were already tagged as such and were simply added to the corpus and labeled with an emotional state of `neutral`.

| | |
|---|---|
| CSSOC: | *This is c s s o c, now operational, over* |
| USER: | CSSOC THIS IS MAGTF FIVE |
| CSSOC: | *Magtf five this is cssoc over* |
| USER: | STAND BY TO RECEIVE RAPID REQUEST |
| CSSOC: | *Roger, ready to receive, awaiting precedence? over* |
| CSSOC: | *attention, alert about status update for b rats request, defer now?* |
| USER: | AFFIRMATIVE |
| CSSOC: | *Magtf five, back to your request, Awaiting precedence? over.* |
| CSSOC: | *Alert! Starboard engine water temperature is 201.* |
| USER: | SAY AGAIN |
| CSSOC: | *Alert! Starboard engine water temperature is 201.* |
| USER: | WHAT IS THE FIRE MAIN HEADER PRESSURE |
| USER: | WHAT IS THE FIRE MAIN HEADER PRESSURE |
| CSSOC: | *Fire main header pressure is 71.* |
| USER: | SEND MY ALERT |
| CSSOC: | *I don't understand your intention.* |
| CSSOC: | *Fire main header pressure is 71.* |
| USER: | BYE |
| CSSOC: | *Goodbye Joe.* |
| CSSOC: | *Magtf five, back to your request, Awaiting precedence? over* |

Figure 2-1: Example dialogue from the Lockheed corpus. This dialogue includes repeat utterances that were later used to investigate acoustic and prosodic cues in emotional speech.

### 2.1.4 Corpus Properties

Most of the work in this thesis was done on the Lockheed lab corpus. For the remainder of this thesis, reference to any data or results should be assumed to be for the Lockheed corpus unless explicitly noted otherwise.

Each utterance in the corpus was given 12 properties to aid in identification and analysis. The following properties were used: `tag`, `waveform_file`, `speaker`, `gender`, `call_id`, `length`, `original_orthography`, `orthography`, `artifact`, `repeat`, `emotional_state`, and `contains_oov`. The property descriptions along with an example of the corpus output for the listed properties can be found in the appendix.

| Corpus | Total | Emotional |
|---|---|---|
| MERCURY | 22,212 | 1,674 |
| Lockheed | 1,739 | 112 |
| Lockheed Lab | 2,392 | 124 |
| LDC | 2,378 | 1,419 |

Table 2.1: List of corpora used during this research. The MERCURY and Lockheed corpora were manually created and utterances were tagged by a single individual for the appropriate emotional state.

### 2.1.5 Transcriptions

Except for the LDC corpus, the transcriptions for all corpora were entered manually after listening to each utterance. All words that were observed in the recorded utterance were transcribed with the appropriate phonetic transcription to alleviate issues associated with out-of-vocabulary words. For instance, some of the utterances contained words such as "moron" and swear words. Utterances such as these were transcribed so that a phonetic and word alignment could be obtained and the utterance could be used in this research, since we needed to make use of all of the data that were available. Table 2.2 lists the acoustic-phonetic symbols used with their ARPA-bet symbols, and example reference expressions are provided as a reference guide for many of the symbols used in chapter 3.

## 2.2 Prosody

Prosody provides additional information about the meaning or encoded message of an utterance, whereas linguistic cues provide information about the literal content. In general, prosody refers to the complex organization of physical, phonetic effects that are used to express attitude and emphasis in human communication. Indeed, both types of information, linguistic and prosodic, contribute to the correct understanding and successful communication of a message. Prior to determining the differences between emotional and non-emotional speech, the prosodic features of the speech signal must be captured for analysis. This thesis focuses on investigating the prosodic

| Label | ARPAbet | Example | Label | ARPAbet | Example |
|---|---|---|---|---|---|
| **Vowels and diphthongs** | | | | | |
| i$^{y}$ (i$^{j}$) | iy | b*ee*t | u$^{w}$ | uw | b*oo*t |
| ɪ | ih | b*i*t | ü | ux | t*oo*t |
| e$^{y}$ (e$^{j}$) | ey | b*ai*t | ɝ | er | b*ir*d |
| ɛ | eh | b*e*t | ɑ$^{y}$ (ɑ$^{j}$) | ay | b*i*te |
| æ | ae | b*a*t | ɔ$^{y}$ (ɔ$^{j}$) | oy | b*oy* |
| ɑ | aa | b*o*b | ɑ$^{w}$ | aw | b*ou*t |
| ɔ | ao | b*ough*t | ə | ax | *a*bout |
| ʌ | ah | b*u*t | ɨ | ix | deb*i*t |
| o$^{w}$ | ow | b*oa*t | ɚ | axr | butt*er* |
| ʊ | uh | b*oo*k | | | |
| **Glides** | | | **Liquids** | | |
| w | w | *w*ay | r (ɹ) | r | *r*ay |
| y (j) | y | *y*acht | l | l | *l*ay |
| | | | l̩ | el | bott*le* |
| **Nasals** | | | | | |
| m | m | *mom* | m̩ | em | bott*om* |
| n | n | *noon* | n̩ | en | butt*on* |
| ŋ | ng | si*ng* | ŋ̩ | eng | Wash*ing*ton |
| ɾ̃ | nx | wi*nn*er | | | |
| **Fricatives** | | | | | |
| s | s | *s*ea | z | z | *z*one |
| š (ʃ) | sh | *sh*e | ž (ʒ) | zh | mea*s*ure |
| f | f | *f*in | v | v | *v*an |
| θ | th | *th*in | ð | dh | *th*en |
| h | hh | *h*ay | | | |
| **Stops** | | | | | |
| p | p | *p*ea | b | b | *b*ee |
| t | t | *t*ea | d | d | *d*ay |
| k | k | *k*ey | g | g | *g*ain |
| p$^{\urcorner}$ | pcl | po*p*s | b$^{\urcorner}$ | bcl | bo*b*s |
| t$^{\urcorner}$ | tcl | po*t*s | d$^{\urcorner}$ | gcl | po*d*s |
| k$^{\urcorner}$ | kcl | po*ck*s | g$^{\urcorner}$ | gcl | bo*g*s |
| ʔ | q | bu*tt*on | ɾ | dx | mu*dd*y |
| **Affricates** | | | | | |
| č (ʧ) | ch | *ch*oke | ǰ (ʤ) | jh | *j*oke |

Table 2.2: Phonetic labels for American English. The **Label** column lists the labels used in SLS, with the corresponding IPA label in parentheses when they differ. The **ARPAbet** column lists the equivalent labels in the ARPAbet ASCII phonetic alphabet.

properties of the speech signal in the hope that these properties will disambiguate between emotional and non-emotional speech.

According to many research studies, there are three physical, prosodic features that are most useful in detecting emotion—fundamental frequency, duration, and energy[10, 27, 50]. The perceptual form of these prosodic features are pitch, speaking rate, and loudness, respectively. This research uses and makes reference to the physical features instead of their perceptual counterparts because the physical features are measurable parameters.

Additionally, the fundamental frequency has been recognized as the most important prosodic feature in distinguishing between emotional states [43]. Consequently, detecting the fundamental frequency then becomes an essential step in the analysis of prosody and emotion in speech signals. One way to accomplish this is by using a pitch detection algorithm.

### 2.2.1 Pitch Detection

This thesis uses a continuous pitch detection algorithm (CPDA), which was designed specifically for prosodic modeling applications for telephone quality speech[49]. The pitch estimation method is based on the harmonic matching approach [16], which uses the relationship of harmonics being spaced a constant distance from the fundamental on a log frequency scale, and a dynamic programming search technique. The dynamic programming search finds a pitch value for every frame, including unvoiced regions. In our research, a segmentation of the waveform was obtained using a speech recognition system, and prosodic features from the vowel segments are analyzed.

Figure 2-2 shows the $F_0$ contour overlaid on a spectrogram for an utterance from the Lockheed corpus. As shown in the figure, the fundamental frequency is significantly weaker than the harmonics and may visually disappear so that an accurate pitch estimation depends on proper utilization of the higher harmonics[3]. To better assess the pitch tracker, its output is doubled so that it should coincide with the

[3] The weak fundamental is due to a high-pass filter in the telephone handset.

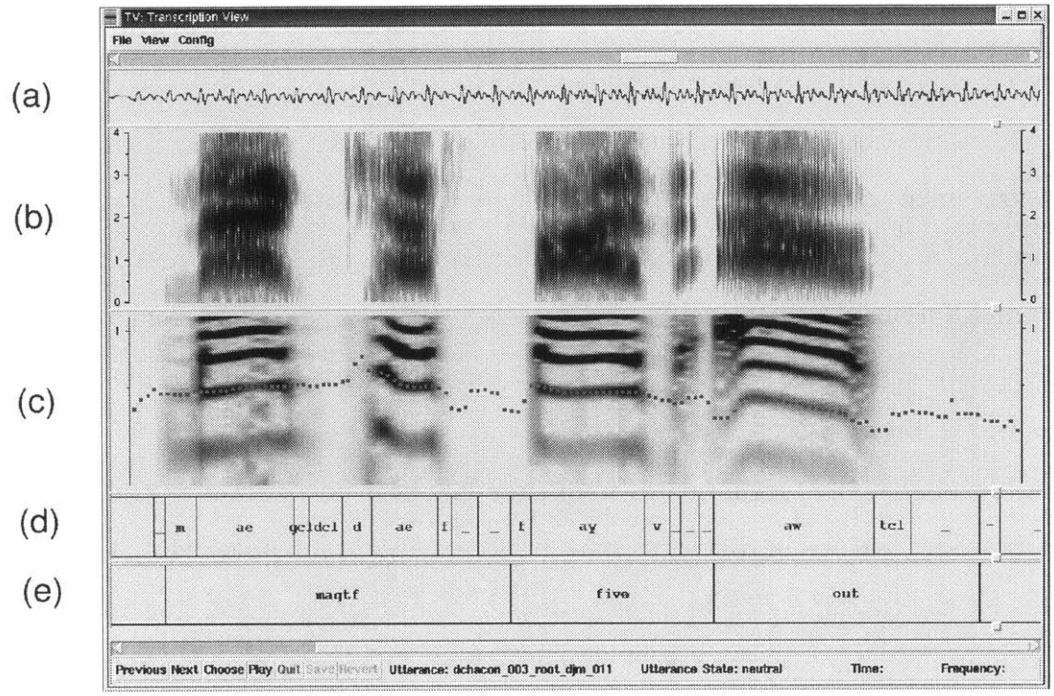



Figure 2-2: (a) waveform, (b) wide-band spectrogram, (c) narrow-band spectrogram in logarithmic frequency scale with pitch extracted using CPDA, (d) phonetic alignment, and (e) word alignment for the utterance "*magtf* (pronounced /mæg´ dæf/) *five out*" spoken by a female speaker.

second harmonic. Therefore, all pitch estimations must first be halved for analysis purposes. Also, notice that the chosen $F_0$ path during unvoiced regions and frication is random and scattered. These values are not meaningful, but are a reflection of the fact that the algorithm always proposes a value for $F_0$.

Using this visual representation as a guide, changes in the prosodic parameters within a repeat $n$-tuple could be observed initially and then later quantitatively analyzed using statistical analysis. Once the pitch estimation was determined, the extracted pitch values allowed pitch contours to be plotted so that any differences within repeat sets were emphasized.

For example, in Figure 2-3, the fundamental frequency for the utterance "*defer*" spoken on three different instances, each consecutively following the other, is plotted. These three utterances form a triple repeat set. The first utterance has a higher overall fundamental frequency at the onset of the /ɝ/ in "*defer*" (0.2 seconds in the figure). Furthermore, the first utterance, which is also designated as the neutral utterance, is shorter than the more frustrated utterances or the utterances immediately following it. This would indicate that there is a decrease in the speaking rate as the utterance is repeated. In more severe cases of vowel or utterance lengthening, the user may be hyperarticulating words, which would certainly lead to performance degradation.

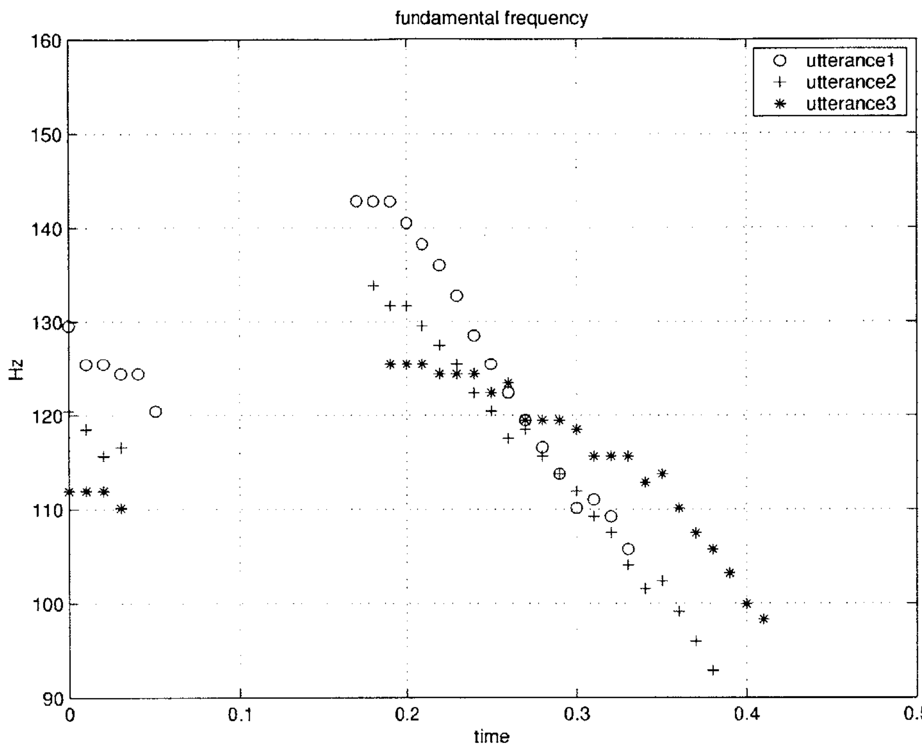



Figure 2-3: Fundamental frequency contours of the utterance "*defer*" spoken three times, sequentially by a male speaker. The /ɝ/ of "*defer*" begins at about 0.2 seconds.

## 2.3 Methods

Using the parallel utterances as the basis for this research, several methods were utilized to analyze the acoustic and prosodic cues for emotional speech. First, a visual inspection of the waveform was done to identify specific features of emotional speech by comparing different parallel utterances with differing emotional states. Second, a statistical analysis was performed using the features that were visually observed. Next, classification experiments were conducted to test how well our algorithm discriminated between the emotional and non-emotional speech using the extracted features. Finally, a human listener study was conducted to assess the reliability of the emotional labels and to provide a comparison to the classification results. These methods will be described in more detail in Chapter 3.

## 2.4 System

The classification and recognition tasks are configured using the SUMMIT system. SUMMIT is an automatic speech recognition system that utilizes acoustic models, language models, and pronunciation models to determine the best hypothesis for a spoken utterance. Unlike many of the current speech recognition systems which use hidden Markov models, the SUMMIT system is segment-based; that is, it focuses on phonetic segment boundaries corresponding to specific acoustic events to extract features.

## 2.5 Tasks

For the classification and recognition tasks, most of the features were characterized initially with respect to the vowels in the signal. Emotion classification involves determining the emotional properties within the utterance, in particular the vowel segments, given the signal and phonetic alignments. Classification, which also involves acoustic modeling and feature extraction, is dealt with during the training phase of the experiment. Emotion recognition consists of determining the emotional state of an utterance by utilizing the acoustic models and combining classification results. For this thesis, classification was a binary decision, i.e., an utterance was recognized as either possessing emotion, frustration in our case, or not possessing emotion.

### 2.5.1 Feature Extraction

The primary goal of feature extraction is to determine a set of features that will maximize the differences among different classes. Six prosodic features were extracted during the classification and recognition tasks. These prosodic features were initially computed as vowel-level features and then later manipulated to corresponding utterance-level features.

The six features investigated for maximum discrimination were average pitch, average energy, maximum pitch, pitch movement, average vowel duration, and the difference between the maximum pitch value in the first and last vowels in the utterance. Of these features, one was related to the extrema of the utterance, three were averages, and the remaining two were features associated with how the pitch moved throughout the utterance. The classification system uses linear discriminate analysis (LDA) to determine which features best discriminate between emotional and non-emotional utterances. These features will be defined in more detail in Chapter 3.

### 2.5.2 Evaluation

The evaluation of the classification and recognition experiments is based mainly on the error rate as it relates to the threshold value that was computed. This thesis focuses more on identifying the acoustic correlates of an emotional utterance as opposed to the more challenging task of improving speech recognition performance. Improvements in recognition performance will hopefully be undertaken in the future.

## 2.6 Summary

In this chapter, the experimental methodology is presented as an approach for identifying and investigating linguistic and prosodic features that are correlated with emotion. Three corpora are highlighted, with their data collection, transcription techniques, and corpus properties also being mentioned. A general overview of the continuous pitch detection algorithm and the classification and recognition systems used in this research are also presented.

# Chapter 3

# Analysis of Speech Variations due to Emotion

The elation of the recent birth of a child or the frustration of experiencing a missed flight can be recognized without visual cues. For instance, imagine that you want to tell a friend who is 800 miles away of the delivery of your firstborn child, using the telephone as the medium for exchange of information. To portray the excitement of the occasion, one might speak with a higher pitch or the speaking rate may increase slightly. Regardless of what actually goes on, there is an extremely high chance that your excitement will be successfully conveyed to the listener.

As illustrated in this simple example, emotion can be perceived not only through gestures and facial expressions, but also with only the presence of auditory information. So the appropriate question then becomes, "How are we able to detect emotion with only audible clues?" There has already been some research done in this area. Schroder and his colleagues showed that the amusement expression, namely smiling, is visual and audible in speech [44].

This thesis explores those parameters that allow a listener to detect a perceived emotion in a speech signal. More specifically, this research will utilize the knowledge of how humans recognize emotion and attempt to characterize emotional speech for a computer. The goal of this thesis is to identify prosodic features and how they affect the perception of a user's emotion.

The methodology developed for this thesis is useful for comparing and analyzing speech variations due to emotion. The repetition of utterances structured essentially identically, by consecutively following each other and spoken by the same speaker, allow comparison of differences caused by variations in prosodic features only. Frequently, these prosodic variations correspond to a change in emotional state. An analysis of these repeat utterances serves as the basis of understanding and eventually characterizing emotional speech.

The remainder of this chapter will cover the various methods employed to identify extractable features to be used during the classification and recognition experiments. A discussion of the results for these experiments will also be presented. Section 3.1 covers the analysis techniques that were used during visual inspection of the data. A discussion of the spectrograms and prosody-based tools, as well as the set of features chosen for maximum discrimination is presented.

Section 3.2 gives an overview of the classification experiments, detailing the training and testing conditions used for both the real and acted data. The results of the recognition tasks are presented as well. Section 3.3 presents ROC plots for the acted data, and the summary in Section 3.4 concludes this chapter.

## 3.1 Visual Inspection

While the presence of emotion may be perceived subjectively by a human listener, a method of visually displaying extractable speech features is important for finding objective measurements, especially for later use in an automatic emotion recognition system. One of the most common means for studying sound is by using a spectrogram. A spectrogram is a visual representation of sound that is created by displaying a frequency-time plot of intensity encoded on a gray-scale.

Another way to display the emotional characteristics, in particular the prosodic characteristics, in the speech waveform is with fundamental frequency contours. In this research, the fundamental frequency contours were computed using a pitch detection algorithm designed for continuous speech. For details about this algorithm, please refer to [49].

Visual inspection of the waveforms was an important step in finding promising features for the classification task. Consequently, hundreds of dialogues were first qualitatively analyzed and utterances inspected using spectrograms and prosody-based tools in order to resolve the subjective properties of emotion into measurable, physical properties.

### 3.1.1 Linguistic Analysis

During the early stages of this thesis, the MERCURY corpus was of primary interest because it was observed to have both linguistic and acoustic cues for emotion. An example dialogue from the MERCURY corpus that was labeled as frustrated is shown in Table 3.1. Using just the context information, hopefully it is fairly evident that the user had some difficulty with the dialogue system. Upon inspection of this dialogue, a number of interesting linguistic features were observed. One feature that was found was the use of negative constructors, i.e, NOT X, Y. In this example, the user uses this sentence structure to change the arrival destination.

Another linguistic feature that was observed during several of the MERCURY dialogues was repeated requests. There are also several examples of repeated requests in the sample dialogue. In general, it was observed that recognition error prompted repetition of information. These repetitions prompted an interesting approach for characterizing emotional speech. For instance, the repetition of "I'm not returning" motivated the study of other parallel utterances. 50 parallel utterances were found in the MERCURY corpus, and it was often found that the repeated utterances had a more frustrated emotional state. Many users emphasized misrecognized words or altered the structure of the sentence. Finally, the example dialogue shows a premature termination command. Although this linguistic feature is useful for analysis purposes, the result usually ends in the dialogue being terminated and the intended goal being unsuccessful. Other linguistic features that were observed in the MERCURY corpus were an increase in the frequency of correction commands, swear words, vocabulary indicating irritation such as "moron" and "frustrating", interrogative statements, and

| | |
|---|---|
| USER: | NEW USER |
| MERCURY: | *Okay, I'll enroll you temporarily as a guest user. How can I help you?* |
| USER: | UH I'D LIKE TO KNOW THE CHEAPEST FLIGHT TICKET FROM BOSTON TO KOREA SEOUL |
| MERCURY: | *I'll look for the cheapest fare when I price your itinerary. Okay from Boston to Riga. What date will you be traveling?* |
| USER: | LET'S SAY JUNE UH FIFTH NO NO JUNE TENTH |
| MERCURY: | *Okay, from Boston to Riga on Sunday June 10. Can you provide the approximate departure time or airline?* |
| USER: | UH NO BUT ACTUALLY IT'S NOT BOSTON TO WHATEVER CITY YOU THINK IT'S BOSTON TO SEOUL KOREA |
| MERCURY: | *Sorry, I'm not sure what you said.* |
| USER: | I WANT TO KNOW THE FLIGHT FROM BOSTON TO SEOUL KOREA |
| MERCURY: | *Okay from Boston to Seoul. What date will you be traveling?* |
| USER: | UH AS I SAID LET'S TRY JUNE TENTH |
| MERCURY: | *Okay, from Boston to Seoul on Sunday June 10. Can you provide the approximate departure time or airline?* |
| USER: | NO |
| MERCURY: | *Okay, flights from Boston to Seoul departing between 5:00 am and 11:00 pm on Sunday June 10. One moment please .... There are no nonstop flights. I have flights on American, Delta, Korean Air, Northwest, Singapore Airlines, and United. The first flight leaves at 9:30 am and arrives at 6:30 pm. You arrive the following day.* |
| USER: | SO THAT'S HOW LONG IT TAKES LIKE HOW LONG IS THE FLIGHT IF YOU START AT NINE AND ARRIVE AT SIX |
| MERCURY: | *Sorry, I'm not sure what you said.* |
| USER: | HOW LONG IS THE FLIGHT |
| MERCURY: | *Okay, flights from Boston to Seoul departing between 5:00 am and 11:00 pm on Sunday June 10. One moment please ... There are no nonstop flights. I have flights on American, Delta, Korean Air, Northwest, Singapore Airlines, and United. The first flight leaves at 9:30 am and arrives at 6:30 pm. You arrive the following day.* |
| USER: | OKAY I'LL ARRIVE ONE DAY LATER I UNDERSTAND ANYWAYS OKAY SO HOW MUCH IS KOREAN AIR |
| MERCURY: | *Sorry, I'm not sure what you said.* |
| USER: | HOW MUCH IS THE FLIGHT TICKET |
| MERCURY: | *I heard you ask about fares. After you have selected the flights for your itinerary, I will be able to lookup the fare. Is there something else I can do for you?* |
| USER: | OKAY KOREAN AIR JUNE TENTH |

| | |
|---|---|
| MERCURY: | *Okay, US Airways flights from Boston to Seoul departing between 5:00 am and 11:00 pm on Sunday June 10. One moment please ... There are no nonstop flights. I have a flight connecting in New Yorks LaGuardia leaving at 9:00 am and arriving at 4:30 pm. You arrive the following day? Would that work?* |
| USER: | OKAY |
| MERCURY: | *US Airways flight 6807 connecting to Korean Air flight 82 in New Yorks Kennedy has been added to your itinerary. You depart Boston at 9:00 am and arrive in Seoul at 4:30 pm. You arrive the following day. What date will you be returning on?* |
| USER: | UH I'M NOT RETURNING |
| MERCURY: | *I'm sorry, but I didn't catch the date. Could you please repeat it please?* |
| USER: | I'M NOT RETURNING |
| MERCURY: | *Okay, from Seoul to Boston on Thursday July 19. Can you provide the approximate departure time or airline?* |
| USER: | NO FORGET IT BYE |
| MERCURY: | *Sorry, I'm not sure what you said.* |

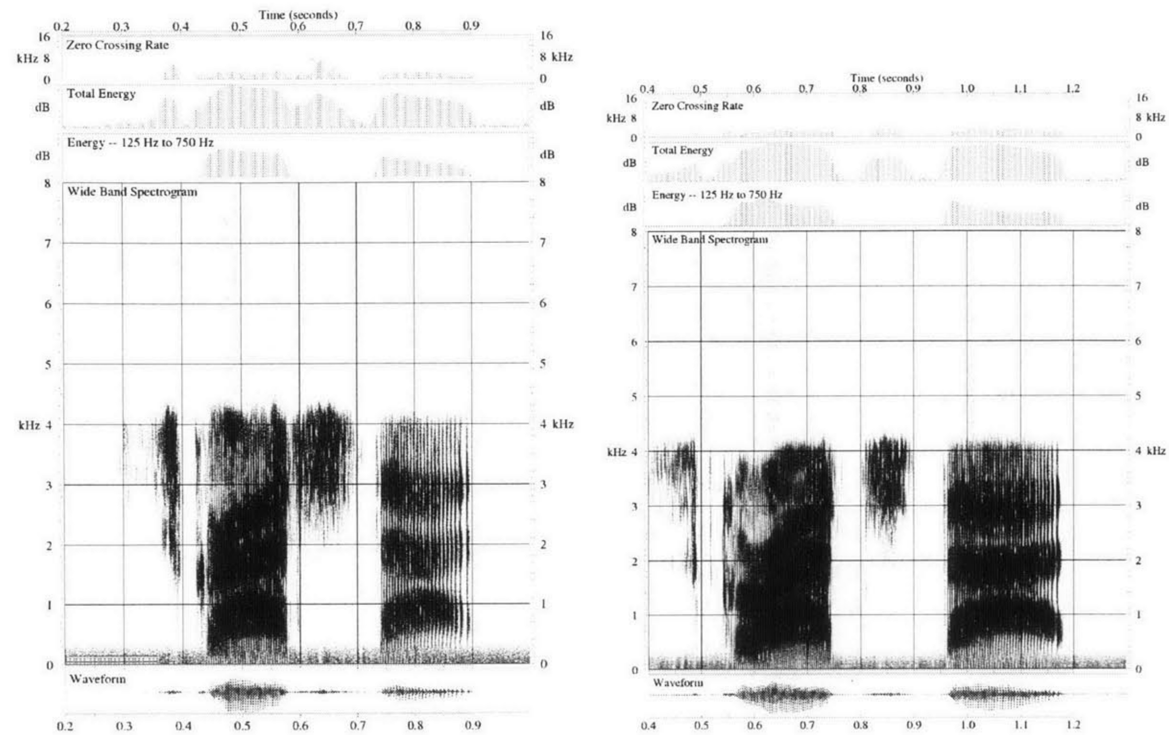
Table 3.1: Example of a frustrated dialogue in the MERCURY flight travel domain. This dialogue includes linguistic features that were helpful in determining promising features for emotion classification.

variations in the MERCURY system commands.

In addition to finding linguistic features that signaled a user was frustrated, acoustic features that were observed also told a similar story. Some of the acoustic features that were present in the MERCURY data and then later investigated in the other corpora were changes in fundamental frequency, energy, hyperarticulation, breathy sighs, and changes in speaking rate. By making use of both the linguistic and acoustic features, we were able to qualitatively assess the effect that emotion has on recognition.

### 3.1.2 Spectrographic Analysis

Spectrograms are a two-dimensional representation of the short-time Fourier transform and provide an effective way to analyze signal characteristics. Figure 3-1 shows spectrograms of the utterance "scratch that" spoken by a female speaker. The vertical axis represents the frequencies of the spectral components, the horizontal axis represents the time, and the gray scale represents the relative intensity of the sound




(a) neutral (b) frustrated

Figure 3-1: Spectrograms of the utterance "scratch that" for different emotional states spoken by a female speaker in the Lockheed corpus. Because spectrogram (b) was a repeat of (a), with the same linguistic content but differing emotional states, this allowed an analysis of variations in prosodic features caused by emotion.

at a particular frequency and time. In general, the darker the color the higher the amplitude of the spectral components at that frequency. In addition, plots of zero crossing rate, total energy, and low-frequency energy are included at the top of the spectrogram, while the waveform is displayed at the bottom. The utterance is transcribed as /sk˺ræč ðæt/.

In comparison to the neutral spectrogram, there is an increase in total energy in the frustrated spectrogram, which is seen in the darker intensity. There is also a significantly longer gap between the two words in the frustrated spectrogram. As a listener, an increased gap would be perceived as a pause. Additionally, the vowel /æ/ in both "scratch" and "that" has been lengthened.

During the course of this research, an interesting phenomenon was observed when utterances were repeated. In general, the repeated utterance was longer in duration than the originally spoken utterance. This occurred because the user clarified or stressed the part of the utterance that was misrecognized, usually speaking in a frustrated emotional state. Often this meant that the user decreased his speaking rate. The increase in utterance duration was the result of a combination of an increase in internal pause duration and vowel lengthening. These events can be seen directly in figure 3-1.

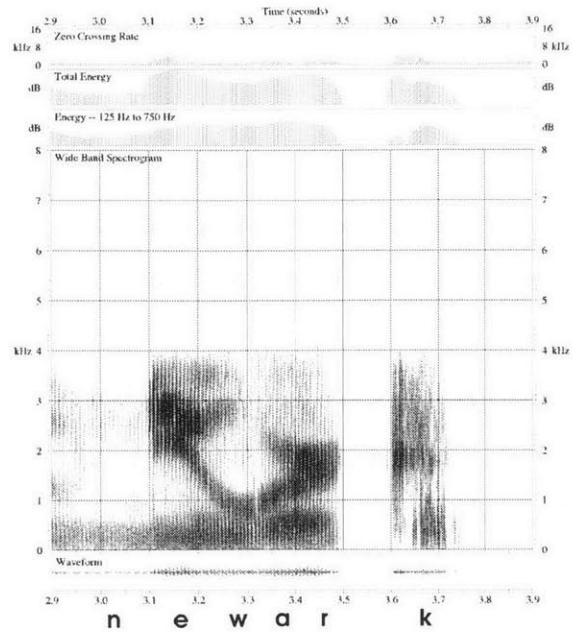
In extreme cases of vowel lengthening, the user over-articulates words, resulting in further system degradation. Figure 3-2 shows three separate spectrograms of the words "newark[1]" and "new york[2]" extracted from utterances spoken by the same speaker in the MERCURY corpus. The transcription for "newark" and "new york" is /nu$^{w}$(ɑ)rk/ and /nu$^{w}$ yɔrk/, respectively.

At some point during the dialogue, the user becomes so annoyed with the system that he begins to hyperarticulate the words in the repeated utterance. In the neutral spectrograms, there is no distinguishable syllable or word boundary. However, in the frustrated spectrogram, there is a clear syllable boundary around 4.1 seconds separating the first syllable from the last syllable in "newark." In comparison to (a), the spectrogram in (b) also has more total energy, especially in the /k/ at the end of the word. Furthermore, the third formant in the vowel at 4.2 seconds has a steeper fall than in (a). When speaking naturally, the "a" in "newark" is pronounced /ɝ/, but when "newark" is hyperarticulated, the "a" now sounds like /ɑ/. The frustrated spectrogram now resembles two separate words which would present difficulty for the recognizer and aggravation for the user if he intended one word, such as in this case. The last spectrogram is presented for comparison of "new york" and the hyperarticulated "newark."

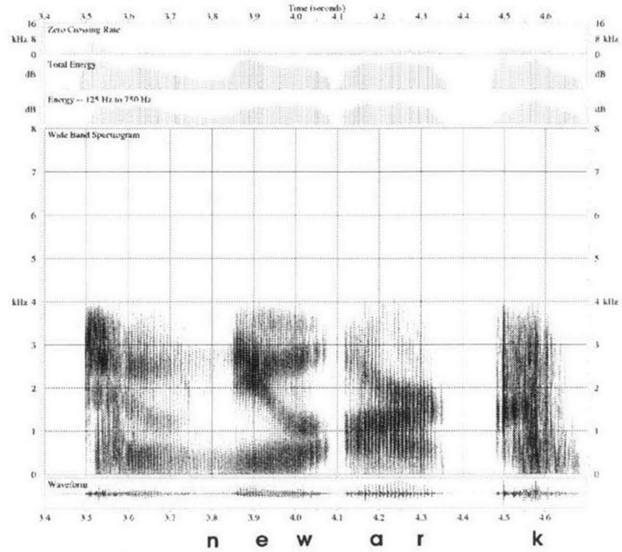
The spectrogram helps to visually display how hyperarticulation alters a word or

[1] The neutral utterance is "i want to go from los angeles to newark then directly on to london on united airlines." The frustrated utterance is "i want to go from los angeles to newark from newark to london heathrow." The speaker is male.

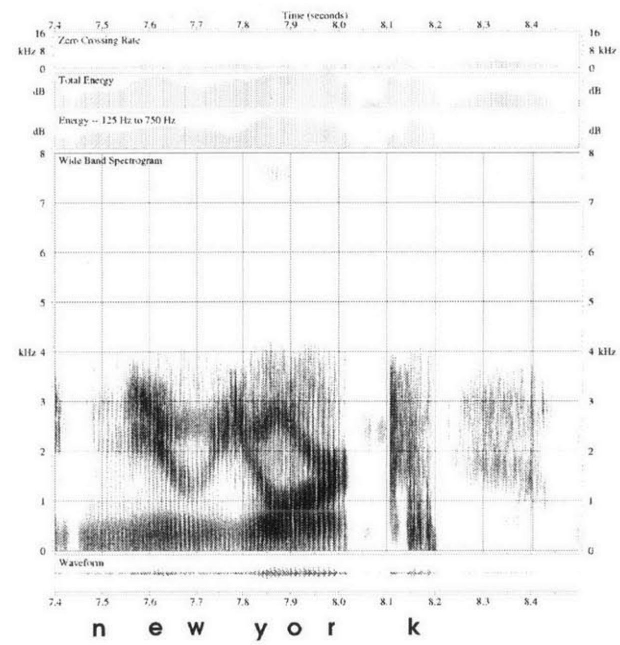
[2] The utterance is "well i wanted to make a connection from los angeles to newark and the system kept hearing me say new york and i finally said newark." This utterance is a comment to the system.




(a) *newark*, neutral




(b) *newark*, frustrated




(c) *new york*, neutral

Figure 3-2: Spectrograms illustrating differences between a normal pronunciation (a) and hyperarticulated pronunciation (b) of the word "newark." Spectrogram (c) shows a normal pronunciation of the word "new york" and is presented for comparison.

utterance. To an enraged user, clearly enunciating his words may seem like a good method for obtaining system recognition; however, as shown in figure 3-2, the result is often unfavorable and actually counteracts system recognition.

Another tool that was utilized during this research also made use of the spectrogram and additional prosodic information as well. Figure 3-3 shows a graphical viewing tool with a smoothed fundamental frequency contour and energy included for the same utterance, "scratch that," as shown in figure 3-1. This *Transcription View* tool shows not only spectrographic characteristics of the waveform but prosodic characteristics as well.

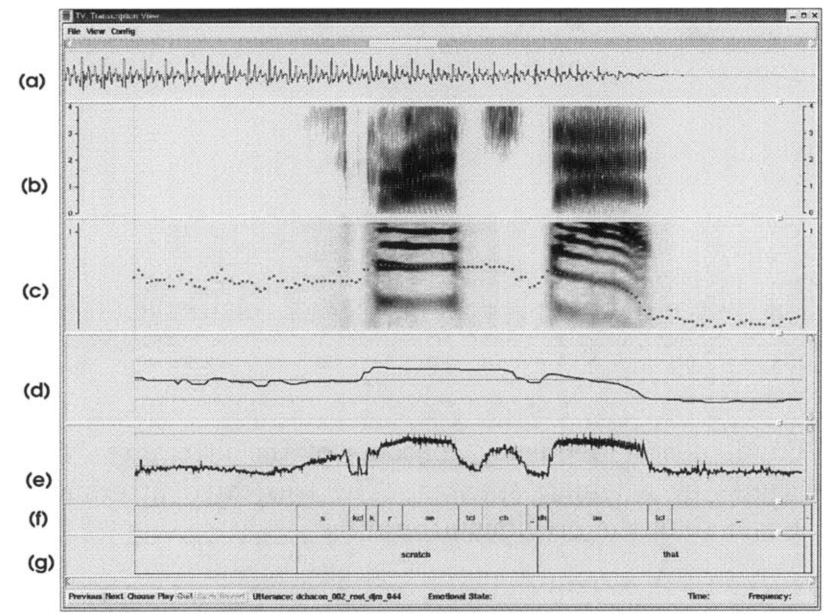



Figure 3-3: (a) waveform, (b) wide-band spectrogram, (c) narrow-band spectrogram in logarithmic frequency scale with extracted pitch, (d) smoothed pitch contour, (e) energy, (f) phonetic alignment, and (g) word alignment for the utterance "*scratch that.*"

### 3.1.3 Prosodic Analysis

The prosodic characteristics of the emotional speech in our corpus were explored using a graphical transcription tool. To compute physical, objective features, the fundamental frequency was extracted and analyzed. In Figure 3-4, the fundamental frequency for the utterance "two one six zero[3]," spoken by the same speaker, is plotted. Utterances two and three are repeated utterances with an emotional state of frustrated, while the first utterance is neutral.

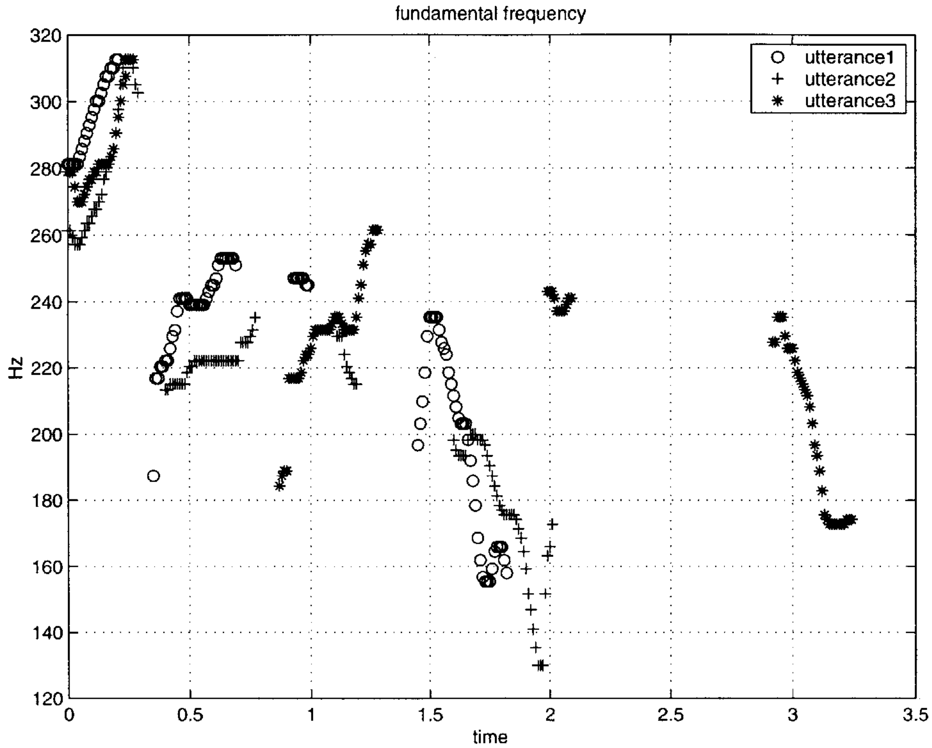



Figure 3-4: Fundamental frequency contours of the utterance "*two one six zero*" spoken three times by a female speaker. The frustrated utterance, utterance3, is significantly longer than the previous utterances.

The first noticeable difference between the three utterances is the utterance duration. As the utterance is repeated, the duration tends to increase. The third utterance is significantly longer than the previous utterances and has clear word boundaries. The second utterance has a slightly lower fundamental frequency than the other two utterances. Interestingly enough, this was the case with several of the repeat groups.

However, some people expressed their frustration by lowering their fundamental frequency, which caused the speaker to sound more depressed rather than frustrated. On the other hand, other speakers increased their fundamental frequency to portray their irritation with the system. In fact, while analyzing utterances that were

[3]The utterance is from the Lockheed corpus. The speaker, dchacon, is female.

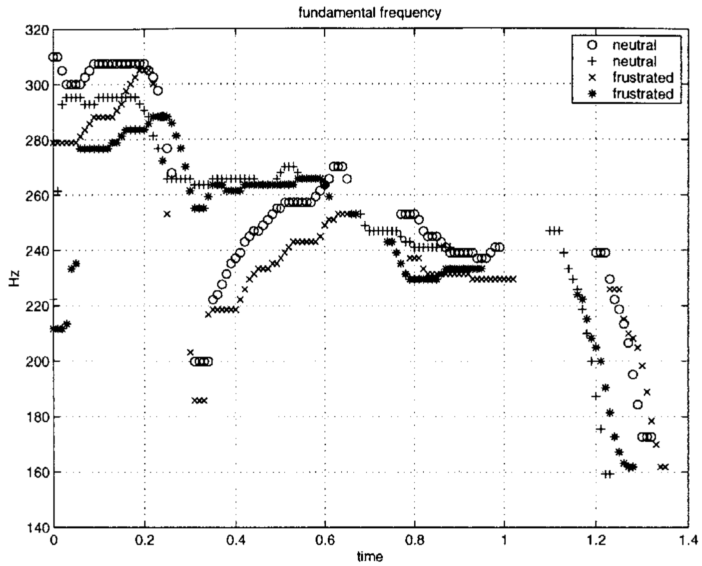
unelicited, there was no noticeable pattern observed for the fundamental frequency. However, when using data collected from the LDC corpus, the fundamental frequency generally increased when the speaker acted in a frustrated emotional state. A few examples from data collected from the Lockheed and LDC corpora are presented to illustrate this finding.

Figure 3-5 shows three fundamental frequency contours of the utterance "two one five six" from different speakers in the Lockheed corpus. In (a), there is very little difference between the fundamental frequencies of the neutral and frustrated emotional states. To detect emotion in the speech of this particular speaker, one would need to rely on other prosodic features besides fundamental frequency.

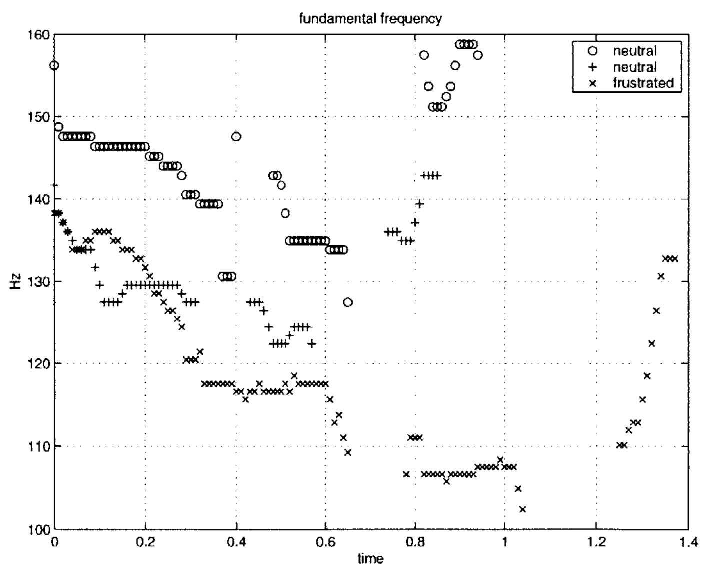
However, in the male speaker, there is a significant difference in the fundamental frequency between emotional states. In contour (b), the frustrated emotional state shows an increase in utterance duration, but the fundamental frequency is lower than in the neutral utterances. Generally, we would expect a frustrated user to shout or speak firmly with the system. Graphically, this would correspond to an increase in the fundamental frequency. However, as shown in (b), with "real-life" data, this may not occur.

In order to make some comparison between the acted and real data, utterances were chosen from the LDC corpus with either exact word combinations or similar word-initial and word-final phones to eliminate differences in fundamental frequencies. For instance, since the utterance "two one five six" was analyzed in the Lockheed corpus, an attempt was made to find an utterance that contained /$u^w$/ and /ɪ/. Instead, an utterance that contained /$u^w$/ and /ɛ/ was chosen from the LDC corpus, since no exact match was found. The resulting utterance was chosen to be "two thousand ten" for comparison.

Upon initial inspection, there are some clear differences between the real data and acted data, as shown in Figure 3-6. As shown in fundamental frequency contours (c) and (d), in the initial /$u^w$/ in the contour plots of the acted data, the fundamental frequency contours of the frustrated utterances are significantly higher than those of the neutral utterances. The large shift in fundamental frequency is indicative of




(a) female speaker




(b) male speaker

Figure 3-5: Two separate sets of fundamental frequency contours for the utterance "two one five six" in the Lockheed corpus. There was no observable pattern for the fundamental frequency of a frustrated emotional state. In contour (a), the speaker does not heavily rely on $F_0$ to express frustration. The speaker in (b) lowers his fundamental frequency to express frustration.

an exaggerated change. These changes are so extreme that a human listener would have no problem identifying the intended emotional state. The /ɛ/ at the end of the utterance appears to be reduced in the plots, and the utterance duration once again is lengthened.

Even in cases where the phones in the utterance were completely different, this finding still held. Figure 3-7 shows $F_0$ contours for the utterance "eight hundred one" for two separate speakers. Once again, the fundamental frequency increases when the emotional state changes from neutral to frustrated, and the utterance duration increases in contour plot (a).

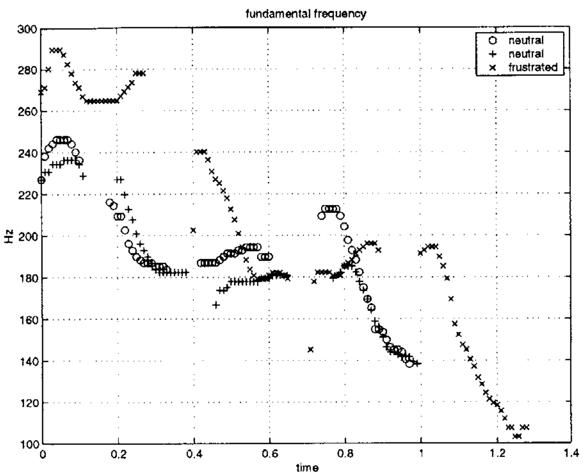
Using similar examples, it became apparent early on that an increase in fundamental frequency, often found in emotional data recorded by an actor, did not accurately categorize much of the data that was collected in our corpus. However, we were able to make use of the variance in the fundamental frequency and propose appropriate features for our classification experiments.

### 3.1.4 Feature Extraction

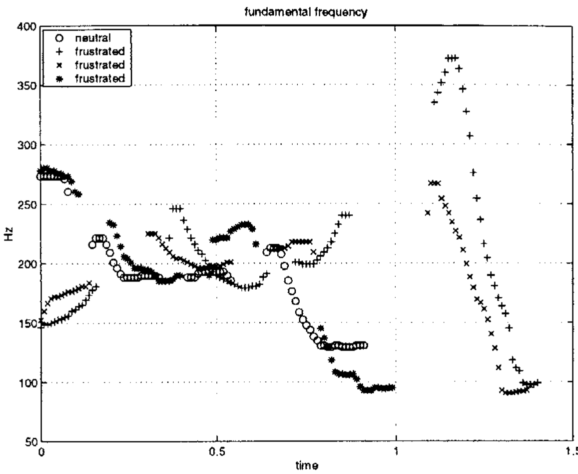
Linear feature extraction can be thought of as finding a set of vectors that effectively capture the information present during an experimental observation while reducing the dimensionality. For this research, we desire to extract features which focus on discriminating among emotional and non-emotional speech.

### 3.1.5 Vowel-level Features

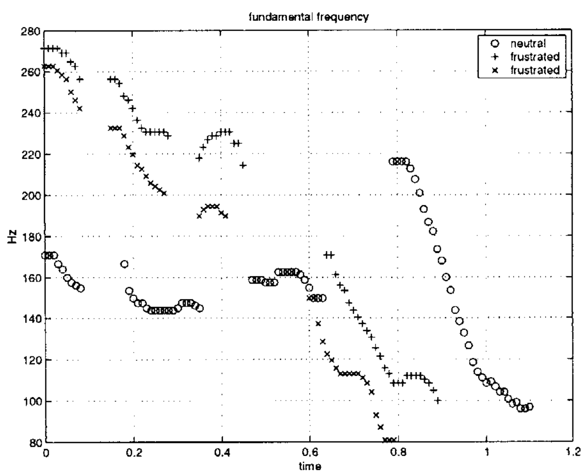
The vowel-level features were computed as an intermediate step toward finding utterance-level features. During the analysis stage of this research, vowel-level prosodic features for each utterance were computed. Using the illustration in Figure 3-8, the prosodic features will be presented for one vowel segment, which is easily extended to other segments in the utterance. The fundamental frequency contour is of the utterance "two thousand nine." The first vowel /uʷ/ is from time 0.0 to 0.2 seconds, the second vowel /ɑʷ/ from time 0.2 to 0.4 seconds, and so on.



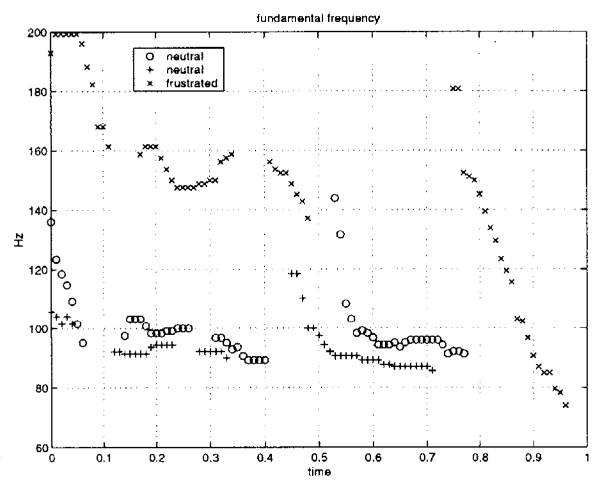

(a) female speaker




(b) female speaker

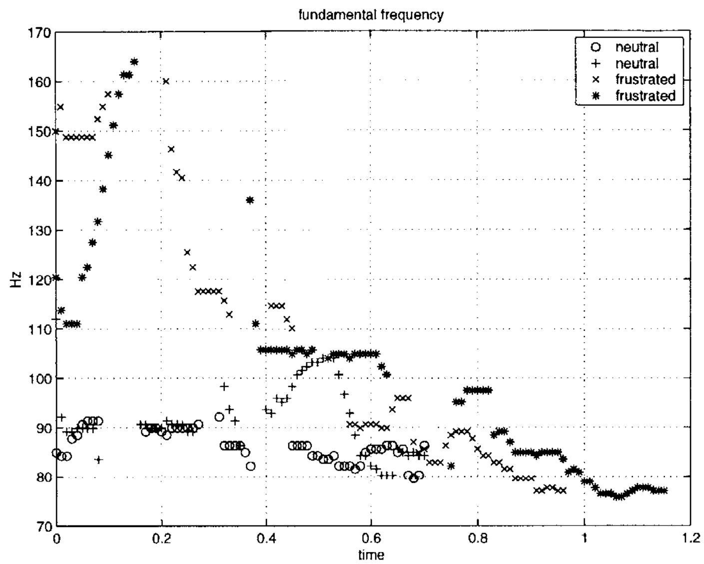



(c) female speaker

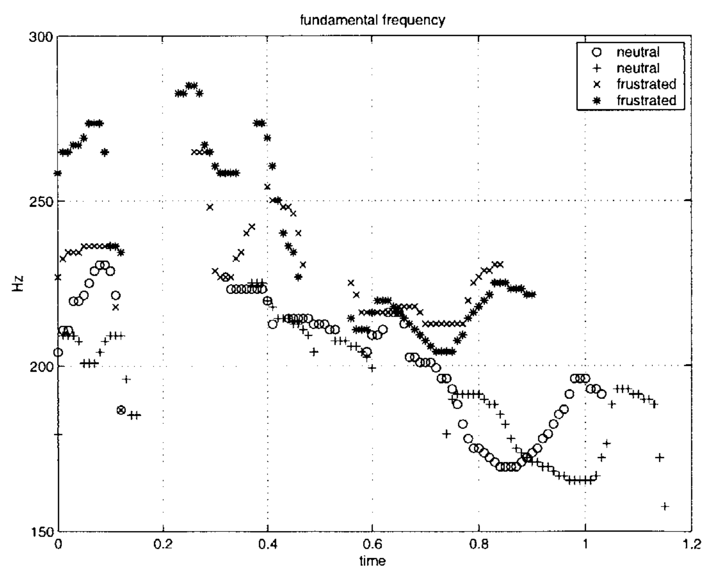



(d) male speaker

Figure 3-6: Four separate sets of fundamental frequency contours for the utterance "two thousand ten" in the LDC corpus. Unlike the real data, there is a significant difference in the frustrated and neutral emotional states, with the frustrated state usually being higher. Contours (a), (b), and (d) also show that the utterance duration increases for frustrated emotion.




(a) male speaker




(b) female speaker

Figure 3-7: Two separate sets of fundamental frequency contours for the utterance "eight hundred one" in the LDC corpus. Contours (a) and (b) show an extreme difference in $F_0$ between neutral and frustrated emotional states.

The four prosodic features that are labeled in the figure are (1) maximum pitch, (2) average pitch, (3) pitch slope, and (4) vowel duration. The pitch-based features were all normalized with respect to the mean pitch of the first utterance in the repeat series. It was felt that this is a realistic model for what could be done in a dialogue setting.

Table 3.2 shows the extracted measurements for each of the vowels in the utterance. The second row corresponds to the illustrated features.

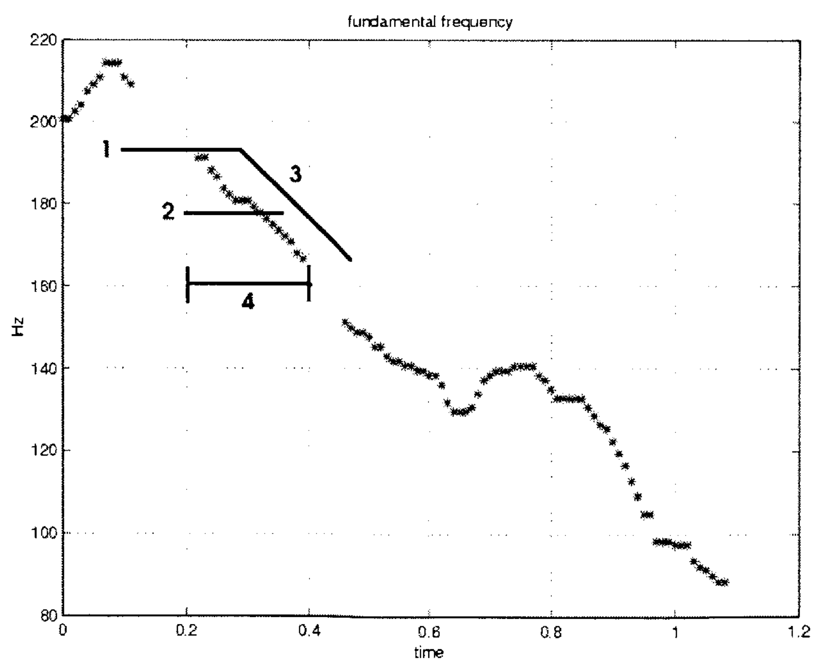



Figure 3-8: Illustration of the vowel-level prosodic features. The prosodic features that are highlighted are (1) maximum pitch, (2) average pitch, (3) pitch slope, and (4) vowel duration.

| max pitch | avg pitch | pitch slope | avg energy | avg vow duration |
|---|---|---|---|---|
| 214.4 | 208.3 | 4.1 | 143.8 | 120.1 |
| 191.3 | 179.2 | -7.1 | 169.0 | 187.5 |
| 151.2 | 140.0 | -6.5 | 127.2 | 229.3 |
| 140.5 | 124.7 | -12.7 | 147.0 | 243.1 |

Table 3.2: Example of extracted vowel-level features for a vowel segment in an utterance. These particular values are for the frequency contour plot in Figure 3-8. Row 2 corresponds to the highlighted vowel, where columns 1, 2, 3, and 5 represent the values of the illustrated features.

### 3.1.6 Utterance-level Features

The vowel-level features were then manipulated to compute the utterance-level candidate features for classification. The utterance-level features were normalized with respect to gender and the first utterance in a repeat group. In general, the first utterance was the most neutral of all the utterances in the repeat group. This normalization was done to attempt to adjust for the speaker dependencies of the pitch value.

A single emotion feature vector is then constructed from the set of utterance-level features that are computed. For this research, six different features which were thought to be correlated with emotion were utilized in the training process. These utterance-level features are as follows:

1. **max_pitch:** The maximum fundamental frequency of the utterance maximized over the maximum fundamental frequency in each vowel.

2. **avg_pitch:** The average fundamental frequency for the utterance derived from the fundamental frequencies of the vowels.

3. **pitch_movement:** The absolute value of the fundamental frequency for each

vowel in the utterance was computed, then averaged to obtain the total excursion of the fundamental frequency throughout the utterance.

4. **avg_energy:** The average energy for the utterance derived from the energy in the vowels.

5. **avg_vow_duration:** The average of all of the vowel durations in the utterance.

6. **delta_first_last :** The difference between the maximum fundamental frequency of the first and last vowels in the utterance.

## 3.2 Classification Experiments

Classification experiments are performed to analyze which prosodic features best discriminate between emotional and non-emotional data. The analysis focuses on six features in which one or more were consistently observed in the frustrated utterances of our corpora.

### 3.2.1 Emotion Classification

The classification method for this project was based on the technique originally used for confidence-scoring in the Spoken Language Systems (SLS) group [15]. Using this technique, a greedy search is performed over all the candidate prosodic features. Features are then iteratively added to a single emotion feature vector from the candidate set until the addition of features no longer improves performance. To produce a single emotion score for a hypothesis, a linear projection vector is trained and used to reduce the emotion feature vector. A probabilistic emotion score is then computed based on a maximum *a posteriori* probability (MAP) classification. During this process, a threshold is computed to achieve a local minimum error. A negative score results in a rejection while a positive score results in an acceptance, where acceptance means the utterance is emotional.

To accomplish the objectives of this project, several preliminary steps were completed to determine a set of prosodic features to use during classification. Simple linear discrimination analysis (LDA) was then used to construct a Gaussian model and obtain a probabilistic emotion score using prior probabilities and the computed threshold.

### 3.2.2 Training and Testing Conditions

Prior to training each utterance, the data are labeled as either *correct* or *incorrect*. In this experiment, the *correct* label corresponds to an utterance possessing emotion and an *incorrect* label to a non-emotional utterance. A cross-validation approach was used to train and test on the data. This approach was chosen since there was a limited amount of data. Our experiments were performed using the Lockheed and LDC corpora. Using the Lockheed corpus, the classification experiment was conducted using 2,247 utterances spoken by 12 different speakers. Eleven speakers were used for training and the remaining speaker was held out for testing.

The LDC corpus of acted data was also used in order to make comparisons between naturally elicited and acted data. The classification and recognition experiments using the LDC corpus consisted of 1,208 utterances spoken by 7 different speakers. To train the accept/reject classifier, a different speaker was held out for testing each time to ensure that the recognizer performance was predicted on unseen speaker data.

The training and testing conditions were first performed using a combination of all six of the utterance-level features. Based on the resulting weights calculated for each of the features, we were able to make an initial assessment of the best discriminating features. To evaluate the discrimination success of each utterance-level feature, each individual feature was trained separately on both of the corpora.

### 3.2.3 Results

Table 3.3 and Table 3.4 show the preliminary results of the classification and recognition tests for the real and acted data, respectively. The results obtained using the Lockheed data was, in general, inconclusive because there was such a limited amount of emotional utterances. In fact, almost half of the speakers had spoken fewer than two emotional utterances. As a result of this, the test set for these speakers was severely biased toward the neutral utterances, and the accuracy rate was not as meaningful. Table 3.3 lists only the speakers that had approximately 10% of their utterances as emotional.

Dch, the speaker with the most frustrated utterances, had an accuracy rate of 71.0%, which is consistent with other literary results on emotion recogniton. This performance rate is still not better than chance, and is most likely due to a combination of the speaker exhibiting emotion and being the only female in the classification and recognition experiments, such that the effect of training on only males and testing on a female hurt performance significantly.

| **Test Speaker** | **Test** | **Frustrated** | **Neutral** | **Accuracy** |
|---|---|---|---|---|
| dch | 290 | 84 | 206 | 71.0 |
| jgr | 258 | 22 | 236 | 91.5 |
| jmc | 170 | 22 | 148 | 88.7 |
| jri | 170 | 22 | 148 | 87.1 |
| kba | 192 | 18 | 174 | 92.7 |

Table 3.3: Performance results in cross validation experiments for Lockheed data. Only 5 of the 12 speakers were reported since the remaining 7 speakers had fewer than 10% of their total utterances as emotional.

For the Lockheed corpus, the emotion classification technique yielded that the `avg_pitch` feature was best at discriminating between emotional states, and `max_pitch` was second best. Generally speaking, although these features had the highest weighted value out of all of the extracted features, their weights were still relatively small.

The results for the acted LDC data are given in Table 3.4. Although the accuracy rates are not as high as those of the real data, these results are more significant and provide a better indication of how emotional utterances affect recognition performance because there is more emotional data. Furthermore, because this data was produced by actors, we can make a direct comparison with results from other literature. The average accuracy for all of the speakers is 70.1%. The performance on each of the speakers is more than 10% better than chance performance.

| Test Speaker | Test | Frustrated | Neutral | Accuracy |
|---|---|---|---|---|
| cc | 166 | 77 | 89 | 71.7 |
| cl | 120 | 48 | 72 | 83.3 |
| gg | 285 | 143 | 142 | 76.1 |
| jg | 219 | 115 | 104 | 64.8 |
| mf | 86 | 33 | 53 | 73.3 |
| mk | 90 | 40 | 50 | 75.6 |
| mm | 242 | 135 | 107 | 57.4 |

Table 3.4: Performance results on cross validation experiments for acted data. Overall, these results are consistent with other published results on emotion. The female speakers are gg, jg, mk, and mm.

It is interesting to note that the speakers who had the worst accuracy rates, jg and mm, are both female speakers. The other two female speakers in the experiment are gg and mk. Three of these female speakers also have the highest number of frustrated utterances. While this corpus contains only acted speech, this situation frequently occurs in real-life scenarios when a user interacts with a dialogue system. As system misrecognition occurs, the number of emotional utterances increases while recognition performance decreases.

For the acted data, the classification experiment yielded that `max_pitch` was the best at discriminating between our emotional states. The feature `avg_pitch` was second best, followed by `avg_energy`.

## 3.3 ROC Plots for Acted Data

It is interesting to examine the emotion recognition results for the acted data using a Receiver Operating Characteristic (ROC) curve. The ROC curve measures the trade-off between the acceptance of correctly identified emotional utterances (i.e. detection rate) and the false acceptance of utterances that were incorrectly identified as emotional (i.e. false alarm rate). Ideally, we would want to see more emotional utterances being correctly identified, and fewer neutral utterances incorrectly identified as emotional. This would correspond to the curve shifting toward the upper left of the plot.

Figure 3-9 shows the ROC plot for the acted data. The three best individual features and a combination of the top two discriminating features are plotted. It is clear from the figure that `max_pitch` does significantly better than `avg_pitch`, and furthermore, does as well alone as any combination of it with other features. Thus, although `avg_pitch` is the second best feature, it is too highly correlated with `max_pitch` to be of further use to the system. The system instead prefers to add the `avg_energy` feature as its second candidate. However, although `avg_energy` does significantly better than chance performance by itself, it is unable to further improve over the `max_pitch` feature acting alone. This can be interpreted as a positive result. That is, for a very simple scoring mechanism on a single feature, significant results can be achieved in discriminating emotion in acted speech.

Choosing an exact operating point on the curve, we can examine the relative capability of the individual features. From the figure, the best operating point for the `avg_energy` parameter is at around 65% correct acceptance. Using the `avg_energy` feature, the system would falsely recognize nearly 38% of the neutral utterances as being emotional. However, using just the `max_pitch` for maximum discriminability, we could reduce the false acceptance rate to 30%.

These results indicating that the maximum pitch, average pitch, and average energy are good discriminators of emotion is not surprising, especially since these findings are a result of using acted speech. Recall from earlier sections in this thesis, that the excursion between the fundamental frequencies of real and acted data was dramatically different. Frequently in acted data, the most heavily relied-upon features for expressing emotion is that of fundamental frequency and energy. This also agrees with other literary works that recognize the fundamental frequency as the most reliable feature for detecting emotion. However, the fundamental frequency, specifically the maximum pitch, is unfortunately not a good discriminator for real data.

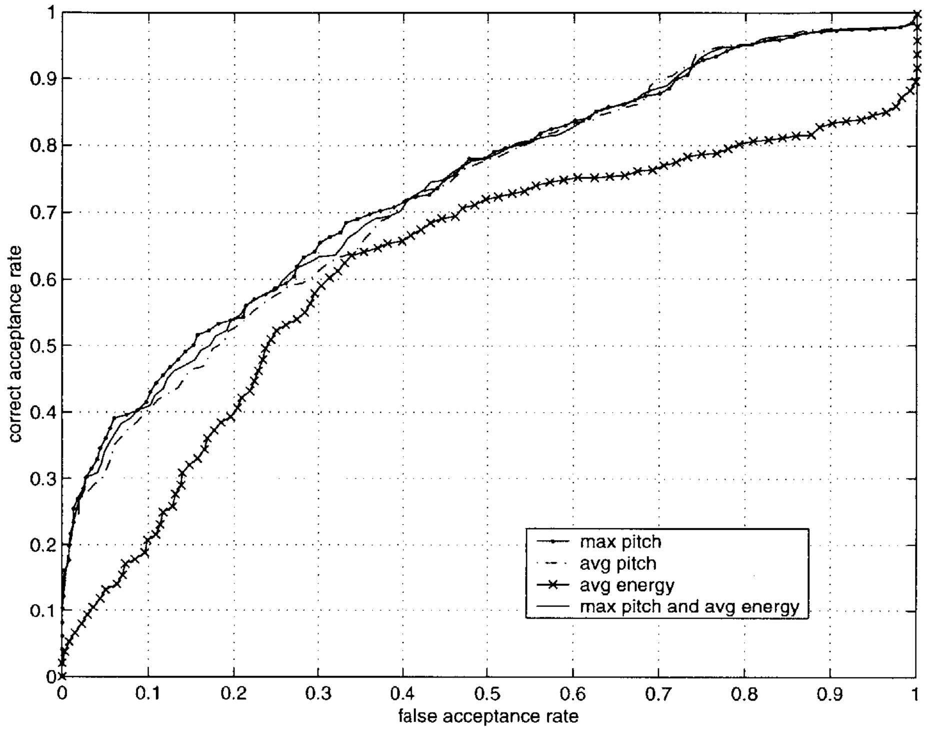



Figure 3-9: ROC curve for the LDC data. This curve was produced by three individual prosodic features and by a full-trained model using the best two discriminating features max_pitch and avg_energy.

## 3.4 Summary

This chapter describes the various methods that were employed in this thesis to identify prosodic features for emotion classification and recognition. Using repeat utterances as the basis of our investigation, an analysis of speech variations due to emotion was performed using dialogues to capture linguistic information, spectrograms to illustrate changes in the energy or utterance duration, and prosody-based tools to emphasize the changes in fundamental frequency and duration. A discussion is given on how the extracted features were chosen as well as a description of each. The classification and recognition tasks are presented and results are discussed and evaluated.

While we were successful in building a model that could discriminate acted emotional and non-emotional speech, the model was unable to make those distinctions on the natural emotional speech. In the next chapter, we will see that human listeners are also not particularly consistent in their judgments of emotional content of natural speech drawn from spoken dialogue interaction, even when given the arguably easier task of discriminating minimal pairs, with the same speaker and linguistic content.

# Chapter 4

# Human Listener Studies

The corpora for this thesis were manually created by relying on a single individual (myself) to classify the appropriate emotional state for each utterance. Since there was no collaboration on ambiguous utterances or subtle differences that may have been perceived, it is quite easy to comprehend that this method of tagging the emotional state of an utterance might have been problematic. An utterance that was perceived to be frustrated was assumed to be so throughout the analysis, then later used to determine the prosodic features for classification. However, if there was a more likely emotional state for the utterance, then the results from the classification experiments would not be as reliable. To counter the subjectivity of one individual tagging all utterances and to assess the quality of the corpora, a listening study was performed with human subjects.

## 4.1 Experimental Setup

Listening experiments were performed with 6 human subjects, three male and three female, who had not previously listened to utterances in the corpus. These subjects were asked to rate the expressed emotion of an utterance by comparing it to the first utterance in its repeat group. Four different repeat groups were used during this experiment.

For any single repeat group, the subject was given a set of instructions concerning their task. A screen snapshot of the interface is displayed in Figure 4-1. Two utterances are played successively, with one utterance guaranteed to be the first utterance of the set. The order in which the utterances are played was randomly chosen. The subject is allowed to listen to the pair of utterances as many times as needed by clicking the *Repeat* button.

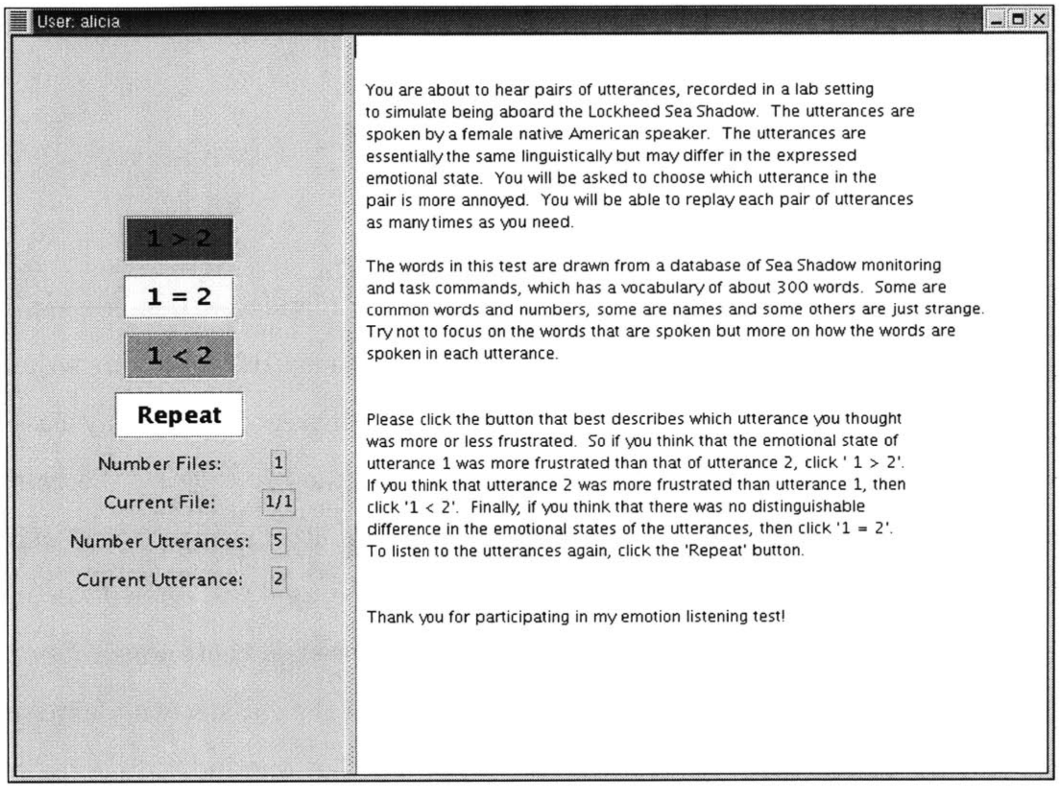



Figure 4-1: Interface for Human-Listener Experiment.

After listening to the utterances, the subject is then prompted to click the corresponding button that best describes which utterance he or she thought was more or less frustrated. The subject is asked to click "1 > 2" if he thought that the first utterance played was more frustrated than the second. On the other hand, if he thought that the second utterance that was played was more frustrated, then he was prompted to click "1 < 2". If the speaker judged that there was no distinguishable difference in the emotional states of the utterances that were played, then he was asked to select "1 = 2".

Sennheiser headsets and the audio capabilities of a Dell computer were used for conducting these experiments. Three of the four lists consisted of utterances that were repeated with almost exactly the same linguistic content. The fourth list, however, consisted of utterances that were fragments that contained the same linguistic message as the first utterance, in an elliptical form.

The goal of this study was to evaluate if other subjects could detect changes in emotional state when utterances were repeated in succession.

## 4.2 Results

This experiment yielded some useful information. First, for all repeat groups except one, the emotional state of the repeats was rarely judged to be less emotional than the first utterance. It is clear from the wide variability in the results that the emotional content of these utterances is not very clearly manifested, and this may partially explain the inability of the system to discriminate the emotional content as well.

| Utterance | More | Less | Same |
|---|---|---|---|
| 2 | 0.50 | 0.33 | 0.17 |
| 3 | 0.67 | 0.17 | 0.17 |
| 4 | 0.67 | 0.33 | 0 |
| 5 | 0.17 | 0.50 | 0.33 |

Table 4.1: Results from group 1 of human-listener study.

| Utterance | More | Less | Same |
|---|---|---|---|
| 2 | 0.33 | 0 | 0.67 |
| 3 | 0.67 | 0 | 0.33 |
| 4 | 0.33 | 0 | 0.67 |
| 5 | 0.67 | 0 | 0.33 |

Table 4.2: Results from group 2 of human-listener study.

| Utterance | More | Less | Same |
|---|---|---|---|
| 2 | 0.83 | 0 | 0.17 |
| 3 | 0.67 | 0 | 0.33 |
| 4 | 0.17 | 0 | 0.83 |
| 5 | 0.50 | 0 | 0.50 |

Table 4.3: Results from group 3 of human-listener study.

| Utterance | More | Less | Same |
|---|---|---|---|
| 2 | 0.33 | 0.33 | 0.33 |
| 3 | 0 | 0.50 | 0.50 |
| 4 | 0.33 | 0 | 0.67 |
| 5 | 0.67 | 0 | 0.33 |
| 6 | 0.17 | 0 | 0.83 |
| 7 | 0 | 0 | 1.00 |
| 8 | 0.33 | 0.17 | 0.50 |

Table 4.4: Results from group 4 of human-listener study.

# Chapter 5

# Conclusion

In this thesis, we explored characteristics of emotional speech by using the spectrogram and prosodic tools. Three corpora were created, of which the Lockheed and LDC corpus were used for a comparative study between real and acted data. Six prosodic features were extracted from the collected data and studied to determine their potential to discriminate among frustrated and non-emotional speech.

An experimental methodology was designed for optimal comparison of the acoustic correlates of emotion. It was observed early on that speakers often repeated words or utterances when misrecognition occurred. The repetitions were usually structured the same linguistically, and frequently a change in emotional state also occurred. Consequently, using these parallel utterances helped isolate the expressed emotional state and eliminated variabilities due to variations in speaker and linguistic content.

Most of the analysis relied on spectrographic and prosodic analysis to determine a set of features for classification. In our investigation, we found that results normally reported in other literature, although encouraging, are an unrealistic comparison for classification and recognition experiments with real data. In particular, the fundamental frequency is a good discriminator for acted speech, but not necessarily for real data.

The classification performance on acted speech was significantly better than that on the natural emotional utterances, suggesting that acted emotion may be an unrealistic model of the variability encountered in real emotional expression. We found that a combination of the maximum pitch and average energy features did the best job of discriminating between emotional and non-emotional speech.

## 5.1 Future Work

Because the data collection required such a great deal of time, we did not get to a point where we could integrate a system component to respond to an emotive user. An effective component could be just a supportive response to indicate to the user that the system understands he or she is having difficulty, or the system could change its mode of operation to a more directed dialogue model, or offer an alternative method to entering the data (e.g. key pad or spelled mode).

The preliminary investigation into reliable features for natural emotion recognition is discouraging. Clearly, more work needs to be done. One issue may be the difficulty in reliably extracting, abstracting, and normalizing pitch information, particularly for telephone quality speech. In a realistic environment, one should be able to compute baseline pitch and energy data from the first few utterances the user utters in a dialogue, and normalize further utterances based on these computations. However, further complicating factors are the users' inclination to perhaps conceal their emotions, or even to be unemotional in the face of difficulties due to a detachment with respect to a simulated task. It may even be worthwhile, once systems are able to respond appropriately to negative emotion, to train the users to *act* emotional when they are having difficulties with communication.

For the acted data, consistent with the observation that maximum pitch and average energy were intuitively good discriminators of emotion, these features were also statistically shown to do well. Further work needs to be done to build confidence in the emotion scoring technique and the classifier. One way to accomplish this is to investigate additional features that may be correlated with emotion. It would also be beneficial to perform additional classification experiments on each feature independently and compare human listener studies with recognition results.

# Appendix A

# Corpus Preparation

The following example output lists the properties of the Lockheed corpus. This is an example for the speech waveform labeled as `jfranke_002_root_djm_030`.

```
(lockheed_lab.corpus)


(datum jfranke_002_root_djm_030: 12 props)
  (tag: jfranke_002_root_djm_030)
  (waveform_file: .../lab_data/jfranke/002/root_djm_030.wav)
  (speaker: jfranke)
  (gender: male)
  (call_id: jfranke_002)
  (length: 10800)
  (original_orthography: roger out)
  (orthography: roger out)
  (artifact: no)
  (repeat: no)
  (emotional_state: neutral)
  (contains_oov: no)
```

A description of the 12 properties of the corpus are as follows:

- tag: the identification label for an utterance

- waveform_file: location of waveform
- speaker: individual user that was recorded speaking utterance
- gender: adult male or female speaker
- call_id: similar to session but includes speaker as well
- length: utterance length in samples
- original_orthography: transcription of the utterance before artifacts and oov words were removed
- artifact: special tag for designating whether the utterances has certain acoustic events such as laughter or coughing
- repeat: indicates whether utterance is a repeat
- emotional_state: indicates whether utterance is emotional (frustrated) or non-emotional
- contains_oov: indicates if the utterance contains out-of-vocabulary words

# Bibliography


[1] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog,", in *Proceedings of the ICSLP*, 2002, vol. 3, pp. 2037–2040.

[2] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," in *Journal of Personality and Social Psychology* vol. 70, 1996, pp.614–636.

[3] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "Desparately seeking emotions or: actors, wizards, and human beings," in *Proc. of the ISCA-Workshop on Speech and Emotion*, Belfast, 2000.

[4] A. Batliner, R. Huber, H. Niemann, E. Nöth, J. Spilker, and K. Fischer, "The recognition of emotion," in *Verbmobil: foundations of speech-to-speech translations*, 2000, pp. 106–121.

[5] K. H. Blocher, "Affective social quest: teaching emotion recognition with interactive media and wireless expressive toys," MIT Masters Thesis, Cambridge, MA, June 1999.

[6] L. Bosch, "Emotions: what is possible in the ASR framework," in *Proc. of the ISCA-Workshop on Speech and Emotion*, 2000.

[7] N. Campbell, "Databases of Emotional Speech," in *Proc. of the ISCA-Workshop on Speech and Emotion*, 2000.

[8] R. T. Cauldwell, "Where did the anger go? The role of context in interpreting emotion in speech," in *Proc. of the ISCA-Workshop on Speech and Emotion*, 2000.

[9] R. Cornelius, "Theoretical approaches to emotion," in *Proc. of the ISCA-Workshop on Speech and Emotion*, 2000.

[10] R. Cowie and E. Douglas-Cowie, "Automatic statistical analysis of the signal and prosodic signs of emotion in speech," in *Proc. of the ICSLP*, 1996.

[11] R. Cowie, "Describing the emotional states expressed in speech," in *Proc. of the ISCA-Workshop on Speech and Emotion*, 2000, pp. 11-18.

[12] N. Dahlbäck, A. Jönnson, L. Ahrenberg, "Wizard of Oz studies - why and how," in *Knowledge Based Systems*, vol. June 4, 1994, pp. 258–266.

[13] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proc. of the ICSLP*, 1996.

[14] K. Fischer, "Annotating Emotional Language Data," in *Verbmobil Report 236*, Dec. 1999.

[15] T. J. Hazen, S. Seneff, and J. Polifroni, "Recognition confidence scoring and its use in speech understanding systems," in *Computer Speech and Language*, vol. 16, 2000, pp. 49–67.

[16] W. Hess, "Pitch determination of speech signals," Berlin, Germany: Springer, Verlag, 1983.

[17] L. Hetherington and M. McCandless, "SAPPHIRE: an extensible speech analysis and recognition too based on Tcl/Tk," in *Proc. of the ICSLP*, Philadelphia, PA, 1996.

[18] H. Holzapfel, C. Fuegen, M. Denecke, and A. Waibel. "Integrating emotional cues into a framework for dialogue management." in Proc. of the 2002 Intl. Conference on Multimodal Interfaces, October 2002.

[19] A. Iida and M. Yasumura, "Designing and testing a corpus of emotional speech," in *ASJ Spring Meeting*, 1998.

[20] M. Kienast and W. F. Sendlmeier, "Acoustical analysis of spectral and temporal changes in emotional speech," in *Proc. of the ISCA-Workshop on Speech and Emotion*, 2000.

[21] D. Kirsch, "The affective tigger: a study on the construction of an emotionally reactive toy," MIT Masters Thesis, Cambridge, MA, June 1999.

[22] D. Litman, J. Hirschberg, and M. Swerts, "Predicting automatic speech recognition performance using prosodic cues," in *Paper presented at NAACL*, Seattle, May 2000.

[23] C. J. Mitchell, C. Menezes, J. C. Williams, B. Pardo, D. Erickson, and O. Fujimura, "Changes in syllable and boundary strengths due to irritation," in *Proc. of the ISCA-Workshop on Speech and Emotion*, 2000.

[24] T. Moriyama and S. Ozawa, "Emotion recognition and synthesis system on speech," in *Proc. of the IEEE ICMCS*, June 1999.

[25] S. J. L. Mozziconacci and D. J. Hermes, "Role of intonation patterns in conveying emotion in speech," in *Proc. of Intl. Congress of Phonetic Sciences*, 1999, pp. 2001–2004.

[26] S. J. L. Mozziconacci and D. J. Hermes, "Expression of emotion and attitude through temporal speech variations," in *Proc. of the ICSLP*, vol. 2, 2000, pp. 373–378.

[27] S. J. L. Mozziconacci, "Study of intonation patterns in speech expressing emotion or attitude: production and perception," *IPO Annual Progress Report*, IPO, Eindhoven, 1998.

[28] S. J. L. Mozziconacci and D. J. Hermes, "Emotion and attitude conveyed in speech by means of prosody," *2nd Workshop on Attitude, Personality and Emotions in User-Adapted Interaction*, Sonthofen, Germany, 2001.

[29] S. L. Oviatt, G. Levow, M. MacEachern, and K. Kuhn, "Modeling hyperarticulate speech during human-computer error resolution," *Proc. of the ICSLP*, vol. 2, 1996, pp. 801–804.

[30] S. L. Oviatt, M. MacEachern, and G. Levow, "Predicting hyperarticulate speech during human-computer error resolution," *Speech Commun*, vol. 24, 1998, pp. 1–23.

[31] A. Paeschke and W. F. Sendlmeier, "Prosodic characteristics of emotional speech: measurements of fundamental frequency movements," in *Proc. of the ISCA-Workshop on Speech and Emotion*, 2000.

[32] C. Pereira and C. Watson, "Some acoustic characteristics of emotion," in *Proc. of the ICSLP*, Sydney, Australia, 1998.

[33] V. Petrushin, "Emotion in speech: recognition and application to call centers,", in *Proceedings of the Conference on Artificial Neural Networks in Engineering*, 1999.

[34] V. Petrushin, "Emotion recognition agents in real world,", in *Papers from the AAAI Fall Symposium on Socially Intelligent Agents: Human in the Loop*, 2000.

[35] R. Picard and J. Klein, "Computers that recognize and respond to user emotion: theoretical and practical implications," MIT Media Lab Tech Report 538.

[36] R. Picard, "Affective computing," Cambridge, MA: MIT Press, 1997.

[37] R. Picard, "Building HAL: computers that sense, recognize, and respond to human emotion," in *Human Vision and Electronic Imaging VI*, January 2001.

[38] H. Pirker and G. Loderer, "'I said "two ti-ckets': how to talk to a deaf wizard," in *Proc. of the ECSA Workshop on Dialogue and Prosody*, Sept. 1999, pp. 181–186.

[39] J. Polifroni and S. Seneff, "Galaxy-II: as an architecture for spoken dialogue evaluation," in *Proc. of the 2nd Intl. Conference on LREC*, Athens, Greece, 2000.

[40] T. S. Polzin and A. Waibel, "Emotion-sensitive human-computer interfaces," in *Proc. of the ISCA-Workshop on Speech and Emotion*, Belfast 2000.

[41] Y. Qi, C. Reynolds, and R. Picard, "The Bayes point machine for computer-user frustration detection via pressure mouse," in *PUI*, 2001.

[42] P. Roach, "Techniques for the phonetic description of emotional speech," in *Proc. of the ISCA-Workshop on Speech and Emotion*, 2000.

[43] K. R. Scherer, R. Banse, H. G. Wallbott, and T. Goldbeck, "Vocal cues in Emotion Encoding and Decoding," in *Motivation and Emotion*, vol. 15, 1991, pp. 123–148.

[44] M. Schröder, V. Aubergé, and M. Cathiard, "Can we hear smile?"

[45] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, and V. Zue, "Galaxy-II: a reference architecture for conversational system development," in *Proc. of the ICSLP*, Sydney, Australia, 1998.

[46] B. Shneiderman, "The limits of speech recognition," in *Communications of the ACM*, Sept. 2000, vol. 43, no. 9, pp. 63–65.

[47] E. Shriberg, E. Wade, and P. Price, "Human-machine problem solving using spoken language systems (SLS): Factors affecting performance and user satisfaction," *Proc. of the DARPA Speech and Natural Language Workshop*, 1992, pp. 49–54.

[48] M. Walker, J. Aberdeen, J. Boland, E. Bratt, J. Garafolo, L. Hirschman, A. Le, S. Lee, S. Narayanan, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. Rudnicky, G. Sanders, S. Seneff, D. Stallard, and S. Whittaker, "DARPA Communicator dialog travel planning systems: The June 2000 data collection," in *Proc. EUROSPEECH*, Aalborg, Denmark, Sept. 2001, pp. 1371–1374.

[49] C. Wang and S. Seneff, "Robust pitch tracking for prosodic modeling in telephone speech," in *Paper presented at the IEEE ICASSP*, Istanbul, Turkey, June 2000.

[50] S. P. Whiteside, "Simulated emotions: an acoustic study of voice and perturbation measures," in *Proc. of the ICSLP*, 1998, pp. 699–703.

[51] V. Zue, J. Glass, D. Goodine, M. Phillips and S. Seneff, "The SUMMIT speech recognition system: phonological modelling and lexical access," in *Proc. of ICASSP*, 1990, pp. 49–52.

[52] V. Zue and J. Glass, "Conversational interfaces: advances and challenges," in *Proc. of the IEEE*, vol. 88, no. 8, Aug. 2000, pp. 1166–1180.