CLASSIFICATION BASED ANALYSIS OF SPEECH UNDER PHYSICAL TASK STRESS

by

Keith William Godin

APPROVED BY SUPERVISORY COMMITTEE:

Dr. John H.L. Hansen, Chair

Dr. Carlos Busso

Dr. William F. Katz

To my parents.

CLASSIFICATION BASED ANALYSIS OF SPEECH UNDER PHYSICAL TASK STRESS

by

KEITH WILLIAM GODIN, B.Sc.

THESIS

Presented to the Faculty of
The University of Texas at Dallas
in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING

THE UNIVERSITY OF TEXAS AT DALLAS
December, 2009

ACKNOWLEDGEMENTS

The completion of this thesis is due also to the work of several others, whom I would like acknowledge. I wish to extend my thanks to my advisor, Dr. John H.L. Hansen, for his advice, attention, and support for the past two and a half years. I would like to thank Dr. Hynek Bořil for practical advice and many supportive discussions on the research methods and results in this thesis. I would also like to thank the Electrical Engineering Department at the University of Texas at Dallas for the financial and logistical support provided for my research and studies.

Finally, to Xing, for her kind words, and warm thoughts.

November, 2009

CLASSIFICATION BASED ANALYSIS OF SPEECH UNDER

PHYSICAL TASK STRESS

Publication No._____

Keith William Godin, M.S.
The University of Texas at Dallas, 2009

Supervising Professor: John H.L. Hansen

The variability in speech production brought on by physical stress causes significant reduction in speech system performance for speech and speaker recognition. This thesis focuses on analysis of speech under physical task stress. An analysis of fundamental frequency, fundamental frequency variance, utterance duration, and the percent of frames voiced in an utterance is performed. F-ratio analyses of the spectrum and cepstrum provide motivation for further research effort. Finally, the aim of the main research effort is to establish which phone classes are the most affected by physical task stress. A new analysis method is proposed which involves interpreting the results of an classification system that is based on Mel-Frequency Cepstral Coefficients using Gaussian Mixture Models. It is shown that nasals and laterals are most affected by physical task stress. These results will find application in speech recognition research and provide new insight in the study of physical task stress speech.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Analysis of speech under physical task stress can result in new understanding of the way that variation affects speech systems. This thesis carries out an analysis of speech under physical task stress and discusses what may be concluded about the behavior of the speech production system and how that affects the design of speech systems. The study of speech variability in general is important for the long-term success of speech systems, as they are deployed in increasingly varying environments (Hansen and Womack, 1996). The focus of this thesis is short-term intra-speaker variability; i.e. the variation inherent in one speaker's speech in the context of one utterance or one or more conversations. Long-term variation includes variation due to aging, or the naturally slow changes in a speaker's accent or dialect, while short-term variation results from changes in emotion, stress, or environmental factors (Benzeghiba et al., 2007). This thesis focuses on the short-term variations in speech production that result from physical task stress, with the intent to uncover insights that may be applicable to the design of speech systems, including speaker identification and verification systems and automatic speech recognition systems, that are robust to such speech production variations.

This thesis contributes a new analysis method, based on the examination of the output of an automatic classification system, to be applied to understanding speech variation. In this thesis traditional analysis methods as well as the new analysis method are applied to physical task stress in an effort to understand the nature of the speech signal variation and speech production variation that occurs as a result of physical task stress on the part of the speaker. The Center for Robust Speech Systems (CRSS) at the University of Texas at Dallas is the only lab actively looking at speech under physical stress at the time of this writing, though research on speech under physical task stress has been published in

the past by researchers at the University of South Dakota. Prior work on speech under physical task stress has not offered theories, methods, experimental results, or definitions of the problems and main ideas directly concerning speech under physical task stress and the design of related systems. This thesis therefore takes a broad view at research in various other styles of speech and discusses what kinds of methods and ideas might be applied to speech under physical task stress, and to contribute relevant theory, method and experimental results.

In the remainder of this thesis is first discussed past work on other types of speech variation in Chapter 2. In Chapter 3 is discussed the speech corpus used for experiments in this thesis. In Chapter 4, data analysis experiments are performed, their results discussed, and then a motivation is presented for the work in Chapter 5. Chapter 5 presents an experiment involving analysis of scores from a stress classification system that is designed to determine which phone classes are most affected by physical task stress. A formal listener test is presented in Chapter 6. Finally, Chapter 7 discusses conclusions that may be made based on the results presented in this thesis, and future research work that may be performed to extend the knowledge presented.

CHAPTER 2

RESEARCH IN SPEECH STYLES

The purpose of this chapter is to synthesize related background research into a basis for and justification for the work that will be performed later in this thesis. Physical task stress has not been the focus of extensive research in the past; this chapter therefore draws from a wider array of research than that specifically considering speech under physical task stress. This chapter begins by introducing the concept of *speech styles*, which will be used as an organizing principle around which research areas similar to speech under physical task stress may be identified. Despite the fact that various types of speech variation are researched in relative isolation by a variety of groups with different motivations, a main goal of this chapter is to present a comprehensive view of the variety of research that has taken place on various speech styles, in the hopes that from a comprehensive view will emerge guidance for research on speech under physical task stress.

Specifically, the following are questions pertinent to a research endeavor in speech under physical task stress that may be answered by a careful review of work on related types of speech:

- What is the best method to collect appropriate speech data?
- How are various types of speech variation defined and distinguished from one another?
- What are the important opening questions for research into a new type of speech variation?

2.1 Speech styles

This section introduces the idea of speech styles as a device with which research into various types of speech variation may be compared and contrasted with research on physical task stress speech. The purpose of this section is not to fully develop the concept of speech styles as a model of speech production or perception. Rather, the purpose of this section is to discuss the concept of speech styles in enough depth that it is reasonable to apply it as a guide in understanding how various research is related. For example, it is clear that speech under physical task stress corresponds with an intuitive notion of a type of speech. But what are speech types? How are types of speech different from each other? It is thus proposed to define a speech style as "a consistent method of producing speech adopted by a speaker for a limited duration that affects the phonetic content of the language". Because they affect the phonetic content of the language, speech styles may thus affect the meaning of the speech produced, and might convey some extra-linguistic information to the listener. Any short term variation in speech may constitute a speech style.

It is important to note that speech styles are an articulatory-phonetic phenomenon, a phenomenon that results in changes to the acoustic-phonetics of the language and may result in auditory-phonetic changes, and may or may not be perceived. This definition of speech styles makes explicit the previously implicit notion that there are distinct ways of speaking that speakers shift into when they change state, and separates the notions of speaker state and the resulting articulatory phenomena. Speech styles are not a phenomenon of the organization of sound units (phonology), or the syntax or morphology of the language. They are also not directly dependent on a speaker's state, including emotions. Rather, various speech styles might result from a single speaker state. Whisper, for example, is an example of a speech style. Whisper is highly definable in terms of properties of the speech production system - it lacks a fundamental frequency. Similarly, other voicing registers, or hoarseness or pressed voicing, would be considered speech styles under this definition. Speech styles are thus collections of properties of the articulatory

system and not necessarily related to speaker intent.

Perhaps, therefore, some distinction should be made between speech styles (such as whisper) and motivations to change speech style (such as emotions). Angry is not a style of speech, under this definition. Emotions are instead a form of speaker state. Rather, one or more speech styles may result from anger. The particular speech style that may result will depend on the speaker's habits, mood, situation, and other factors. Whisper, as a speech style, could result from a wide range of situations and decisions on the part of the speaker. The purpose of the definition of speech styles is to move the center of focus in understanding research in speech variation away from motivation or context surrounding a speaker, and towards a particular speech production process.

The primary motivation of the speech styles concept is to assist the speech researcher determining more specific definitions of what constitutes the speech under consideration. Classic papers considering speech under stress or emotion assume that a particular type of constant speech production process is associated with, for example, anger, or at least a certain subtype of anger. But what constitutes hot anger? Is acted speech similar enough to base conclusions on? The speech styles model attempts to resolve such ambiguity by shifting the problem of definition to one that may offer more possibilities for objective measurement than is possible with speaker state. The speech styles concept defines as styles of speech a consistent set of articulatory properties. For example, one style of speech that might result from anger on the part of the speaker might involve increased pitch, a certain hoarseness, and faster speaking rates. Another style of anger might have more in common with loud speech, involving formant variations due to a more open vocal tract, pitch, and intensity changes.

The speech styles concept represents a departure from the past model for short-term speech variation, the ordered-stressors model presented in Murray et al. (1996). The ordered stressors model focuses on the sources of speech variation, termed *stressors*, rather than on their results, and presents a way to taxonomize the sources of speech variation. It too has not seen application in the literature on short-term speech variation. The or-

dered stressors model is discussed further in Section 2.2.1. Whether or not research into speech variation and how it affects speech systems has explicitly employed the ordered stressors model, it has in the past employed the perspective of the source of the speech variations (such as anger) and examined the speech that results. Some recent research has begun to consider a perspective more related to that of the speech styles concept. Hansen and Varadarajan (2009), for example, considered whether there may be different types of Lombard speech, which might result from different types of noise the speaker may be exposed to.

However, despite the fact that it is at odds with the way research has been conducted in this area in the past, the concept of speech styles is useful as a way view research into various sources of speech variation because it turns the focus of the research away from the source of variation, which is quite different from physical task stress, towards the variation that results from that source, which may or may not be in common with the speech that results from physical task stress and may thus provide some insight into research in speech under physical task stress.

2.2 Past work on speech variation

Past work on various types of short-term speech variation has generally focused on parameter analysis, the design of detection systems, and the design of improvements to achieve robustness of speech and speaker recognition systems. Early papers (Williams and Stevens, 1972) often emphasized visual observation of spectrograms and qualitative reasoning. More recent papers emphasize automatic measurements and employ statistical tests to draw conclusions, such as Maniwa et al. (2009). However, the measurement focus has not changed significantly. Parameters including fundamental frequency, formant locations, and phone durations are generally the focus of study. There has been some discussion in the literature seeking to consider other aspects of the speech signal, in an effort to gain deeper insight into the nature of speech variation and its relationship to speech system performance (see e.g. Cheang and Pell (2008)), but these parameters

remain the primary objects of study in the context of speech systems. Perhaps the reason is that few other parameters form such direct and fundamental connections between the acoustic speech signal and the behavior of the speech production and perception systems.

The term "speech under stress" has been found in the literature of speech systems for more than two decades. The term is difficult to define, and working definitions have evolved over time. As speech systems reached an initial level of success, research engineers began branching out in the mid-1980's into studying other types of speech than the laboratory, low-noise, non-emotional speech that had been the focus of system development up until that time. Any type of speech that deviated in its speech production from this norm was referred to as "speech under stress" by researchers at that time (see Chen (1988) and Hansen (1988) for examples of this usage). Speech referred to as "speech under stress" included speech differing in rate (fast, slow) and effort (loud, soft), differing due to task demands (cognitive demands, physical stress, fatigue, and others), and differing due to chemical inducement. Later, efforts were made to more formally define speech under stress and to create theoretical models of stress types for use in research (Murray et al., 1996). As the area diversified to include more stress types, researchers began focusing their work on a single type of speech variability or small set of types, rather than broadly considering speech variation. Work has been published recently on Lombard effect (Boril and Hansen (2009), Lu and Cooke (2009)), emotion (Burkhardt et al. (2009), Ijima et al. (2009)), cognitive load (Lindstrom et al. (2008)), and fatigue (Greeley et al., 2006), among other areas. Especially in research on emotional speech, we find that researchers from several domains, such as psychology, engineering, and phonetics now collaborate on new results. Work also continues intermittently on refining definitions of speech under stress and emotion (e.g. Scherer (2003) and Godin (2009)). As the term "speech under stress" has evolved, it has become more specific; now speech production variation due to emotion, rate or style changes, and Lombard effect are rarely referred to as speech under stress, the term being reserved for speech under cognitive task stress and physical task stress.

The remainder of this section discusses specific insights from past work on short-term speech variation that are relevant to the study of speech under physical task stress. Past work has discussed theoretical models that frame research and experimental design, a large body of past work has been published on the parameter analysis of various types of short-term speech variation, and past work has discussed several automatic classification methods and their applications.

2.2.1 Theoretical models: taxonomies of variation

There has been research into a few theoretical models that may be applied to research in short-term speech variation. Models are important because they assist in answering the fundamental questions of organization. They explain, for example, why one should collect data in a certain way, what kinds of research questions one might ask. Models drive the fundamental breakthroughs, because they form the underpinnings of experimental design. Descriptions of two models applied to the study of speech variation follow. One is the Brunswickian lens model, developed by psychologists for the study of perception, and applied to the study of emotions in speech by Scherer (Scherer, 2003). Another model is an ordered-stressor model proposed in (Murray et al., 1996), in which the speaker is modeled as having four levels at which stressors can affect his/her speech.

Emotions in speech are a well studied source of speech styles. Intuitively we know that when people are angry, or sad, or currently experiencing a wide variety of other emotions, their speech is affected. One model that has been applied to the study of emotions in speech is the Brunswickian lens model, as applied by Scherer (Scherer, 2003). The lens model as applied by Scherer is designed to study the vocal communication of emotion. It breaks the vocal communication of emotion into several steps: speakers have an emotional state, that state has effects on the speech production system, the produced speech may change due to transmission and the hearing process, and finally the listener attributes the sounds they hear to emotions in the speaker. The utility in the Brunswickian lens model is that it provides for the systematic study of the various stages of vocal

emotion communication. The lens model is not focused on the determination of the types of variation but on the study of the results of vocal emotion.

Another important model of the speech types considered in this chapter was first described in detail in (Murray et al., 1996) and was left unnamed by its original authors. It is referred to in this thesis as the ordered-stressors model.

Neither the Brunswickian lens model nor the ordered-stressors model have been extensively applied to research in their respective domains. Most publications in the area of speech variation assume an intuitive definition of the speech variation to be studied. This can lead to problems where research results from different studies are not comparable because in practice they assume fundamentally different definitions of a particular type of speech variation, while on the surface they may assume what appear to be equivalent intuitive definitions. An important conclusion to be drawn that is relevant to physical task stress speech research is that research should be guided by theoretical models that are continually refined as new results are achieved.

2.2.2 Data collection

Data collection is an important problem in the study of speech variation, because it directly affects the validity of the results of any experiments. Fortunately, data collection for physical stress avoids much of the ambiguity associated with the collection of, say, emotion data, where researchers disagree on whether acted speech is ecologically valid, and whether it is possible to design experiments to elicit real emotions in a laboratory setting. The collection of physical task stress data does encounter some specific questions. How should the level of physical task stress be defined? Should it be an absolute heart rate? Should it be a measure of heart rate as a percentage greater than resting heart rate? Or should it be a specific exertion level? Or calibrated to body weight, or to Body Mass Index (BMI)? The most relevant conclusion from past research on short-term speech variation is that it is important to clearly define what is the ground-truth of what constitutes the source of the speech variation to be studied. However, while it is important to de-

fine and fix the definition of ground-truth within the context of one study, it is important to examine, in other studies and throughout the course of research, what results from other definitions of ground-truth. In this thesis, the available corpus has taken the level of physical task stress to be the based on how quickly the experimental subject operates an exercise machine, irrespective of the fitness level of the subject. It should thus be noted that the results from this studies are based in this particular context. Future studies on the nature of physical task stress speech can and should also consider the physical task stress speech that results from other definitions of the level of physical task stress, making explicit the definitions used, and comparing the results to those of this study.

2.2.3 Acoustic data analysis

Many past works have investigated various sources of speech styles. A complete literature review may be found in Benzeghiba et al. (2007). The following is a list of representative examples:

- Angry (Hansen, 1988), (Scherer, 2000), (Cummings and Clements, 1995), (Williams and Stevens, 1972)
- Apache helicopter
- Clear speech (Hansen, 1988), (Cummings and Clements, 1995), (Maniwa et al., 2009)
- Cockpit conditions
- Fast (Hansen, 1988), (Cummings and Clements, 1995)
- Fatigue (Whitmore and Fisher, 1996)
- Fear (Williams and Stevens, 1972)
- Lombard (Hansen, 1988), (Bond et al., 1989), (Cummings and Clements, 1995), (Hansen and Varadarajan, 2009)
- Loud (Hansen, 1988), (Holmberg et al., 1988), (Cummings and Clements, 1995)

- Physical task stress (Godin and Hansen, 2008), (Patil and Hansen, 2008)
- Question (Hansen, 1988), (Cummings and Clements, 1995)
- Rollercoaster
- Sarcasm (Cheang and Pell, 2008)
- Slow (Hansen, 1988), (Cummings and Clements, 1995)
- Soft (Hansen, 1988), (Holmberg et al., 1988), (Cummings and Clements, 1995)
- Sorrow (Williams and Stevens, 1972)
- Whisper (Fan and Hansen, 2009)
- Workload task stress (Hecker et al., 1968), (Hansen, 1988), (Cummings and Clements, 1995), (Lindstrom et al., 2008)

It is clear from this list the variety of short-term speech variability that has been studied in the past. However, in most studies of short-term speech variation, acoustic data analysis generally takes the form of comparing the mean values of parameters measured from the speech signal in neutral and measured under stress or speech variation. Often the following parameters are investigated include fundamental frequency, formant locations, formant bandwidths, speech rate, and intensity. Each type studied is sought to provide a description of the acoustic correlates of the stress type. Data analysis studies though are often more than simply descriptive of the variation found in a set of parameters. For example, it has been shown in (Hansen and Varadarajan, 2009) that there are different types of Lombard speech, depending on the type of noise in the speaking environment. The best work on speech styles seeks to be more than simply descriptive: it strives for more an interpretation of the results in terms of how the speech production system is behaving or how speech systems may specifically be affected. The best work strives to characterize the speech style as a whole: How does the speech style fit into the phonetic sequence? Into everyday experience? How much variation across speakers is seen from

the same source of variation? Might different speech styles be driven by the same emotion? Do different emotions result in the same speech style? What are the implications for the semantic meaning of the utterance when a speech style is adopted?

2.2.4 Automatic classification

Besides data analysis, a number of studies in the literature on speech variability have considered classification systems, such as Cairns and Hansen (1994), Womack and Hansen (1996), Lee and Narayanan (2005), Patil and Hansen (2008), and Sethu et al. (2009), and others. Classification systems have various applications, including as front-ends to robust systems or as part of a segmentation or summarization system. Classification systems have been of interest to build and analyze because in so doing a researcher can examine the effects of speech variability on feature extraction and modeling technologies outside the complicating context of a speech or speaker recognition system. In this thesis it will be shown that to build and analyze a classification system can have an additional purpose. A classification system for physical task stress speech will be analyzed in this thesis to draw conclusions about the effects of physical task stress on the speech production system.

2.3 Past work on physical task stress

Little work has been performed specifically on speech under physical task stress. Publications so far include Entwistle (2003), Entwistle (2005), Godin and Hansen (2008), Godin (2009), and Patil and Hansen (2008). From Entwistle's work it can be concluded that automatic speech recognition (ASR) systems are negatively impacted by physical task stress (referred to as "human exertion" in her work). Patil and Hansen (2008) described classification experiments performed on the UT-Scope corpus. Two features were employed, a stress detection feature (TEO-CB-AutoEnv), and a classic speech feature (MFCCs), and speech data from two sensors was examined, the standard close-talking acoustic microphone, and the physical microphone (P-MIC). The statistical classifiers were Gaussian Mixture Models (GMMs). It was found that the system classifying data from the P-MIC

performed better than the system classifying data from the acoustic microphone. There are various possible reasons for the improvement in performance seen. The classification performance might be improved in the P-MIC due to heartbeats heard in the P-MIC of increased amplitude or frequency of occurrence, or the P-MIC may capture more of the (assumably modified) respiration noise of speech, or the P-MIC may capture more clearly glottal sounds that differentiate physical task stress speech from neutral speech. The performance improvement may serve as evidence that many of the articulatory effects of physical task stress lay with the glottal structures, a hypothesis in line with other observations including the reduction in voicing and the change in fundamental frequency seen in Godin and Hansen (2008).

2.4 Work on physical task stress

From the literature discussed in this chapter a guide may be formulated to further research into physical task stress speech. First, it is clear that the ground truth of what constitutes physical task stress and the level of physical task stress speech must be defined and investigated. Second, the literature suggests that a three part process is most effective for beginning work on a new type of short-term speech variation, when the purpose of that work is to advance the design of speech systems. That process involves defining a theoretical model and ground-truth, performing and interpreting data analysis experiments, and investigating the behavior of speech systems through the behavior of classification systems based on relevant speech technologies. Finally, the literature has suggested that data analysis should strive to be more than descriptive of specific parameters of the speech signal, especially because those parameters are not directly employed in relevant, contemporary speech systems. Instead, interpretation of the results of data analysis in terms of the overall behavior of the speech signal is important to ensure relevance and applicability of the analysis results. As these three parts of the investigation together lay the groundwork for future studies that develop robustness improvements to speech systems, this thesis makes these three parts its focus.

CHAPTER 3

CORPUS

This thesis employs the UT-Scope corpus (Ikeno et al., 2007) for experiments on speech under physical task stress. The corpus includes neutral speech, speech under physical task stress, speech under cognitive task stress, and Lombard effect speech. This thesis contains some of the first results of experiments performed on the physical task stress segment of the UT-Scope corpus. Other recent work on the physical task stress portion of the UT-Scope corpus includes Godin and Hansen (2008), which this thesis draws from, and Patil and Hansen (2008), Patil (2009), and Sangwan (2009), each of which applied the UT-Scope corpus to speech systems to directly examine the effect of physical task stress. This chapter discusses the nature of the physical task stress portion of UT-Scope and addresses its suitability for research into physical task stress speech.

3.1 Overview

The UT-Scope corpus was collected at the University of Texas at Dallas and includes speech from 77 speakers, 51 of which are native speakers of American English, with 42 female speakers of Am-English and 9 male speakers. For each stress type was recorded a 35 sentence prompted speech segment (prompted through headphones) and a 3 minute spontaneous speech segment involving a conversation between the experimenter and the subject. The prompted segments of the recordings of the native speakers have full sentence and word level segmentations available, with phone-level segmentations available for 38 of the 42 female native speakers. The phone-level segmentations were generated using forced-alignment, with segmentations for 10 of the speakers having been hand corrected by a researcher.

The physical task stress for the speech recordings was induced using an elliptical stair

Figure 3.1. Recording setup for the physical task stress segment of the UT-Scope corpus.

stepper. The speakers were directed to maintain an approximately 10 mph speed on the stairstepper. The recording setup for the physical task stress segment of UT-Scope is shown in Fig. 3.1.

The experiments performed in Chapter 4 were performed on the prompted segments of the physical task stress and neutral recordings of the 51 native Am-English speakers. The experiments performed in Chapter 5 were performed on the prompted neutral and physical task stress segments of the 38 female native speakers for which phone segmentations are available. Relevant aspects of the portion of UT-Scope used in this thesis are described in Table 3.1.

3.2 Heart rate analysis

As a corpus designed to include speech under physical task stress, it is important to gauge the actual exertion level of the speakers as recorded. Heart rate (including for the neutral

Table 3.1. Aspects of the subset of UT-Scope used in this thesis.

Parameter	Male spkrs	Female spkrs
# of speakers	9	42
Average age (yrs)	22.3	23.6
Age range	19-33	18-45
Sentences/task	35	
Tasks	Neutral, Physical exertion	
Native language	American English	
Microphone	Close-talking	
Speech style	Prompted	
Av. exertion level	43%	
Sampling rate	16kHz	

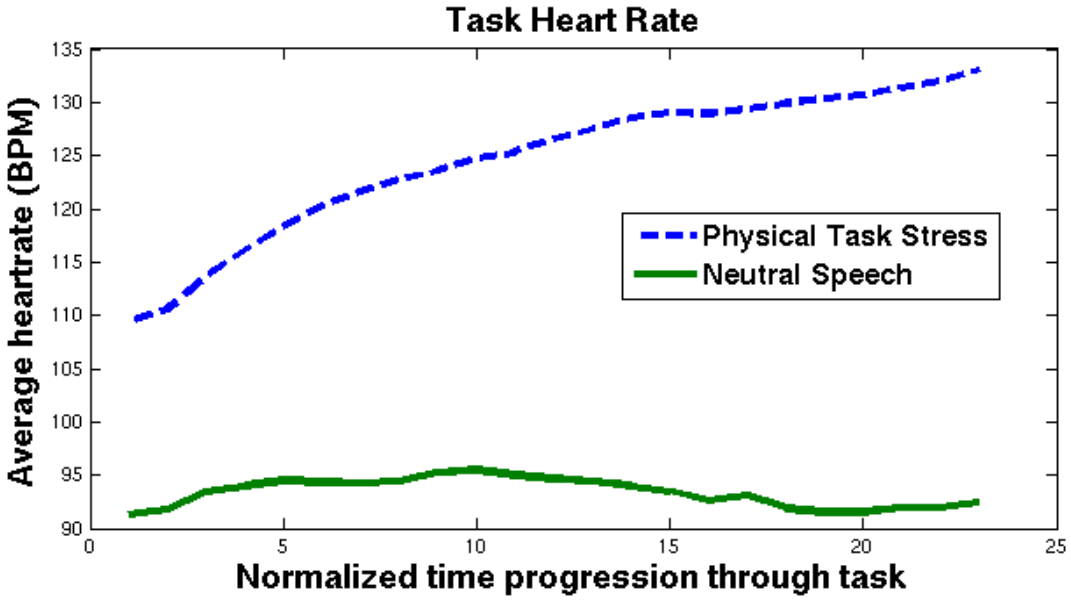

Figure 3.2. Average heart rate through time for the neutral and physical task stress recordings.

condition, and sampled each 15s) and age data are available for all of the speakers. The average heart rate across all of the speakers for both the neutral and physical task stress recordings is shown in Fig. 3.2. It may be seen from the figure that the heart rate for the physical task stress recordings is higher than the heart rate for the neutral recordings.

The available heart rate and age data may be used to estimate exertion level. Though it will tend to overestimate the resting heart rate, a rough estimate of the resting heart rate of each speaker is obtained by averaging the speaker's heart rate during the neutral segment. Then, one of many popular, and sometimes controversial, and generally similar

formulae can be used to estimate the exertion level. One estimate is the Karvonen formula

$$HR = (MHR - RHR)l + RHR \tag{3.1}$$

where HR is the current heart rate, RHR is the resting heart rate as estimated, l is the exertion level (ranging from 0 to 1) and MHR is the person's maximum heart rate, estimated according to (Tanaka et al., 2001):

$$MHR = 208.9 - 0.7A \tag{3.2}$$

where A is the age of the person. Applying this formula to the 51 native speakers of the UT-Scope database, having had to disregard 7 speakers due to missing heart rate data (6) or error in recording data (1), an average exertion level of 0.43, or 43%, is found, with standard deviation of 12%. We consider this an appropriate level of exertion. A high level of exertion (perhaps in the range of 60-80%) would not produce exemplars similar to those likely to be seen in real life. Those exercising highly are not likely to be interested in speaking for extended periods of time. A low exertion level is thus more ecologically valid.

CHAPTER 4

PARAMETER ANALYSIS

This chapter discusses an analysis of physical task stress speech across several speech parameters, in an effort to understand the ways that physical task stress affects speech production and the acoustic speech signal. There are two parts to the following parameter analysis. The first part of the analysis examines four parameters of the speech signal on a per-speaker basis. For each speaker, statistical tests are used to determine whether the measured parameter undergoes a significant shift in mean from that measured in neutral speech to that measured under physical task stress. The second part of the analysis examines the spectrum and cepstrum to determine where, on average across all speakers, the most variance due to physical task stress may be found. An important result of the second part of the analysis is that it motivates the use of the cepstrum for further analysis purposes in Chapter 5.

4.1 Per-speaker parameters

Four parameters of the speech signal were measured and compared between neutral and physical task stress: fundamental frequency, standard deviation of fundamental frequency within an utterance, utterance duration, and the percent of frames voiced within an utterance. To measure fundamental frequency (F_0), the F_0 for each 10ms frame was computed with WaveSurfer (Sjolander and Beskow, 2000), using the ESPS algorithm with an analysis window of 75ms. The F_0 minimum was set to 120 Hz for females and 80 Hz for males, and the maximum to 400 Hz. A distribution was formed for each condition (physical and neutral) for each speaker, selecting only those F_0 values lying within a prompted utterance as indicated by the utterance segmentations available with UT-Scope. Two 1-sided t-tests, with a 99% confidence level, were used to compare the distribution means

of F_0 for each speaker. Results for F_0 and the other parameters are shown in Table 4.1. The results are discussed at the end of this section.

The standard deviation of F_0 within an utterance, $F_0\sigma$, was measured to determine whether F_0 varies more or less under physical task stress. The standard deviation of the F_0 within each utterance was computed, yielding 35 measurements of the distribution of σ per condition per speaker. Two 1-sided student-t tests were used to compare the means of these distributions for each speaker. The third parameter examined was the difference in duration of utterances between physical task stress and neutral speech as spoken by the same speaker. This was compared by applying two one-sided student-t tests to the set of 35 measurements of difference in duration to determine whether the difference in duration had a distribution with mean statistically greater or less than zero at the 99% confidence level.

Finally, the percentage of voiced frames in an utterance was measured as for F_0 above, by using WaveSurfer to identify which 10ms frames of the recordings were voiced and which were unvoiced. The percentage of voiced frames in each utterance was computed, yielding 35 measurements per speaker per condition. For each speaker, distributions for neutral and physical task stress speech were formed and the means of these distributions compared using two 1-sided student-t tests.

Table 4.1 summarizes the results. Approximately 88 % of the speakers showed a statistically significant decrease in the percentage of frames in an utterance considered voiced, and just 1.96 % showed a statistically significant increase. We conclude that a reduction in the amount of voiced speech is a primary indicator of physical task stress in speech. Considering the mean F_0, 60.8 % of the speakers had a statistically significant increase in their mean F_0 under physical task stress, 13.6 % of the speakers had a statistically significant decrease in their mean F_0 under physical task stress, and 25.5 % of the speakers had no statistically significant change. Thus, for most speakers, a change in mean F_0 is associated with physical task stress.

The tests performed on utterance F_0 σ showed that only 25.4 % of the speakers showed

Table 4.1. Summary table of parameter analysis results

Statistical test result on physical stress speech	% of speakers
F_0 greater in stress	60.8
F_0 lower	13.6
F_0 same	25.5
F_0 σ greater in stress	1.96
F_0 σ lower	23.5
F_0 σ same	74.5
Duration shorter in stress	43.1
Duration longer	31.4
Duration same	25.5
Greater % voiced in stress	1.96
Lower % voiced	88.2
Same % voiced	9.80

a statistically significant difference in the mean of utterance F_0 σ, with 23.5 % of the speakers having a lower utterance F_0 σ. For most speakers (74.5 %), no statistically significant change in utterance F_0 σ was found. We conclude that physical task stress has a negligible effect on the short term variability most speakers impart in their F_0. Finally, the results show that 25.5 % of the speakers have no statistically significant difference in the duration of their utterances, while 31.4 % of the speakers spoke longer under physical task and 43.1 % spoke shorter. We conclude that duration of prompted sentences is often affected by physical task stress, but that the manner of the effect is speaker dependent.

4.2 F-ratio analysis

In this section two analyses are performed, one of the influence of physical task stress on the variance in the speech signal spectrum, and another of the influence of the stress on the speech signal cepstrum. The measurements specifically are designed to measure the effects of physical task stress on each frequency or quefrency bin of a DFT/real cepstral analysis. This measurement is accomplished by forming the ratio of the variance of each frequency/quefrency bin between tasks to the average variance for that bin within the tasks:

$$F_j = \frac{\textit{inter-category variance of bin } j}{\textit{average intra-category variance of bin } j} \tag{4.1}$$

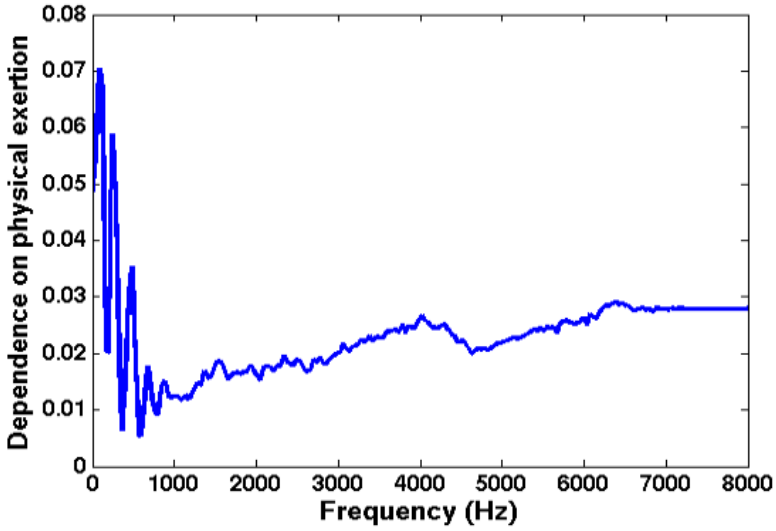

Figure 4.1. F-Ratio showing dependence of each frequency bin on physical task stress

for each frequency bin i. This is an application of Fisher's F-ratio, which in speech literature has been applied similarly to compare general intra-speaker variability to inter-speaker variability (Lu and Dang, 2008), and has been applied to the measurement of the discriminative capability of features for speaker identification (Campbell, 1997). To determine the overall F-ratio of each frequency bin, each $F_{i,j}$ is averaged across all speakers j resulting in an average F_i for each bin of the DFT analysis. This parameter is also measured for each bin of the real cepstrum, the real cepstrum being computed by taking the inverse DFT of the natural log of the magnitude of the DFT spectrum (Deller et al., 2000).

The results of the F-ratio measurements are shown in Figs. 4.1 and 4.2. From Fig. 4.1 it can be observed that the effects of physical task stress are on average not concentrated in specific areas of the spectrum, though greater effects may be observed on the lower frequencies than on the higher frequencies. This may be attributed to fundamental frequency changes, though that is not certain. However, it is clear from Fig. 4.2 that the effects of physical task stress concentrate in the middle of the cepstrum, which is associated with mid-size structural variations in the spectrum, and at the high end, associated with the smallest structural variations of the spectrum. These variations at the high-end of the cepstrum may be evidence of noisiness induced by breathing or breathiness at the

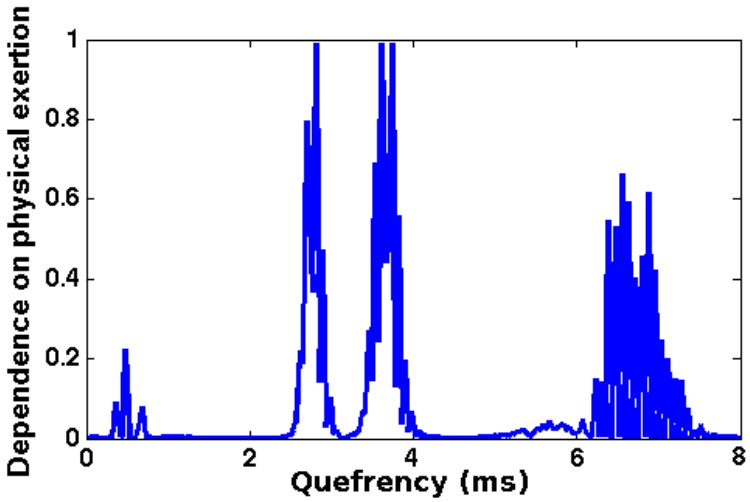

Figure 4.2. F-Ratio showing dependence of each cepstral bin on physical task stress

glottis, though that remains to be determined.

However, the most striking observation from Fig. 4.2 is the scale of the effects of physical task stress on individual areas of the cepstrum as compared to areas of the spectrum. It may be observed that the magnitude of the effects on the cepstrum can be as much as 10 times as high as that of the effects on the spectrum, and in some areas of the cepstrum the variance due to physical task stress is equivalent in magnitude to the variance due to all other factors combined (i.e. in some places the F-ratio approaches 1). Because the magnitude spectrum and the real cepstrum have, mathematically speaking, the same information, it may be concluded that the effects of physical task stress so clearly evident in the cepstrum are to be found not in specific frequency bins, but in the structure of the spectrum or in the time variation of the spectrum. Further investigation is warranted both into how physical task stress affects the structure of the spectrum, and into the source of the variation seen so clearly in the F-ratio of the cepstrum.

4.3 Discussion

Three observations may be made based on the measurements presented in Section 4.1. First, it is clear that the effects of physical task stress vary by speaker. Thus, there must be a variety of strategies that speakers may employ to cope with physical task stress, and the

design of any experiment to determine the effects of physical task stress on speech must explore or account for the speaker variability of the effects. Furthermore, any speech system designed to be robust to speech under physical task stress must also account for variations between speakers in how their speech production processes are affected by physical task stress. Second, it is clear that most speakers reduce the percentage of time during an utterance that is phonated. This is evidence for one or more of three possible effects. One, the excitation process may be affected in such a way that vowels are either given a breathier quality or otherwise modified to confuse the automatic voicing detection algorithm. Two, the reduction in measured voicing may be evidence for in-utterance breaths. Three, the reduction in measured voicing may be evidence for increased durations of unvoiced segments of speech. Whether any of these possibilities account for the reduction in measured voicing cannot be determined from available data and could be the subject of future research.

Finally, it should be noted that for most speakers, statistically significant changes in the average measured fundamental frequency are observed. This constitutes evidence that physical task stress affects speech production in the vicinity of the glottis, which coincides with the measured decrease in voicing. There are two possible mechanisms by which fundamental frequency varies during the speech production process. Fundamental frequency may vary due to a change in transglottal pressure, and it may vary due to changes in the stiffness of the vocal folds, with changes in the stiffness of the vocal folds being the dominant means by which fundamental frequency is varied (Stevens (1998), p.73). Future research could investigate which of these mechanisms results in the observed changes in fundamental frequency. It should also be noted that, despite the observed changes in average fundamental frequency, most speakers demonstrate an ability to maintain some control over the variation of fundamental frequency within an utterance, as for most speakers it can be observed that the variation in fundamental frequency in an utterance is not found to change.

It is clear from the measurements presented in this chapter, and from previous re-

search, that physical task stress affects the speech production process is measurable ways, but that a full understanding of the effects of physical task stress on speech production remains distant. In the next chapter another step is taken to broaden the understanding of what aspects of speech production may be affected the most by physical task stress. The experiments of this chapter have relied on analysis of various parameters of the speech signal. In Chapter 5 a different approach is taken. The cepstrum, shown in Fig. 4.2 to capture more variation in the speech signal in individual quefrency bins than the spectrum, is applied in an analysis to determine which phone classes are most affected by physical task stress, regardless of the particular parameter affected. In the conclusion, Chapter 7, the results of the two different approaches from Chapters 4 and 5 are discussed to draw new conclusions about the nature of physical task stress speech.

We close this discussion of the parameter analysis of physical task stress speech with some points about automatic parameter analyses in general. Automatic methods have both limitations and advantages as compared to making measurements by hand, perhaps by using a spectrogram. Automatic measurements may have more errors, or their error behavior on styled or stressed speech may not be well understood. Also, because the audio and spectrograms are not examined by humans, unforeseen effects of the stress or variability may remain undetected. However, automatic methods support experiments on very large databases at a fraction of the cost of by-hand analysis. This fact has been exploited in this thesis to support drawing broadly speaker independent conclusions about certain aspects of speech under physical task stress.

CHAPTER 5

PHONE CLASS ANALYSIS

In this chapter is discussed the design, results, and interpretation of an experiment that determines which phone classes are most affected by physical task stress. As discussed in Chapter 2, studies of stressed speech, including that presented in Chapter 4, generally attempt to discern and understand the causes of variations in one or more parameters including pitch, formants, or various durations. Some also study the relationship between the variations of those parameters and various phones or phone classes. By contrast, the design of the following experiment is motivated by a desire to explore the relative effects of physical task stress on various phone classes, irrespective of any specific parameter of the speech - i.e. to determine, in the most general sense possible, which phone classes are most affected by physical task stress. Such knowledge could be applied in several ways. One is to direct future exploration into the effects of physical task stress to those phones that are most affected. Another is to apply the knowledge to focus improvements in robustness of recognition systems on those phone classes that are most affected. At the conclusion of this study, the results of this analysis will compared to the parameter analysis discussed in Chapter 4 in order to draw new insight into what speech production mechanisms might be most strongly influenced by physical task stress, and what that may imply for the design of speech systems.

5.1 Overview

As noted, the basic purpose of the following experiment is to rank order the phone classes of American English in terms of the overall effect that physical task stress has on their production. The basic structure of the experiment, shown in Figure 5.1, is to first apply a speaker independent neutral/stress classification system to the speech. Then, from the

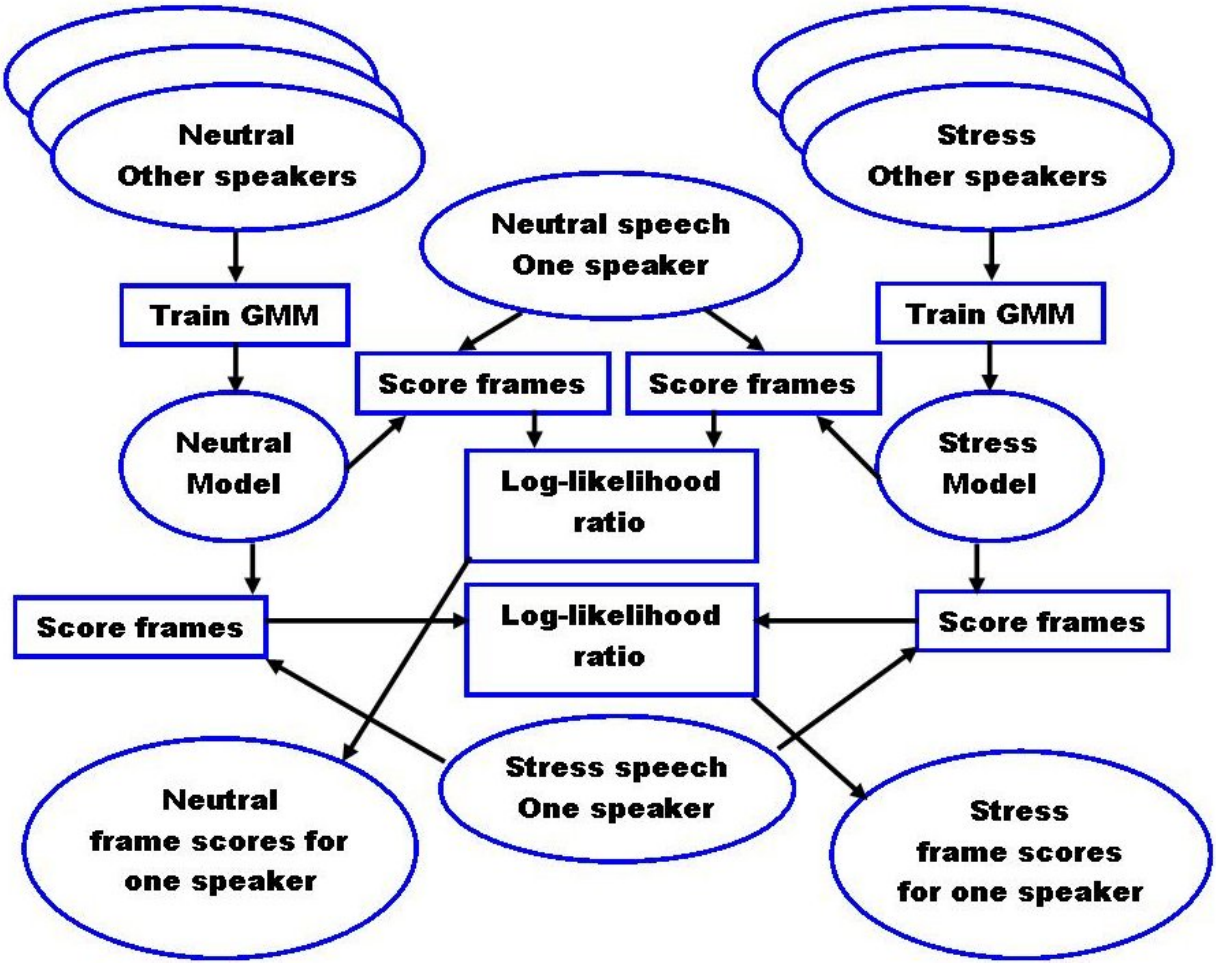

Figure 5.1. Process of computing the frame scores for one speaker (256 mixture GMMs, 15-dim MFCCs, with 15-dim delta and double-delta coefficients, C0 included).

classification results it is estimated which phone classes are most affected by physical task stress. It is important to note that the particular interpretation of the results presented is dependent on the particular structure of the classification system employed. The structure of the classification system discussed shortly results in a meaningful classification score for very short time durations, which in the experiment will be applied to durations as short as one phone.

The basic structure of the classifier is the common two stage process of feature extraction followed by statistical modeling. Gaussian Mixture Models (GMMs) (Reynolds and Rose, 1995) are the statistical modeler employed. GMMs function by determining a summation of a finite set of multidimensional Gaussian distributions that closely approx-

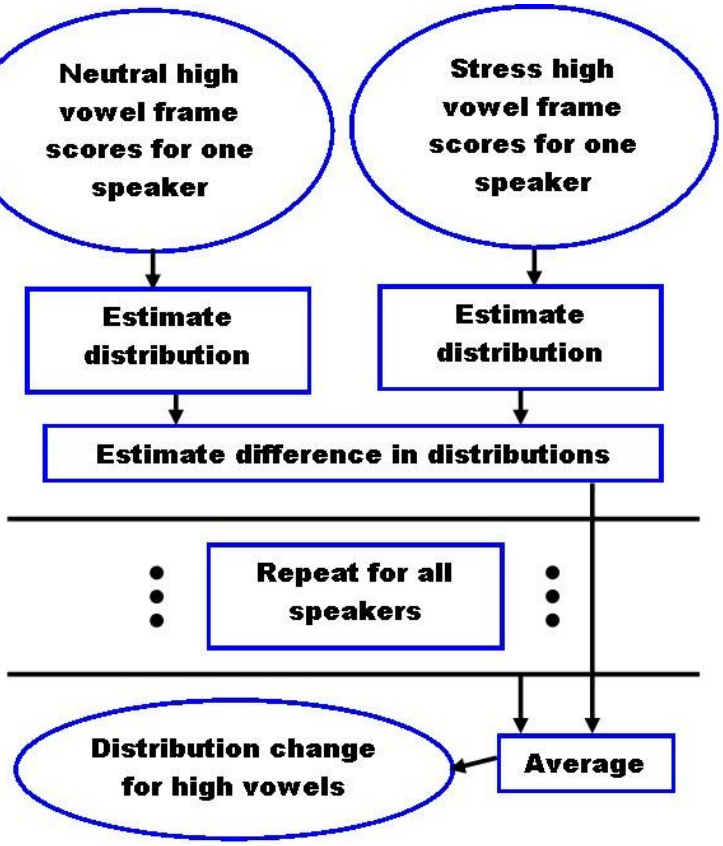

Figure 5.2. Process of computing the classification strength of neutral high vowels.

imates the distribution of the feature data of the training process. Each frame of the test data is scored against the models for neutral and physical task stress; in essence, frames that have higher scores more closely fit the differences in the distributions modeled by the GMMs. In a complete classifier, i.e. one not being used for such experiments, frames are scored individually and their scores averaged and compared to a threshold to make a final classification decision for a given utterance. Because the individual frames are assumed independent and their individual score is not dependent on other frames, it is possible in the following experiment to examine individual frame scores outside of the context of their temporal neighbors. Thus the second part of the experiment is to group all the frame scores for one speaker by the phone class from which that score originated, and then to examine the statistical distribution of the scores within each group. This process is described by Fig. 5.2.

These groups of scores, one group for each combination of phone class, speaker and speech type (neutral/stress), may each be modeled by a probability distribution. In the following experiment, the probability distribution of frame scores for one phone class for one speaker in neutral will be compared to the probability distribution of frame scores for one phone class for one speaker in physical task stress. It will be argued that those phone classes for which the distribution of the frame scores undergoes the most change from neutral to physical task stress must be those phone classes that are most affected by physical task stress.

The following analysis relies on the cepstrum as represented by Mel-Frequency Cepstral Coefficients (MFCCs). This is because the cepstrum is mostly parameter independent; as seen in Chapter 4 individual parts of the cepstrum encompass the variation in speech due to physical task stress, rather than the variation being captured in the structure of the cepstrum, as is the case for the spectrum. Therefore by use of the cepstrum the experiment closely approximates an analysis of the total variation of the speech signal within each phone class, irrespective of the nature of the variation.

5.2 Alternative experiment designs

In this section two principle concerns are addressed regarding the design of the experiment. First, are the automatic classification methods reliable enough to build an experiment upon? And second, are there other ways to formulate the classifier?

To evaluate whether the classification method employed for analysis is reliable enough to result in meaningful conclusions, the system performance was evaluated on a per-speaker basis. The per-speaker accuracy of the classification system used in the experiment, when employing optimum speaker-specific classification threshoulds, is shown in Table 5.1. For 25 of the 42 speakers, the system has 100% accuracy at classifying 10 second utterances as either neutral or stress tokens, when employing the optimum threshold for each speaker. For most of the remaining speakers, the classification accuracy is very high, with a total average of 95%. It is argued that this implies that the classification systems have modeled the most significant and common effects of physical task stress on speech production, and are therefore suitable for the following experiment. It should be noted that speaker dependent thresholds are justified here, on the grounds that the purpose of this evaluation is solely to determine the ability of the system to separate neutral frames from physical task stress frames for a specific, known speaker.

In order to examine whether the low performance observed for some of the speakers is related to the exertion level, a scatter plot of this data is shown in Fig. 5.3. It can be seen from the plot that there is a trend towards lower exertion levels for those speakers which are classified poorly. It can also be seen that there are speakers for whom high classification performance is observed but who demonstrated a low average exertion level. It may be concluded that there is some factor that relates system performance to exertion level for certain speakers.

The second concern addressed is whether there may be other ways to formulate the experiment. One alternative formulation would be to train one GMM for each phone class, for both neutral and physical stress, and to compare the classification accuracy for each

Table 5.1. Accuracy for each of the 42 per-speaker classification systems. Each system was trained with all corpus data except that of the indicated speaker, then tested with the data for the indicated speaker.

Spkr	Class. acc.	Spkr	Class. acc.
fac1	100%	frt1	100%
fad1	100%	fss1	100%
fah1	100%	fth1	100%
fch1	100%	fts1	100%
fdb1	100%	fam1	95%
ffl1	100%	fat1	95%
fjc1	100%	fjs1	95%
fjf1	100%	fjw2	95%
fjw1	100%	fms3	95%
fkc1	100%	fmt1	95%
fli1	100%	fsm1	95%
flk1	100%	ftk1	95%
flk2	100%	fcb1	91%
flm1	100%	fmb1	91%
fml1	100%	fjf2	86%
fmp1	100%	fnw1	86%
fms1	100%	fms2	82%
fmw1	100%	fsm2	82%
fnc2	100%	fep1	77%
fnh1	100%	fae1	68%
fnm1	100%	fmf1	68%
Average:			95%

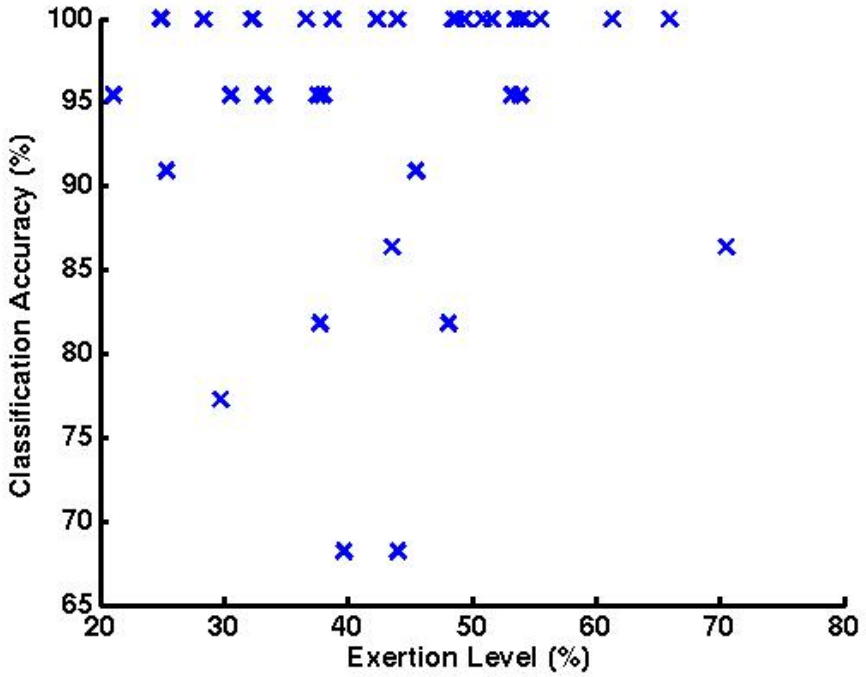

Figure 5.3. A relationship between exertion level and performance for some speakers may be observed from this scatter plot.

phone class, reasoning that the phone classes which may be used in the best classifiers must be those that are most affected by physical task stress. This is not the way the following experiment is formulated because the the performance of the phone-class based classifiers could be dominated by a ceiling affect, in which several of the phone classes result in classification accuracy near 100%. In that case it would be difficult to determine which of the phone classes is most affected by physical task stress.

5.3 Method

This section discusses the design of the experiment in sufficient detail for the experiment to be replicated. The experimental method has two parts: the classification system, which generates a score for each frame of the speech, and distribution comparison.

5.3.1 The classification systems

The classification systems are comprised of feature extraction (MFCCs) and statistical modeling (GMMs), as shown in Figure 5.1. One classification system is built for each speaker, employing as training data the speech from all other speakers. Gaussian Mixture Models (GMMs) result in a meaningful score generated for each frame of the speech; this fact is exploited in the experiment to examine groups of frames outside of their temporal context. Several details of the construction of the classification systems should be noted, including the number of cepstral coefficients, whether delta and double-delta coefficients are included, the number of mixtures that are employed, and the software used; these are summarized in Table 5.2, with a ✓ indicating included features. To determine these parameters, they were varied until the best overall equal error rate (EER) was found (when using the same decision threshold for all 42 systems), or, in the case of the number of mixtures, the performance became negligibly higher for an increasing number of mixtures.

It should be noted that the purpose of the classification system development is to build the best possible classification systems *for this dataset* rather than to build a classification system that could be deployed in the real world. It therefore is fair to use all of the available data in the development of the classifier, while this would not be the case were the purpose to estimate the classification systems' performance in real-world deployment. It should also be noted that scoring code to compute scores for individual frames was developed in-house. To verify the accuracy of the frame score computation, the average scores for utterances computed by the frame score code were checked against the results from HTK and found to be equivalent.

5.3.2 Comparing score distributions

In this section is described the method to compare the distribution of the frame scores for one speaker for one phone class in neutral to the distribution of the frame scores for

Table 5.2. Aspects of the classification system used in this thesis.

Parameter	Value used in experiment
Features used	MFCCs
# of mixtures	256
# of cepstral coefficients	15
Delta coefficients	✓
Double-delta coefficients	✓
Training software	HTK
Training speakers	41
Test speakers	1
Testing style	Round-robin
Equal error rate	15%
Global threshold	-0.1670

one speaker for one phone class in physical task stress. The comparison is performed in two steps. In the first step, the frame scores are modeled with a particular probability distribution. This provides a closed form expression of the probability distribution, facilitating comparison. In the second step, the distributions are compared using the Kullback-Leibler Divergence (KL Divergence) (Kullback and Leibler, 1951).

The KL Divergence is an expression designed to compare two probability distributions $p(x)$ and $q(x)$:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) log \frac{p(x)}{q(x)} dx \tag{5.1}$$

In this form the KL Divergence is not symmetric. For the purposes of this experiment the distance between probability distributions will be measured by:

$$D_{KL}(P,Q) = D_{KL}(P||Q) + D_{KL}(Q||P) \tag{5.2}$$

Figure 5.4 shows a histogram of frame scores from fricatives spoken by speaker fmb1 under physical task stress. A Laplacian probability distribution is shown fit to the histogram. It will be assumed that the distribution of frame scores for all phone classes is Laplacian. Table 5.3 shows justification for this assumption. It can be seen in the table that for most phone classes, frame scores from over 75% of the speakers are consistent with a Laplacian distribution. Statistical tests for the fit of a Gaussian distribution to the frame

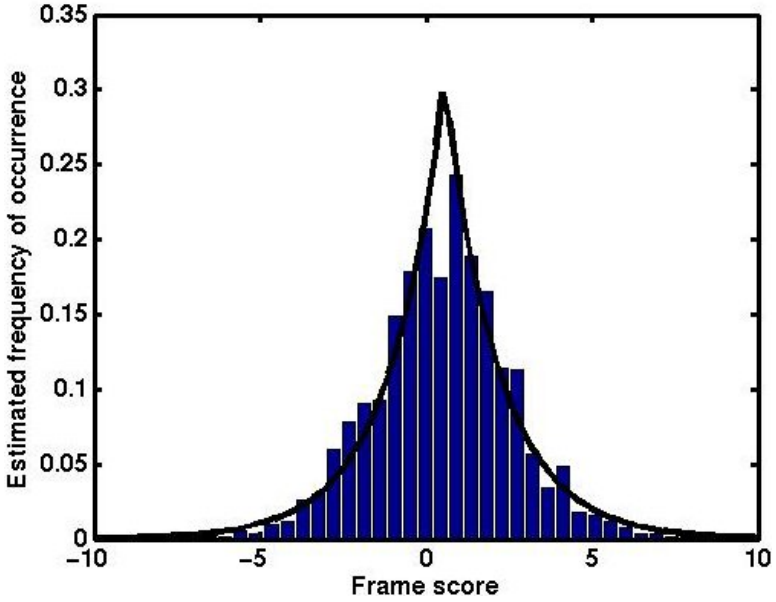

Figure 5.4. Speaker fmb1 physical stress fricative frame scores, with estimated Laplacian distribution.

scores found that assuming a Gaussian distribution for frame scores was a reasonable assumption for less than 10% of speakers, for most phone classes.

The Laplacian probability distribution is:

$$p(x) = \frac{1}{2b} exp(-\frac{|x-\mu|}{b}) \tag{5.3}$$

Having assumed that $p(x)$ and $q(x)$ in Eq. 5.1 are Laplacian and thus take the form of Eq. 5.3, it remains to find a closed form expression of Eq. 5.2. This is accomplished by substituting Eq. 5.3 into Eq. 5.1, solving the integration, and substituting the result into Eq. 5.2. This substitution and solution process is straightforward but tedious; a complete derivation may be found in Appendix 7.2. The final result is:

$$\begin{aligned} D_{KL}(N,S) = b_n + b_s &- b_n \exp\left(\tfrac{-|\mu_n-\mu_s|}{b_n}\right) \\ &- b_s \exp\left(\tfrac{-|\mu_n-\mu_s|}{b_s}\right) \end{aligned} \tag{5.4}$$

where parameters b_n and μ_n are the parameters of the Laplacian distribution for the neutral frame scores, and parameters b_s and μ_s are the parameters of the Laplacian distribu-

Table 5.3. Results of statistical tests to determine whether frame scores within each phone class for each speaker are Laplacian distributed at the 99% confidence level.

Phone class	% of spkrs with scores Laplacian dist.
Neutral - low non-R vowels	87%
Phy - low non-R vowels	89%
Neutral - high non-R vowels	59%
Phy - high non-R vowels	66%
Neutral - laterals	100%
Phy - laterals	100%
Neutral - stop plosives	87%
Phy - stop plosives	92%
Neutral - diphthongs	97%
Phy - diphthongs	87%
Neutral - R vowels	100%
Phy - R vowels	97%
Neutral - fricatives	84%
Phy - fricatives	76%
Neutral - glides	97%
Phy - glides	97%
Neutral - nasals	82%
Phy - nasals	89%
Neutral - combo consonants	100%
Phy - combo consonants	100%

tion for the physical task stress frame scores. The parameter μ is estimated as the median of the available data samples, and the maximum likelihood estimator for b is:

$$b = \frac{1}{M} \sum_{fric} |s[n] - \mu| \tag{5.5}$$

where M is the number of frames of that phone class.

Eqs. 5.4 and 5.5 therefore form the measurement that is applied to compute the changes affected by physical task stress on the distribution of frame scores within one phone class.

5.4 Discussion on method

It is important to group the phone classes in such a way that they may be modeled by a unimodal probability density function for both conceptual and practical reasons. For practical reasons, it is important that a closed form solution to Eq. 5.1 be found. For

conceptual reasons, modeling the frame scores with a unimodal density function assures that the phone class under examination represents a group of phones that are affected similarly by physical task stress. When that is the case, it is reasonable to assume that the quantity being measured is the effect of physical task stress on the members of that class, rather than separate, averaged effects. This is the reason that the vowels have been split into three groups: R-colored vowels, high vowels, and low vowels. Grouping all the vowels together does not result in a unimodal distribution, while separating out the R-colored vowels results in a unimodal Laplacian distribution for the R-colored vowels. For the non-R-colored vowels, other separations, such as dividing the vowels into front and back vowels, do not result in unimodal distributions. It may thus be hypothesized that physical task stress affects high and low vowels differently.

5.5 Results and discussion

The experiment was conducted as described and the results are shown in Figure 5.5. The experimental results show that nasal phones undergo the greatest change in score distribution from neutral to physical task stress. Laterals, high vowels, and diphthongs could be grouped together as being the next most affected phone classes after nasals. The score distribution, and therefore the speech signal, and, it is thus argued, the production process of fricatives and plosives, are affected the least by physical task stress.

Voiced sounds, with the exception of glides, show a trend of being more affected by physical task stress than unvoiced sounds. This may be related to the results on fundamental frequency and voicing changes seen in Chapter 4. Glides are somewhat of a surprise; it is not clear for what reason they appear to be affected less by physical task stress than other voiced sounds. It may be that the specific position of the articulators in glides, one that is more constricted than that of vowels, results in a speech production process that is more easily controlled when the speaker is under physical task stress than the relatively more open vowels. Further investigation into this possibility, or into alternate explanations for the differences seen between nasals, diphthongs, and vowels is

clearly warranted.

Considering fricatives and plosives, their noise character when produced in neutral speaking conditions may make them similar to the breath noise that we expect to see in physical task stress utterances. Thus perhaps the acoustic signal resulting from their production in both neutral and physical task stress is not clearly distinguishable from other noises produced under physical task stress.

It is clear from the figure, however, that all phone classes underwent a change in score distribution from neutral to physical task stress, and that therefore all the phone classes must undergo some change from neutral production to physical task stress production. It should also be noted that the experimental results give only a measure of the overall amount that physical task stress affects the production of each phone class. R-vowels and low vowels are, according to the results, affected in some sense a similar amount by physical task stress, but the experimental results are silent on the question of whether the mechanism by which they are affected is the same.

The conclusions that may be drawn from the experimental results are discussed further at the conclusion of this thesis.

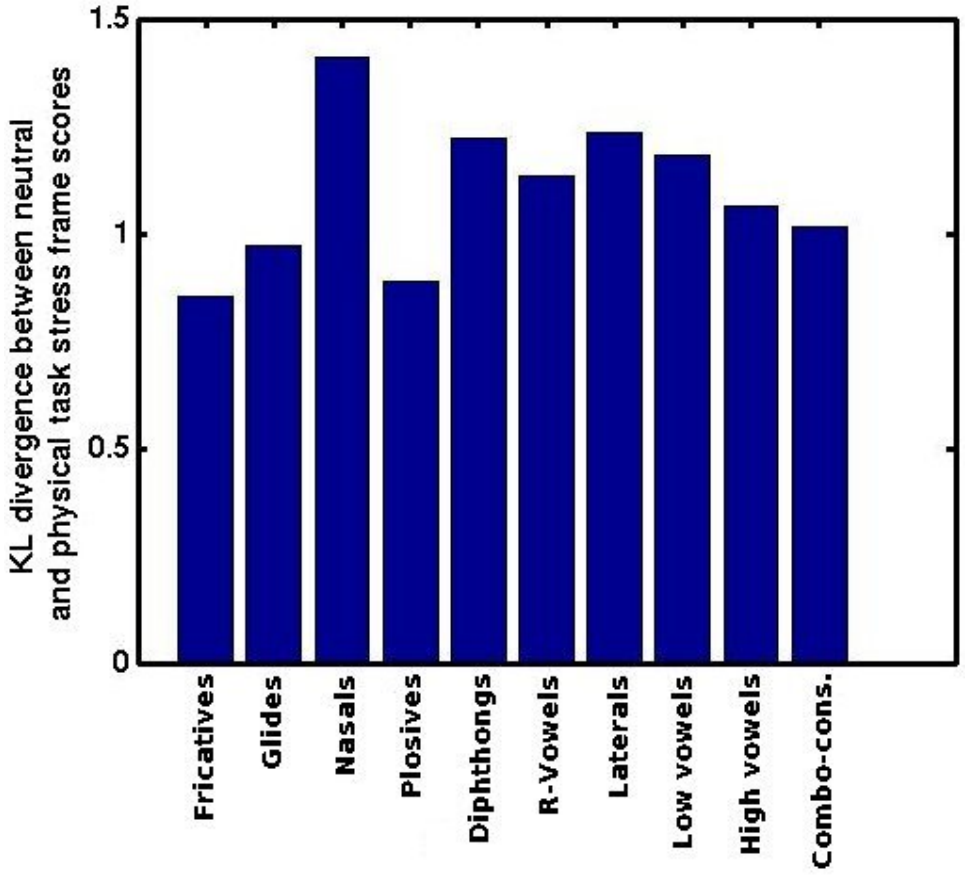

Figure 5.5. Overall results of the phone class classification power comparison.

CHAPTER 6

PERCEPTUAL STUDY

A listener test was conducted to examine some aspects of the perception of speech under physical task stress. The analyses described in Chapter 4 revealed that changes in F_0 and the glottal waveform are strong correlates with physical task stress. Listener tests were therefore performed to determine the strength of these acoustic correlates as perceptual cues, and also to establish listener performance on a stress classification task. Here, 10 subjects were asked to classify 84 utterances, describing the speaker as either "exercising", or "seated, resting".

6.1 Listener test procedure

The physical task stress speech had heavy breathing surrounding the utterances, so as a first step all of the utterances were closely cropped in time to remove their context. This helped to ensure that listeners made their classification decision based on the speech itself, and not on surrounding breath sounds. To test the strength of F_0 and the glottal waveform as perceptual cues, some of the utterances were modified so that they were in some way shifted towards the counterpart utterance in the opposite condition. To shift the F_0 of some utterances, a PSOLA technique was applied so that the given utterance had a mean F_0 equal to the mean F_0 of the same speaker's utterance in the opposite condition. For the glottal waveform tests, the glottal inverse filtering method described in Childers and Lee (1991) was used to extract glottal waveforms from the voiced portions of the speech, which were then replaced with waveforms extracted from the opposite condition. The utterances were then reconstructed by inverting the process and concatenating them with the unmodified unvoiced portions of the utterances.

The 84 utterances were partitioned into 8 groups. Two groups of 10 utterances, one

Table 6.1. Results of listener tests and statistical comparisons. ✓ indicates statistical significance. W.R.T. means with respect to.

Category	Perf.	Sig., W.R.T.
Unprocessed neutral	84.4 %	N/A
Unprocessed physical	68.9 %	N/A
Neut. $< F_0 >$ shifted to phy.	82.8 %	-, unproc. neut
Phy. $< F_0 >$ shifted to neut.	44.4 %	✓, unproc. phy
Replace neut. glottal waveform w/ neut.	48.9 %	✓, unproc. neut
Replace phy. glottal waveform w/ phy.	66.7 %	-, unproc. phy
Replace neut. glottal waveform w/ phy.	61.6 %	- proc. neut
Replace phy. glottal waveform w/ neut.	71.7 %	-, proc. phy

from each condition, were left unprocessed. Two groups of 11 utterances were processed as described above to shift their pitch. Two groups, each comprising 10 utterances, served as control groups for the glottal processing technique. These were processed using glottal waveforms from the same condition. Two groups served as experimental groups, each of 11 utterances. These utterances were synthesized using glottal waveforms extracted from that speaker's opposite condition utterance. The order of all 84 utterances was randomized, and then presented to listeners in a formal test.

6.2 Listener test results

The listener test results are summarized in Table 6.1. The table shows that the listeners correctly classified 84.4 % of the unprocessed neutral utterances, and 68.9 % of the physical task stress utterances. Student-t tests were also used to make comparisons between the test results. The results for pitch shifted utterances were compared with those from unprocessed utterances of the same condition. Shifting the pitch of the physical stress utterances caused a statistically significant decrease in listener performance of more than 20 %. Shifting the pitch of the neutral speech did not have an effect on performance.

The results for the utterances which underwent glottal waveform replacement from the same condition were compared with unprocessed utterances to determine if the processing method had an effect on listener performance. The processing method did not have a statistically significant effect on the listeners' ability to mark utterances as physical

task stress, though the processing decreased performance on neutral utterances to chance levels. In comparing the utterances with glottal waveforms swapped from opposite conditions to those with waveforms from the same conditions, no statistically significant shift was found.

6.3 Discussion

In retrospect, the results of the glottal waveform tests are not reliable indicators of perceptual behavior regarding glottal waveforms. This is because processing artifacts likely dominated the listeners' perceptual experience, which has been confirmed by informal listening. This is also supported by the fact that processing the neutral utterances to remain neutral resulted in a statistically significant shift in performance; no shift should be observed if the processing method did not result in clearly audible artifacts. These results are included here because they were part of the experiment as conducted. The results for the other types of processing remain valid, as they were presented independently to the listeners.

We can conclude from the results of this experiment that listeners are able to discern speech production differences that result from physical task stress, even when the utterance is presented out of context. That the speech production process is affected by physical task stress in fundamental ways is thus corroborated by the acoustic parameter analysis of Chapter 4, the phone class analysis presented in Chapter 5, and the results of the formal listener just presented.

CHAPTER 7

CONCLUSION

This thesis has presented and discussed experimental results concerning the nature of speech under physical task stress. It is clear from these results that physical task stress has an affect on many aspects of the speech production process.

7.1 Results summary

Chapter 4 presented results on a parameter analysis. It was found that three of the four parameters varied. Additionally, f-ratio analysis showed that the variation due to physical task stress does not concentrate in specific frequency bins but must instead by associated with the temporal structure of the spectrum, but that variance due to physical task stress may be found in specific areas of the cepstrum. Thus we may hypothesize, based on this result, that specific speech production processes are consistently affected by physical task stress.

Chapter 5 presented a new analysis method, and the results thereof, that showed that production of nasals and laterals are the most affected by physical task stress, and that the production of fricatives and plosives are the least affected. Chapter 6 presented results of a listener test that showed that listeners are able to discern physical task stress in speech from parameters beyond the breaths inherent in the speech. We can thus conclude that physical task stress affects several aspects of the speech production process.

This has implications for the development of speech systems as well as for further scientific inquiry into the nature of physical tasks stress speech production and perception. We have sought to describe physical task stress in a way that goes beyond description of the changes in measured parameters to discuss possible ways that physical task stress affects speech production processes.

7.2 Future work

It is clear that the research performed thus far has resulted in some general insights into the nature of physical task stress speech. However, it is only a beginning as many more general and specific questions remain to be asked and answered. For example, a variety of further analysis questions might be investigated, such as How many speakers add intra-utterance breaths and how often is this employed? Does it depend on the linguistic structure of the utterance or on the particular phones employed? What is the mechanism by which fundamental frequency is varied? Additionally, as discussed in Chapter 2, further investigation of physical task stress should consider other definitions of the ground-truth level of physical task stress, and continue investigation into whether there is one or more specific speech styles associated with physical task stress and the nature of those styles.

DERIVATION OF THE KL DIVERGENCE FOR TWO LAPLACIAN PDFS

In Chapter 5 the KL divergence between two Laplacian PDFs was employed as a measurement of how much each phone class was affected by physical task stress. In this Appendix this divergence is derived. The derivation of the divergence itself is presented in Section A.1. The derivation relies on other formulations that are presented in subsequent sections. All of the derivations in this Appendix employ commonly known techniques for integration.

A.1 Derivation of main result

This section presents the derivation of the main result, seen previously in Eq. 5.2. The symmetric KL divergence between $p(x)$ and $q(x)$ is defined to be:

$$D_{KL}(p,q) = D_{KL}(p||q) + D_{KL}(q||p) \tag{A.1}$$

where the KL divergence of $p(x)$ from $q(x)$ is defined as:

$$D_{KL}(p||q) = \int_{-\infty}^{\infty} p(x) ln \left(\frac{p(x)}{q(x)} \right) dx. \tag{A.2}$$

The Laplacian probability density functions $p(x)$ and $q(x)$ take the form:

$$p(x) = \frac{1}{2b_p} \exp \left(\frac{-|x - \mu_p|}{b_p} \right) \tag{A.3}$$

and therefore:

$$D_{KL}(p||q) = \int_{-\infty}^{\infty} \frac{1}{2b_p} \exp\left(\frac{-|x-\mu_p|}{b_p}\right) \ln\left(\frac{b_q}{b_p} \exp(|x-\mu_q| - |x-\mu_p|)\right) dx \tag{A.4}$$

$$= \frac{1}{2b_p} \int_{-\infty}^{\infty} \exp\left(\frac{-|x-\mu_p|}{b_p}\right) [\ln b_q - \ln b_p + |x-\mu_q| - |x-\mu_p|]\, dx \tag{A.5}$$

$$\begin{aligned} = \; & (\ln b_q - \ln b_p) \frac{1}{2b_p} \int_{-\infty}^{\infty} \exp\left(\frac{-|x-\mu_p|}{b_p}\right) dx + \frac{1}{2b_p} \int_{-\infty}^{\infty} \exp\left(\frac{-|x-\mu_p|}{b_p}\right) |x-\mu_q| dx \\ & - \frac{1}{2b_p} \int_{-\infty}^{\infty} \exp\left(\frac{-|x-\mu_p|}{b_p}\right) |x-\mu_p| dx \end{aligned} \tag{A.6}$$

Clearly the first term equals $\ln b_q - \ln b_p$, as the PDF integrates to 1. The second and third terms will be integrated separately. For the remainder of the derivation, it will be assumed, without loss of generality, that $\mu_p \geq \mu_q$. For the second term,

$$\begin{aligned} \frac{1}{2b_p} \int_{-\infty}^{\infty} \exp\left(\frac{-|x-\mu_p|}{b_p}\right) |x-\mu_q| dx \; = \; & \frac{1}{2b_p} \Big[\int_{-\infty}^{\mu_q} \exp\left(\frac{x-\mu_p}{b_p}\right) (x-\mu_q) dx \\ & + \int_{\mu_q}^{\mu_p} \exp\left(\frac{x-\mu_p}{b_p}\right) (\mu_q - x) dx \\ & + \int_{\mu_p}^{\infty} \exp\left(\frac{\mu_p - x}{b_p}\right) (\mu_q - x) dx \Big] \end{aligned} \tag{A.7}$$

$$\begin{aligned} = \; & \frac{1}{2b_p} \Big[\int_{-\infty}^{\mu_q} x \exp\left(\frac{x-\mu_p}{b_p}\right) dx - \mu_q \int_{-\infty}^{\mu_q} \exp\left(\frac{x-\mu_p}{b_p}\right) dx \\ & + \mu_q \int_{\mu_q}^{\mu_p} \exp\left(\frac{x-\mu_p}{b_p}\right) dx - \int_{\mu_q}^{\mu_p} x \exp\left(\frac{x-\mu_p}{b_p}\right) dx \\ & + \mu_q \int_{\mu_p}^{\infty} \exp\left(\frac{\mu_p - x}{b_p}\right) dx - \int_{\mu_p}^{\infty} x \exp\left(\frac{\mu_p - x}{b_p}\right) dx \Big] \end{aligned} \tag{A.8}$$

Applying supporting results from Section A.2 results in:

$$= \left(\frac{\mu_q - b_p}{2}\right) \exp\left(\frac{\mu_q - \mu_p}{b_p}\right) - \mu_q \left(\frac{1}{2}\exp(\left(\frac{\mu_q - \mu_p}{b_p}\right)\right) + \frac{\mu_q}{2} - \frac{\mu_q}{2}\exp\left(\frac{\mu_q - \mu_p}{b_p}\right)$$
$$- \frac{\mu_p}{2} + \frac{b_p}{2} + \frac{\mu_q}{2}\exp\left(\frac{\mu_q - \mu_p}{b_p}\right) - \frac{b_p}{2}\exp\left(\frac{\mu_q - \mu_p}{b_p}\right) + \frac{\mu_q}{2} - \frac{\mu_p + b_p}{2} \tag{A.9}$$
$$= \mu_q - \mu_p - b_p \exp\left(\frac{\mu_q - \mu_p}{b_p}\right) \tag{A.10}$$

Now to derive the third term from Eq. A.6:

$$\frac{1}{2b_p}\int_{-\infty}^{\infty} \exp\left(\frac{-|x-\mu_p|}{b_p}\right)|x-\mu_p|dx$$
$$= \int_{-\infty}^{\mu_p} \exp\left(\frac{x-\mu_p}{b_p}\right)(x-\mu_p)dx + \int_{\mu_p}^{\infty} \exp\left(\frac{\mu_p - x}{b_p}\right)(\mu_p - x)dx \tag{A.11}$$
$$= \int_{-\infty}^{\mu_p} x \exp\left(\frac{x-\mu_p}{b_p}\right)dx - \mu_p \int_{-\infty}^{\mu_p} \exp\left(\frac{x-\mu_p}{b_p}\right)dx$$
$$+ \mu_p \int_{\mu_p}^{-\infty} \exp\left(\frac{\mu_p - x}{b_p}\right)dx - \int_{\mu_p}^{\infty} x \exp\left(\frac{\mu_p - x}{b_p}\right)dx \tag{A.12}$$

Applying supporting results for the first and last integral of Eq. A.12 and noting that the middle two integrals cancel results in:

$$\frac{1}{2b_p}\int_{-\infty}^{\infty} \exp\left(\frac{-|x-\mu_p|}{b_p}\right)|x-\mu_p|dx = b_p \tag{A.13}$$

Thus substituting Eqs. A.10 and A.13 into Eq. A.6 results in:

$$D_{KL}(p||q) = \ln b_q - \ln b_p + b_p + \mu_p - \mu_q - b_p \exp\left(\frac{\mu_q - \mu_p}{b_p}\right) \tag{A.14}$$

Then $D_{KL}(q||p)$ for the case where $\mu_p \geq \mu_q$ may be derived by similar techniques. The result is:

$$D_{KL}(q||p) = \ln b_p - \ln b_q + b_q + \mu_q - \mu_p - b_q \exp\left(\frac{\mu_q - \mu_p}{b_q}\right) \tag{A.15}$$

Finally, substituting Eqs. A.14 and A.15 into Eq. A.1, and noting that the symmetricity

of the equation allows us to include the absolute value function to generalize away our assumption that $\mu_p \geq \mu_q$, results in the formula employed in this thesis:

$$D_{KL}(p,q) = b_p + b_q - b_p \exp\left(\frac{-|\mu_q - \mu_p|}{b_p}\right) - b_q \exp\left(\frac{-|\mu_q - \mu_p|}{b_q}\right) \tag{A.16}$$

A.2 Supporting results

In this section several supporting results for the derivation of the main result are listed. Their derivation is straightforward from elementary calculus. The first such result is the identity:

$$\lim_{x \to -\infty} x \exp(x) = 0 \tag{A.17}$$

The second useful identity is:

$$\int x e^{cx} dx = \frac{e^{cx}}{c^2}(cx - 1) \tag{A.18}$$

Also,

$$\int_a^b \exp\left(\frac{a-x}{c}\right) dx = c - c \exp\left(\frac{a-b}{c}\right) = \int_a^b \exp\left(\frac{x-b}{c}\right) \tag{A.19}$$

And finally,

$$\int_a^b x \exp\left(\frac{a-x}{c}\right) dx = c^2 + ca - bc \exp\left(\frac{a-b}{c}\right) - c^2 \exp\left(\frac{a-b}{c}\right) \tag{A.20}$$

$$\int_a^b x \exp\left(\frac{x-b}{c}\right) dx = cb - c^2 - ca \exp\left(\frac{a-b}{c}\right) + c^2 \exp\left(\frac{a-b}{c}\right) \tag{A.21}$$

BIBLIOGRAPHY

M. Benzeghiba, R. De Mori, O Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens. Automatic speech recognition and speech variability: A review. *Speech Comm.*, 49:763–786, 2007.

Z. S. Bond, Thomas J. Moore, and Beverley Gable. Acoustic-phonetic characteristics of speech produced in noise and while wearing an oxygen mask. *JASA*, 85:907–912, February 1989.

Hynek Boril and John H. L. Hansen. Unsupervised equalization of lombard effect for speech recognition in noisy adverse environments. *IEEE Trans. on Audio, Speech, and Lang. Proc.*, In Press, 2009.

Felix Burkhardt, Tim Polzehl, Joachim Stegmann, Florian Metze, and Richard Huber. Detecting real life anger. In *IEEE Intl. Conf. on Acoustics, Speech, and Sig. Proc.*, 2009.

Douglas A. Cairns and John H. L. Hansen. Nonlinear analysis and classification of speech under stressed conditions. *Journal of the Acoustical Society of America*, 96(6):3392–3400, December 1994.

Joseph P. Campbell. Speaker recognition: A tutorial. *Proc. of the IEEE*, 85:1437–1462, Sept. 1997.

Henry S. Cheang and Marc D. Pell. The sound of sarcasm. *Speech Comm.*, 50:366–381, 2008.

Yeunung Chen. Cepstral domain talker stress compensation for robust speech recognition. *Acoustics, Speech, and Sig. Proc., IEEE Trans. on*, 36(4):433–439, April 1988.

D. G. Childers and C. K. Lee. Vocal quality factors: Analysis, synthesis, and perception. *JASA*, 90(5):2394–2410, Nov. 1991.

Kathleen E. Cummings and Mark A. Clements. Analysis of the glottal excitation of emotionally styled and stressed speech. *JASA*, 98(1):88–98, July 1995.

John R. Deller, John H. L. Hansen, and John G. Proakis. *Discrete-Time Processing of Speech Signals*. IEEE Press, 2000.

Marcia S. Entwistle. *Training methods and enrollment techniques to improve the performance of automated speech recognition systems under conditions of human exertion*. PhD thesis, Dept. of Psychology, U. of South Dakota, 2005.

Marcia Seivert Entwistle. The performance of automated speech recognition systems under adverse conditions of human exertion. *Human Computer Interaction, Intl. J. of*, 16(2): 127–140, 2003.

Xing Fan and John H. L. Hansen. Speaker identification with whispered speceh based on modified LFCC parameters and feature mapping. In *Acoustics, Speech, and Sig. Proc., IEEE Intl. Conf. on*, 2009.

Keith W. Godin. An explanation of physical stress classification performance in terms of observed production changes. Master's thesis, Univ. of Texas at Dallas, Richardson, TX, USA, Dec. 2009.

Keith W. Godin and John H. L. Hansen. Analysis and perception of speech under physical task stress. In *Interspeech*, Sep. 2008.

H. P. Greeley, E. Friets, J.P. Wilson, S Raghavan, J. Picone, and J. Berg. Detecting fatigue from voice using speech recognition. In *IEEE International Symposium on Signal Processing and Information Technology*, pages 567–571, August 2006.

John H. L. Hansen. *Analysis and compensation of stressed and noisy speech with application to robust automatic recognition*. PhD thesis, Georgia Inst. of Tech., July 1988.

John H. L. Hansen and Vaishnevi Varadarajan. Analysis and compensation of lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition. *Audio, Speech, and Lang. Proc., IEEE Trans. on*, 17:366–378, 2009.

John H L Hansen and Brian D. Womack. Feature analysis and neural network-based classification of speech under stress. *Speech and Audio Processing, IEEE Transactions on*, 4 (4):307–313, July 1996.

Michael H.L. Hecker, Kenneth N. Stevens, Gottfried von Bismark, and Carl E. Williams. Manifestations of task-induced stress in the acoustic speech signal. *Journal of the Acoustical Society of America*, 44(4):993–1001, October 1968.

Eva B. Holmberg, Robert E. Hillman, and Joseph S. Perkell. Glottal airflow and transglottal air pressure measurements for male and female speaker in soft, normal, and loud voice. *JASA*, 84:511–529, 1988.

Yusuke Ijima, Makoto Tachibana, Takashi Nose, and Takao Kobayashi. Emotional speech recognition based on style estimation and adaptation with multiple-regression hmm. In *IEEE Intl. Conf. on Acoustics, Speech, and Sig. Proc.*, 2009.

Ayako Ikeno, Vaishnevi Varadarajan, Sanjay Patil, and John H. L. Hansen. UT-Scope: Speech under lombard effect and cognitive stress. In *Aerospace Conference 2007, IEEE*, 2007.

S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

Chul Min Lee and Shrikanth S. Narayanan. Toward detecting emotions in spoken dialogs. *Speech and Audio Proc., IEEE Trans. on*, 13(2):293–303, March 2005.

Anders Lindstrom, Jessica Villing, Staffan Larsson, Alexander Seward, Nina Aberg, and Cecilia Holtelius. The effect of cognitive load on disfluencies during in-vehicle spoken dialogue. In *Interspeech*, 2008.

Xugang Lu and Jianwu Dang. An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. *Speech Comm.*, 50:312–322, May 2008.

Youyi Lu and Martin Cooke. Speech production modifications produced in the presence of low-pass and high-pass filtered noise. *J. of the Acoustical Soc. of Am.*, 126:1495–1499, 2009.

Kazumi Maniwa, Allard Jongman, and Travis Wade. Acoustic characteristics of clearly spoken english fricatives. *J. Acoustical Soc. Am.*, 125:3962–3973, June 2009.

Iain R. Murray, Chris Baber, and Allan South. Towards a definition and working model of stress and its effects on speech. *Speech Comm.*, 20(3):3–12, November 1996.

Sanjay A. Patil. *Alternate Sensor Based Speech Systems for Speaker Assessment and Robust Human Communication*. PhD thesis, The Univ. of Texas at Dallas, 2009.

Sanjay A. Patil and John H. L. Hansen. Detection of speech under physical stress: Model development, sensor selection, and feature fusion. In *Interspeech*, 2008.

Douglas A. Reynolds and Richard C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *Speech and audio proc., IEEE Trans on*, 3: 72–83, Jan. 1995.

Abhijeet Sangwan. *Speech system advancements based on phonological features*. PhD thesis, The Univ. of Texas at Dallas, 2009.

Klaus R. Scherer. A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology. In *Int. Conf. on Spoken Lang. Proc.*, pages 379–382, 2000.

Klaus R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Comm.*, 40(1):227–256, 2003.

Vidhyasaharan Sethu, Eliathamby Ambikairajah, and Julien Epps. Speaker dependency of spectral features and speech production cues for automatic emotion classification. In *IEEE Intl. Conf. on Acoustics, Speech, and Sig. Proc.*, 2009.

Kare Sjolander and Jonas Beskow. Wavesurfer - an open source speech tool. In *Proc. ICSLP*, 2000.

Kenneth N. Stevens. *Acoustic Phonetics*. MIT Press, 1998.

H. Tanaka, K. D. Monahan, and D. R. Seals. Age-predicted maximual heart rate revisited. *J. Am. Coll. Cardiology*, 37:153–156, 2001.

Jeffrey Whitmore and Stanley Fisher. Speech during sustained operations. *Speech Comm.*, 20(3-4):55–70, November 1996.

Carl E. Williams and Kenneth N. Stevens. Emotions and speech: some acoustical correlates. *J. of the Acoustical Soc. of Am.*, 52(4B):1238–1250, Oct. 1972.

Brian D. Womack and John H. L. Hansen. Improved speech recognition via speaker stress directed classification. In *Acoustics, Speech, and Sig. Proc., IEEE Int. Conf. on*, pages 53–57, 1996.

VITA

Keith William Godin was born in Minneapolis, Minnesota, U.S.A. on August 2nd, 1985, to William Keith Godin and Cynthia Lynne Godin. He attended St. Albert the Great and St. Mark's grade schools in Minneapolis and St. Paul, and De La Salle High School in Minneapolis. He was awarded the degree of Bachelor of Science in Computer Engineering in May of 2007 by Rose-Hulman Institute of Technology in Terre Haute, Indiana. He is the author of "Analysis and Perception of Speech Under Physical Task Stress," published at *Interspeech* in Brisbane, Australia, in September 2008.