Speech Analysis Methodologies towards Unobtrusive Mental Health Monitoring

by

Keng-hao Chang

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor John F. Canny, Chair
Professor Nelson Morgan
Professor Allison Harvey

Spring 2012

Speech Analysis Methodologies towards Unobtrusive Mental Health Monitoring

Abstract

Speech Analysis Methodologies towards Unobtrusive Mental Health Monitoring

by

Keng-hao Chang

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor John F. Canny, Chair

The human voice encodes a wealth of information about emotion, mood and mental states. With the advent of pervasively available speech collection methods (e.g., mobile phones) and the low-computation costs of speech analysis, it suggests that non-invasive, relatively reliable, and modestly inexpensive platforms are available for mass and long-term deployment of a mental health monitor. In the thesis, I describe my investigation pathway on speech analysis to measure a variety of mental states, including affect and those triggered by psychological stress and sleep deprivation.

This work has contributions in many folds, and it brings together techniques from several areas, including speech processing, psychology, human-computer interaction, and mobile computing systems. First, I revisited emotion recognition methods by building an affective model with a naturalistic emotional speech dataset, which is consisted of a realistic set of emotion labels for real world applications. Then, leveraging the speech production theory I verified that the glottal vibrational cycles, the source of speech production, are physically affected by psychological states, e.g., mental stress. Finally, I built the AMMON (Affective and Mental health MONitor) library, a low footprint C library designed for widely available phones as an enabler of applications for richer, more appropriate, and more satisfying human-computer interaction and healthcare technologies.

To my dear family and my lovely friends

I'd like to dedicate this thesis to my family, who has been extremely supportive throughout my journey in the Ph.D. program. As an international student studying abroad, I can always sense the warmth sent overseas from my dear parents at Taiwan and the occasional but caring phone calls from my brother studying at UT Austin. I do want to apologize for my infrequent calls back to home as things get busy, but from the bottom of my heart I always know it, I miss the time that we spent together.

As the time progresses in the Ph.D. study, it is not only an intellectual but a psychological challenge. I know it was my lovely friends staying around me, listening to me that gives me the strength to move on, on those sometimes small but occasionally big obstacles. Friends in Taiwan and in the United States, I dedicate this work to you.

Contents

List of Figures

List of Tables

Acknowledgments

I want to thank first all the members of my thesis committee for their invaluable guidance and contribution to this work. My advisor professor John Canny has advised over the years a great intellectual support for the investigation and a significant tolerance for my mistakes. Professor Nelson Morgan has provided insightful feedbacks to the speech rate tracking techniques I built in his speech processing class, and valuable expert knowledge in speech processing while I am writing the thesis. Professor Allison Harvey has openly supported me to collaborate with her students on the sleep deprivation projects. I feel fortunate to have benefited from the advice, supervision, and contributions of the great committee.

I also owe thanks to other collaborators who have provided great insights and devoted significant energy to this research. Graduate student researcher Ellie McGlinchey from the department of Psychology inspired me with her rigorous analytic and experimental skills in the sleep deprivation project. In addition, I want to thank my group member Reza Naima from the department of Bioengineering for unselfishly showing his guru hardware skills in the development of the physiological sensing platform so that I can fulfilled my responsibility as a graduate student researcher. My undergraduate assistants Matthew Chan and Melissa Lim also spent significant time in conducting the studies and labeled the speech dataset, so I'm grateful for their contribution.

In addition, I feel extremely grateful to my former internship mentors for giving me excellent opportunities to work in the top-notch industrial research centers: Thomas Zimmerman from IBM Research Almaden, Jeffrey Hightower from Intel Research Seattle, and Tim Paek from Microsoft Research at Redmond. The experience I learned with the variety of research projects greatly shaped my research interest and opened up my mind. I also appreciate their continuous support in the process of job hunting.

Chapter 1

Introduction

Emotion, mood and mental health are key determinants of quality of life. *Affect* is a term used to cover mood and emotion, and other non-cognitive phenomena such as arousal. Mental health, especially depression, has close ties with emotion and e.g., is often first manifest as persistent negative mood. Affective computing has a variety of applications: computers may adapt based on affect to improve learning [47], work performance [83], and communication [51]. Healthcare technologies can be made more intelligent to help people regulate emotions, manage stress, and avoid mental illness [63]. But capture of affect can be quite challenging, e.g., GSR sensors must be worn in the periphery of the body and primarily capture arousal, heart rate variability primarily captures stress and confounds with physical activity, and facial and body gesture conveys rich emotion but requires a camera pointing at the subject and real-time image analysis. On the other hand, voice is easily captured and has proved to be a surprisingly accurate tool for mental health evaluation, e.g., showing 90% classification accuracy for depression from a few minutes of voice data [59]. Voice analysis for emotion recognition [81] is somewhat less accurate (accuracies 70-80%) but should be usable for many applications. Thus voice seems to be an excellent choice for everyday affect/mental health estimation.

1.1 Affect and Mental Health Monitor

The World Health Organization has reported that four of the ten leading causes of disability in the US and other developed countries are mental disorders. Depression has become a major financial burden for the world's economy; in the US alone, it is estimated to cost as much as $83.1 billion dollars in year 2000. By the year 2020, depressive illnesses are expected to become the second most costly health problem and the leading cause of disability for women and children worldwide [64]. In addition, the social ramifications associated with depression are just as staggering; depression is a major contributor to suicide, which takes about 850,000 lives each year [66].

Fewer than 25% of depression patients are currently receiving the necessary treatment and an even smaller percentage of at-risk groups are getting adequate preventive care [66]. A multitude of barriers exist for depression prevention and treatment, including the lack of trained professionals, the lack of resources for long-term treatment and monitoring, and often, the social stigma associated with mental disorders. Identifying effective ways for early detection of warning signs, persuading at-risk groups to seek for help, and establishing long-term monitoring plans (to avoid relapse of depression) are major areas for improvement to move towards a more holistic approach for treating depression related mental disorders.

Were one to design an ideal device for affect/mental health monitoring by voice, it would probably look a lot like a cell phone. A small, handheld device that is regularly used for other voice-based tasks (i.e., calling others), and which helps to distinguish a particular user's voice from those around them (phones have a variety of noise-canceling and directional features built in). What is lacking for developers are the speech features needed for applications or better still, binary or real values that denote emotion or depression strengths - i.e., emotion classifier outputs.

Automatic recognition of affect in speech is a problem with a very large scope. Solving this problem goes beyond the work of a single doctoral thesis. Therefore, this work is focused on the acoustic, non-content-wise properties of speech to build a statistical model for monitoring affective states. The computation costs of processing the acoustic parameters are significantly lower than extracting and evaluating the semantic content, which is an advantage towards modestly inexpensive platforms for mass and long-term deployment of a mental health monitor.

Scenarios

The value of speech monitoring of mental health may not be immediately obvious so we describe some scenarios, illustrated in Figure 1.1:

- Cell-phone monitoring of healthy subjects as part of a health-care package. Using client code on the phone itself, the voice is analyzed during cell phone conversations. Privacy is maximized in this way, and subjects are directly informed if problems emerge. In the early stages, they are likely to seek treatment in many cases. Rather than individual sign-up, this would be part of a health "package" from a provider which includes physical health as well. This de-stigmatized mental health, and presents it as part of comprehensive health care.

- Subsidized "calling-card" number for at-risk populations. Subjects can make free calls on a normal phone using a special access number. This number routes calls through a cloud of servers where they are analyzed. Distinct access codes would allow per-patient tracking. This method should be cost effective for many chronic conditions such as AIDS where mentally ill patients add severe cost overhead due to wasted (not taken) medication and (corollary) drug-resistance strains of the virus.

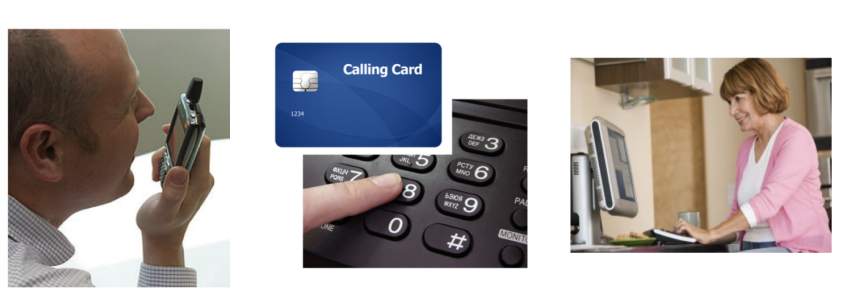


Figure 1.1: Scenarios for Speech Monitoring

- Monitoring of human-computer speech interfaces and interpersonal speech for elders in assisted or independent care. Environmental monitors (array microphones that work 10-20 feet from subjects) can be used to gather incidental speech between subjects. For subjects living alone, speech interfaces may be introduced as a convenient way for subjects to access email, text messages, news, weather and personal schedule information. These routine services provide regular opportunities to intercept and analyze their speech for mental health purposes.

Motivating Applications

In this section we describe several applications that an affect monitor should be able to support (application mockups are available in Appendix A).

- Improving emotional intelligence. This application monitors the user's emotion continuously in order to improve the user's ability to identify, assess, and control their emotions. Even if users are good at assessing emotions over the short term, this application would allow visualization of frequency and intensity of emotions over the long term to expose trends in mood. By integrating contextual information like the user's calendar and location, the application can correlate emotions with possible triggers and allow the user to better manage those effects.

- Managing social relationships. This application would measure emotions and detect positive affect or conflicts during phone conversations. While users are generally aware of their emotions during a conversation, they are also cognitively loaded with the subject matter of the conversation. They may also fall without realizing into counterproductive roles (e.g., mutual victim roles in close relationships) which induce a variety of negative emotions (frustration, defensiveness, anger) that are incorrectly attributed to the partner in the conversation. Emotion monitoring can help users better understand what they were actually feeling and expressing during a conversation with another.

- Computer-assisted psychotherapy. Almost all psychotherapies attempt to track patient's mental state in between therapy sessions. This includes mood, triggers to emotion (the first bullet above), and direct cues to mental health. In conventional therapy this is limited to patient self-reports, which are often irregular and subject to a variety of biases. Monitoring of phone conversations should provide a more fine-grained and diverse sample.

1.2 Thesis Outline

In the thesis, I describe my investigation pathway on speech analysis methods to measure a variety of mental states, including affect and those triggered by psychological stress and sleep deprivation.

The thesis is organized as follows. The thesis starts by providing the theoretical background for speech analysis methods and associated psychology studies, serving as the foundation for the rest of the thesis. This includes the emotion theory, speech production theory, and several relevant work indicating that speech features are useful to measure affective states and mental illness. Chapter 3 describes the process of building speech analysis methods for emotion recognition, using a naturalistic emotional speech dataset with a variety and realistic set of emotion labels. Chapter 4 presents the a light-weight method to recognize phonemes. The method was applied for the recognition of speech rate (rate of phoneme transitions). The method is useful in a sense that the speech rate is an important feature to associate different mental states (e.g., fast speech in anger and slow speech in sadness). Chapter 5 describes the source of speech production and its applications in two domains. We believe that glottal activities in speech production can be a good indicator of physical change induced by mental states. It demonstrates the hypothesis by showing that the features depicting the glottal vibrational cycles are effective in improving the classification of speech in pathology and mental stress, where mental stress often manifests physical response in the autonomic nervous system. Acting on the glottis is a muscle that is activated entirely by the ANS (Autonomous Nervous System), this muscle responds directly to stress. Moreover, Chapter 6 extends this idea by describing a thorough analysis on two applications in which the speech features are effective to reflect the trigger of the physical body. Firstly it presents a project where the impact of sleep deprivation on vocal expression of emotion was investigated. Results for the computerized acoustic properties indicate decreases in pitch, intensity in certain high frequency bands and vocal sharpness. Secondly, a speech dataset under simulated and actual stress will be revisited, and it describes several critical techniques to improve the classification of stress, including user normalization and the constraints of test speech length. Chapter 7 describes the AMMON (Affective and Mental-health MONitor) library, a low footprint C library designed for widely available phones. To comfortably run the library on feature phones (the most widely-used class of phones today), we implemented the routines in fixed-point arithmetic, and minimized computational and memory footprint by algorithmic improvement and code optimization. Finally the thesis draws conclusion and

describes potential work in the future.

Chapter 2

Theoretical Foundation and Background

2.1 Recognition of Affect

What makes our conversation with the others more human-like is the affect we associate within. Affective computing is the study and development of systems that can recognize, process, and simulate human emotions [68]. In the vein of affect recognition, psychologists, computer scientists, and bioengineers have been exploring this in different aspects, be the construction of cognitive models of affect, or the measurement of the bodily expressions.

The Model of Affect

The section does not intent to be an overview of the vast literature on emotion theory. Its goal is to present a simplified and intuitive distinctions to the mental states that my study is referring to. Justin and Scherer suggests using affect as a general, umbrella term that subsumes a variety of phenomena such as emotion, stress, mood, interpersonal stance, and affective personality traits [43]. All of the states share a special affective quality that sets them apart from "neutral" states. We often use the phrase to describe that a person is "affected" by something (e.g., an event, a thought, a social relationship etc), which in fact defines clearly the root of the word "affect". The influence by something gives an affective episode standing out from the neutral baseline states both in the subjective experience of the person and in the perception of the person by an observer.

Scherer [75] differentiated affective states by seven dimensions, including intensity, duration, event focus, rapidity of change, etc. This is called a design-feature approach. This approach suggests three broad classes of affective states:

1. emotions and stress

2. moods and interpersonal stances

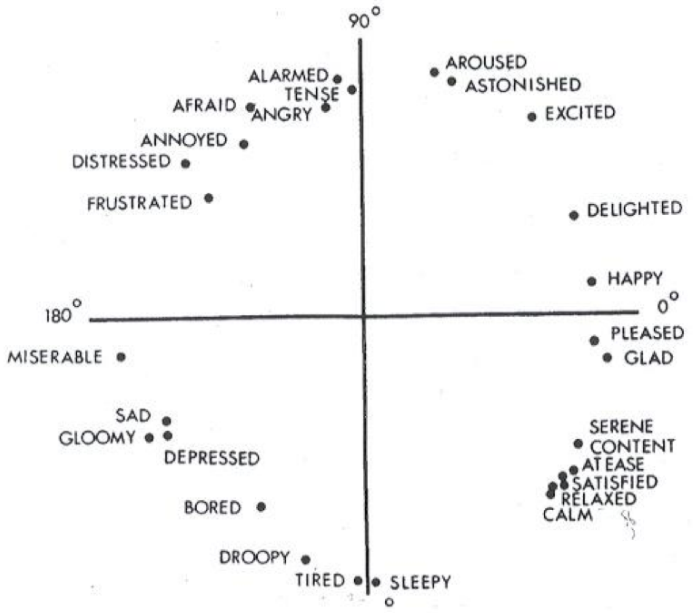


Figure 2.1: Arousal-valence theory with discrete emotions. Arousal increases vertically, valence is positive to the right and negative to the left.

3. preferences/attitudes and affect dispositions (e.g., personality)

Using the dimensions to describe, emotions and stress are quite short but intense reactions to specific events of high pertinence to the individual. Evolutionarily, it is believed that stress played a role in survival by increasing arousal and through activation of the fight-or-flight response in the presence of danger [7]. These reactions are generally powerful and have strong impact on behaviors, which enables studies in the following section using the strong behavioral/expressional cues to sense the affective state of an individual as we can perceive as an observer.

Moods and interpersonal stances are rarely generated by specific events or objects. Moods may occur for many different reasons, often unknown to the individual, triggered by factors such as fatigue, hormonal influences, or even the weather. Along with interpersonal stances, these states may last for hours or days and change only slowly, and the intensity is low. Lastly, preferences/attitudes are long-term affective evaluations of objects or people that have low intensity. Probably due to the nature of low intensity, there are relatively less literature leveraging physiological arousal to sense these affective states, but growing recently [41].

That is the basis for distinguishing affective states, at least from the bodily expression point of view. Looking a step further at the work of recognizing *emotions*, two approach of differentiating emotions were adapted and each of them has their proponents.

- A **discrete emotions approach** argues that one should distinguish among a limited number of "basic emotions" (e.g., angrer, happiness etc) [20].
- A **dimentional approach** defines the emotions as points in a two-dimensional space formed by *valence* (pleasant and unpleasnt) and *activation* (aroused/sleepy) [72]. This approach involves traditional expectations about arousals might affect the bodily expressions.

Figure 2.1 shows the relationship of the two models. One thing to note is that chapter 3 will leverage both approaches to present my investigation in speech analysis methods.

Vocal Expression of Affect

Studies have shown that facial expression [44], speech [43], and biosignals [39] including galvanic skin response, heart rate, breathe rate, and brain activity are strong indicators for inferencing the emotional states.

The branch of emotional speech processing recognizes the user's emotional state by analyzing speech patterns [43, 28]. To be clear there are linguistic *content*- and *prosody*-based aspects of speech that we can leverage to perceive others' emotional states. A intuitive distinction is that the content is about the words in speech, i.e., *what we are saying*, whereas the prosody is about the sound characteristics of speech, i.e., *how we say it*. In terms of acoustics, the prosodics of oral languages involve variation in syllable length, loudness, pitch, and the formant frequencies of speech sounds. Further distinction in speech can be made upon the linguistic, paralinguistic, and extralinguistic information within speech utterances. However, the description is beyond the scope of this thesis, but can be referenced in the linguistic literature, although some distinction is still in debate. Research using the speech content to distinguish emotions is known as "sentiment analysis", which works by counting frequency of sentimental terms (e.g., I am 'excited'). A relevant and currently very popular topic is to analyze large scale web data for consumer behavioral analysis [67].

However, this thesis is geared towards the speech analysis on the "non-content" acoustic parameters to model affect and emotions. Previous studies directed their effort to identify the optimal set of acoustic features to classify a set of emotions. These works were often based on psychological studies that, some prosodic features such as pitch variations and speech rates associate well with emotional changes [43]. In linguistics, prosody is the rhythm, stress, and intonation of speech. Prosody reflects whether an utterance is a statement, a question, or a command. In the mean time, prosody is also applied to reflect the emotional state of a speaker, either consciously or unconsciously.

First of all, *pitch* (i.e., physiological pitch[1], F0, fundamental frequency) represents the rate at which vocal folds open and close across glottis. A sudden increase in pitch can often

[1]Strictly speaking, pitch is a psychological quantity, while F0 is physiological. For instance, pitch can change from the frequency F0 depending on the amplitude of the signal. Furthermore, the fundamental can be entirely missing (as it often is on telephones) and still the listener will perceive the same pitch. So we distinguish the pitch here as "physiological pitch" to represent F0.

be perceived as high activation (e.g., anger), whereas low variance of pitch is often conceived as low energy (e.g., sadness). *Intensity* reflects the effort to produce speech. Studies showed that *angry* utterances usually display rapid rise of energy, and on the contrary *sad* speech usually has characteristics of low intensity. In temporal aspects, speech rate and voice activity (i.e., pauses) are also affected by emotions. For example, sadness often results in slower speech and more pauses.

In addition to discrete emotions, a number of studies have obtained results regarding to affect dimensions, i.e., activation and valence. Activation has been studied the most, and the results are fairly consistent. High activation is associated with high mean F0, large F0 variability, fast speech rate, short pauses, increased voice intensity, and increased high-frequency intensity [43]. The results for valence are much more inconsistent unfortunately.

A representative landmark for emotion recognition was the Interspeech Emotion Challenge 2009 [78]. This challenge included a standard dataset of emotion-tagged speech, and a "baseline" implementation of feature analysis, known as openSMILE. Surprisingly, while some more sophisticated algorithms improved on the baseline system, the improvements were very small, and it is fair to say that the baseline implementations achieved state-of-the-art performance. A second surprising result was that the use of segmental features (phone-level features) did not improve on "suprasegmental" primitive features (MFCCs, pitch, dynamics, energy). This may change in the future, but for now it means that state-of-the-art emotion recognition is much simpler than phonetic analysis. Expressed in terms of speech recognition components, that means that fully-accurate emotion analysis requires only the front-end of a speech recognizer and not the (memory and compute-intensive) acoustic model or later stages.

As a quick reference, the state-of-the-art recognition accuracy is about 70% for five-way classification of emotions (happy, sad, fear, anger and neutral) in a standard database with actors expressing emotion portrayals [70]. On the other hand, for the Interspeech challenge, naturalistic transcripts were recorded and hand annotated. Accuracy was only 70% for two-way classification [79]. An important question concerns the extent to which such portrayals differ from natural vocal expressions. However, a preliminary view can be offered that acted emotional speech may be more exaggerated than natural vocal expressions so that it allows higher recognition accuracies.

2.2 Diagnostic Cues in Vocal Expression

Mental illness is one of the most undertreated health problems worldwide. Previous work has shown that there are remarkably strong cues to mental illness in short samples of the voice. Mounting evidence from the literature suggests a critical role for speech in the clinical aspect of affective states. The section gathers research suggesting the critical role of vocal expression for standardized diagnosis, emerging research of psychopathological signs, and common practice using mental status exam.

Table 2.1: DSM-IV: Diagnosing Criteria for Major Depressive Disorder (for a minimum of two week duration)

Major Criteria	Plus 4 or More
Depressed mood	Feelings of worthless or guilt
Loss of interest	Impaired concentration
	Loss of energy or fatigue
	Thoughts of suicide
	Loss or increase in appetite
	Insomnia or hypersomnia
	Retardation or agitation

Psychomotor Symptoms

A growing body of scientific research points towards psychomotor disturbances as consistent indicator (also known as prodrome [42]) of the onset of depression [76, 85]. It was also reported that psychiatrists routinely monitor these prodromes in patients during the diagnostic period and as measures for assessing treatment progress. For example, depressed patients often express slow responses (longer response time to questions and pause time within sentences), monotonic phrases (less fundamental frequency variability), and poor articulation (slower rate of diphthong production) [29, 48, 65, 87]. These factors indicate signs of retardation. On the other hand, the spectrum of agitated behavior includes expansive gesturing, pacing and hair twirling [89]. Lemke and Hesse [49] stressed the importance of developing a monitor of psychomotor symptoms. Moreover, they stated that the technology should not be constrained to research purposes only: "the development of clinical instruments for evaluation of motor symptoms in psychiatric patients is necessary to differentiate more clearly between observed psychopathological signs and experienced symptoms in clinical psychiatry."

Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR)

Persistent depressed mood is the main criteria for diagnosing the existence of major depressive disorder (DSM-IV-TR) [1]. Coupled with lost of interest and four additional criteria for a minimum of 2-week duration, psychiatrists may diagnose existence of the depressive illness (Table 2.1). Note that sleep deprivation (insomnia) and stress (agitation) are criteria relevant to depression and the recognition of these mental states are explored in the thesis.

The Diagnostic and Statistical Manual of Mental Disorders (DSM) published by the American Psychiatric Association provides a common language and standard criteria for the

Table 2.2: Speech Descriptors in Mental Status Exam

Category	Patterns
Rate of speech	slow, rapid
Flow of speech	hesitant, long pauses, stuttering
Intensity of speech	loud, soft
Clarity	clear, slurred
Liveliness	pressured, monotonous, explosive
Quantity	verbose, scant

classification of mental disorders. It is used in the United States and in varying degrees around the world, by clinicians, researchers, psychiatric drug regulation agencies, etc. Current version is DSM-IV-TR (fourth edition, text revision). It is worthwhile noting that the current standard does not involve any criterium described with speech (Table 2.1). Nonetheless, the human speech is the expression of our bodies and minds. It is the focus of the thesis to bridge the vocal expression with the criteria stated in the standard. This goal is similar to that of actiwatch, an accelerometer-enabled watch to measure gross motor activity.

Mental Status Examination

The mental status examination in the United States or mental state examination in the rest of the world, abbreviated MSE, is an important part of the clinical assessment process in psychiatric practice. It is a structured way of observing and describing a patient's current state of mind, under the domains of appearance, attitude, behavior, mood and affect, speech, thought process, thought content, perception, cognition, insight and judgment [88]. In particular, practitioners pay special attention to abnormal speaking styles listed in the MSE (Table 2.2) and relate the descriptors to certain mental status. The MSE allows the clinician to make an accurate diagnosis and formulation, which are required for coherent treatment planning.

Relationship with Emotion Research

Results derived from affective computing share a similar set of acoustic features in psychopathology research, including pitch, intensity, speech rate etc (i.e., prosodic features). The methodology to extract acoustic features has been studied more extensively in emotion recognition research than in the mental illness setting, so acquaintance in the emotion research helps us to proceed in the clinical mental health setting.

2.3 Theory of Speech Production

The thesis focuses on the meaningful parts of vocal expression for mental health monitoring, so understanding the process of speech production helps develop effective speech analysis routines. The following description is adapted from [26, 43]. The basis of all sound making with the human vocal apparatus is air flowing through the vocal tract powered by respiration. The type of sound produced depends on whether the air flow is set into vibration by rapid opening and closing of the glottis, producing quasi-periodic *voiced sounds*, or whether it passed freely through the lower part of the vocal tract and is transformed to turbulent noise by friction at some obstruction (e.g., the lips), i.e., non-periodic, *unvoiced sounds*. The characteristics of a speech wave form (and of its spectrum) are determined by two quite different and largely independent factors: the glottal wave or pulse (determined by the subglottal pressured and the laryngeal setting) and the vocal tract resonance characteristics (transfer of filter function, mainly determined by the supralaryngeal articulatory setting). The process is shown in Figure 2.2, which illustrates the *source-filter theory* of speech production [26]. The human speech production system is illustrated in Figure 2.3.

The Glottal Wave

At the beginning, the vocal folds are set into a closed position by the muscular action of the laryngeal muscles. The continuous respiratory air flow compresses the air in the column below the glottis and builds up subglottal pressure. When the pressure exceeds the closing force of the muscles, the vocal folds open for a fraction of a second to release some of the pressure. The reclosing of the vocal cords is achieved by the elastic recoli of the folds themselves. Both the overall tension of the vocal folds are regulated by a large number of extra- and intralaryngeal muscles (laryngeal setting). The most important factors are the length, thickness, mass and tension of the vocal folds. The greater the length and the tension, the faster they will open and close. Both F0 and voice quality (e.g., breathiness, roughness, sharpness) are strongly influenced by the timing of the glottal cycle (e.g., the relative duration of closing, closed, opening and open phases).

The Vocal Tract

As a result of the glottal pulse's passage through the transfer function of the vocal tract, some of the harmonics in the spectrum of the pulse are amplified (producing local energy maxima called *formants*) and attenuated. Both effects depend on the resonance characteristics of the vocal tract. Figure 2.2.a-c show the result of this filtering process in the time domain and wave forms in Figure 2.2.d-f, its equivalent in the frequency domain. Radiating at the mouth of a speaker, the waveform serves as the basis for the objective measurement of acoustic parameters.

Basic speech acoustics include simple parameters of wave forms including amplitude and frequency, and complex characteristics such as spectral decomposition, fundamental fre-

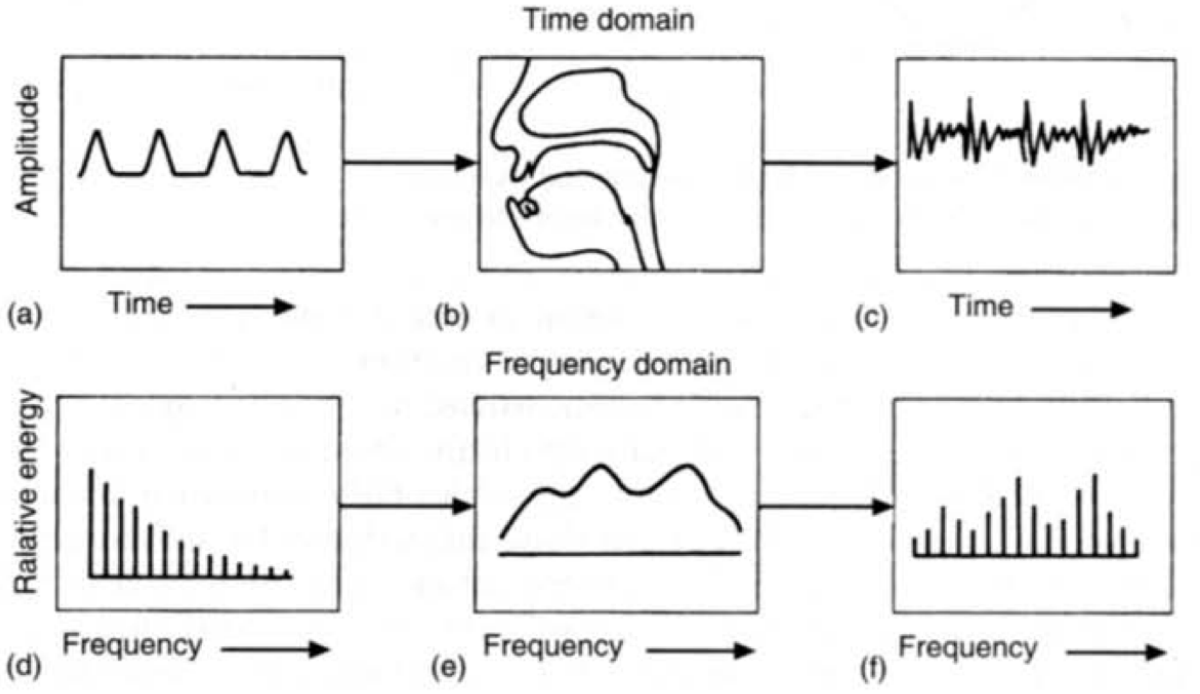


Figure 2.2: The source-filter theory of speech production: (a) glottal wave, (b) vocal tract shape, (c) radiated sound wave, (d) glottal spectrum, (d) vocal tract transfer function, (f) acoustic spectrum at mouth opening (adapted from [26])

.

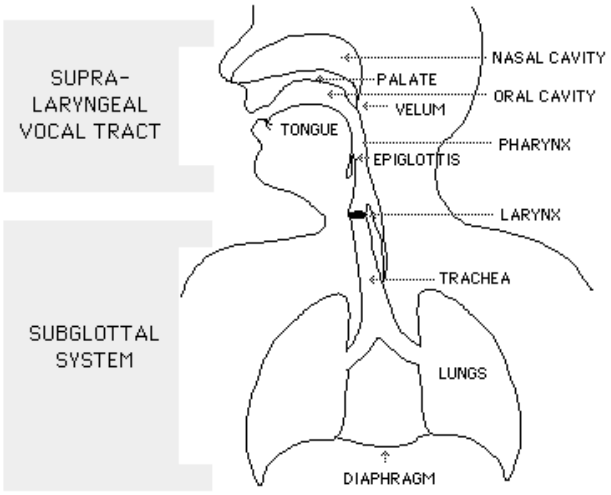


Figure 2.3: The human speech production system

quency (the lowest harmonic that would correspond to the series of harmonics associated with a periodic source) and harmonics (higher-order components of complex waves occurring at integral multiples of the fundamental frequency). The reader is referred to classic textbooks in the field (e.g., [4]).

2.4 Long-term Monitor and Healthcare Applications

Morris conducted ethnographical studies to understand the acceptance of technologies for early detection of health conditions [62]. They suggested that adoption of such health technologies will be increased if monitoring is woven into preventive and compensatory (i.e., intervention feedback) health applications. An integrated system should provide values beyond assessment. Some related work followed this idea. An example UbiFit [17] utilized a wearable multi-sensor device to infer physical activities and created a graphical application to promote physical health. Moreover, an abundance of research was dedicated to address healthcare problems from different perspectives. Abaris [45] supports therapy for children with autism. Ramachandran [71] applied information and communication technologies (ITCs) to support health workers for improving maternal health in rural India.

Chapter 3

Emotion Recognition

We built an automatic emotion monitor via voice, which will serve as a key component to promote emotional awareness. The work was based on the Belfast Naturalistic Emotion Database, which was chosen to provide a realistic model for monitoring everyday conversations. While the database offered a holistic solution to approach "naturalistic-ness", including a variety of emotion labels, a flexible labeling scheme, and a complete set of natural recordings, we discovered that some points need to be considered during the modeling process, in order to fully benefit from those great characteristics. For example, we experimented with methods for feature normalization and proposed methods to address data imbalance and shared-meaning emotion labels. In this work, we achieved 0.7 unweighted average (UA) F-measure in a two-way classification task and 0.4 UA F-measure in a four-way classification task. Here we discuss the experimental result and the learned lessons of a generalized model, which was designed to simultaneously classify a multitude of emotion labels. In the end, we report the useful features of vocal expression.

3.1 Introduction

Persistently depressed mood is one of the major criteria for the existence of major depressive disorder [1]. However, the depressed often lose their ability to recognize and manage the onset of harmful emotions [6], so they often fail to participate the control of the illness. Affective computing research [81][43] has proved that vocal expression is often encoded with a wealth of information, which is suitable to automatically *infer the emotion* that a user is currently experiencing and *prompt the user* of the recent onset of emotions. With strategic feedback, it is anticipated to help users build their ability to retrospect the momentary emotions and develop coping skills. In this chapter, we focus on building an automatic monitor, which can analyze users' vocal expression and then identify the *appearance of emotions*.

With the goal of building an application that will work in people's everyday lives, we considered it necessary to make use of a naturalistic emotion database. We chose *The Belfast Naturalistic Database* [23], which consists 298 audiovisual clips from 125 speakers, 31 male,

94 female. The database provided a holistic solution to approach *'naturalistic-ness'*. For example, it provided a set of 40 emotion labels, which were much richer than the prototypical emotions, such as sad, happy, and angry emotions. In addition, the clips themselves were descriptive enough in showing the necessary context for researchers to understand the local peak of emotions and the development over time. Moreover, it provided a flexible yet detailed labeling mechanism, where the judges were asked to mark down more than one emotion that they can perceive within a clip, along with a three-point intensity value (weak, medium and intense).

During the course of the data analysis, we found it necessary to be more careful processing the data, in order to fully benefit from the great characteristics. For example, out of the rich set of emotion labels, some emotion labels have shared meaning, such as *happy* and *pleased* emotions, possibly causing judges to mix the use of labels. More seriously, the similar emotions may not have distinctive vocal expression. This creates a modeling challenge since clips sounding similar may be labeled as subtly different emotions and will create confusion when training a model. For this problem, we discussed an "objective" way to merge similar emotions as one to avoid the confusion during modeling, instead of judging which two emotions are similar "subjectively".

It is also difficult to obtain a balanced dataset from a naturalistic database, because there are rarely-appearing emotion labels in a rich label set. This problem often misleads training algorithms to create an incorrect model favoring the majority class [12] while optimizing objectives such as accuracy or alike. Instead of applying the popular solutions such as resampling or downsampling, the just-mentioned method which merges similar emotions into a larger class can also interestingly play a role here. A larger class can partially reduce the magnitude of data imbalance, and therefore help create a better model.

Furthermore, leveraging the Belfast database offers great opportunity to explore several useful mechanisms for creating better emotion recognition models. For instance, we proved the effectiveness of a *user-based feature normalization*. We also found that this method is effective even if there are fewer than five data points per user. In addition, we experimented with a *generalized model* that is capable of simultaneously classifying a multitude of emotion labels. Although in the end the accuracy of the model was not satisfying, we were able to study the feasibility of identifying rare emotions. Moreover, we adopted a *label aggregation* method to combine the labels created by multiple judges, which was made possible by the detailed labeling scheme with a three-point intensity scale.

Combining the efforts mentioned above, we delivered classifiers that were able to distinguish prototypical emotions with a reasonable performance. In short, we achieved 0.7 unweighted average F-measure in a two-way classification task and 0.4 unweighted average F-measure in a four-way classification task. We also discussed the features that were useful in identifying emotions, and showed that voice quality features, pitch accent, and first-order measure of contours were the most outstanding ones.

The rest of the chapter will be presented in the following outline. First, we will provide related work specifically to emotion recognition in Section 3.2. Then, we describe the Belfast database in Section 3.3, which includes data characteristics, a label aggregation method, an

objective way to evaluate the similarity of emotions, and the classification tasks that we plan to investigate. In Section 3.4, we will describe the feature set. Finally, we will provide experimental results in Section 3.5, and draw a conclusion in Section 3.6.

3.2 Related Work

In this section we describe some related work for discussing our contribution to the area.

Constructing an automatic emotion recognizer depends on a sense of what emotion is [19]. Most psychological studies have been trying to provide a holistic framework for the portrayal of the source and expression of emotional states [18]. Nonetheless, for emotion recognition research, it comes naturally to develop a simple scheme that can be applied directly to statistical machine learning models. In the end, assigning categorical labels to speech becomes the most popular one, because it is natural and convenient to reduce the recognition problem into a well-studied classification problem in the machine learning area [81]. However, as humans use a daunting number of labels to describe different emotions, the simple reduction advantage leads to a challenging question: 'could we create a system that can recognize emotions with the same level of granularity people apply'?

For modeling emotions, researchers usually follow the same modeling process. It usually starts by extracting a significant amount of features. Since there is not yet a final conclusion of the most effective feature set [43][52][28], researchers usally take a shut-gun approach where they include as many features as possible if the features may be helpful for recognition. Then, some may apply an intermediate feature selection step to reduce the dimensionality of features [28]. Finally, researchers choose a machine learning model of interest to classify emotions, including support vector machines [21], artificial neural network [91], gaussian mixture models [56], or a combination of these [31]. Again, researchers have not concluded which method is superior. Since classes were often unbalanced, the primary measure of performance was often chosen as unweighted average recall or F-measure. A recent work showed that it achieved in the level of 70% UA recall for two emotion classes and 45% UA recall for five emotion classes (Interspeech 2009 Emotion Challenge [81]).

3.3 The Naturalistic Belfast Emotional Database

We made use of *The Naturalistic Belfast Emotional Database* [23] to model vocal expression of affect. The database consists of 298 audiovisual clips from 125 speakers, 31 male and 94 females. These clips were collected from a variety of television programs and studio-recorded conversations. The television programs consist of chat shows, religious programs, programs tracing individuals' lives and current affairs programs. Studio recordings were based on one-to-one interactions between a researcher with field work experience and close colleagues or friends.

Table 3.1: Categorical labels used in Belfast database

affecting, afraid, agreeable, amused, angry, annoyed, anxious, ashamed, bored, calm, confident, content, despairing, disappointed, disapproving, disgusted, embarrassed, excited, guilty, happy, hopeful, hurt, interested, irritated, jealous, joyful, loving, nervous, panicky, pleased, proud, relaxed, relieved, resentful, sad, satisfied, serene, surprised, sympathetic, and worried.

The database contains conversations covering a wide range of emotional states that occur in everyday interactions as well as more archetypal examples of emotion such as full-blown anger. In addition, for each speaker there is at least one clip showing him or her in a state judged relatively emotional, and also one clip judged relatively neutral. Clips range from 10 to 60 seconds in length. Also it is worth noting that, in addition to the main speaker of interest, other speakers such as the host of TV shows were recorded as well. To avoid confusion while building classifiers, we did some preprocessing to segment out the voice of the other participants.

The Belfast database provides both categorial labels and two-dimensional activation-evaluation space coding [72]. Nonetheless, we only focused on the categorical labeling. The database utilized an intuitive and generalized labeling scheme. First, it accommodates the *coexistence* fact that in our everyday conversations more than one emotion may show up simultaneously, each with different intensity. This is different from that of traditional emotional databases, in particular acted emotional databases, where each clip is coded with (or designed to display) only one categorical emotional label [35]. In addition, it considers the *variety* fact that the emotions appearing in everyday conversations are more than just the basic ones, i.e., happy, sad and angry. The forty different categorical labels allowed us to study how well a computerized algorithm can match the perception of humans.

The labels were labeled by judges. A judge was asked to assign up to three emotion labels that the she can perceive from a clip. In addition, she was asked to describe the intensity level for each emotion, which may be weak (1), medium (2), or strong(3). In this way, a label instance may look like (*Clip*: 001a, *Judge*: bob, *Emotion 1*: sad, *Intensity of Emotion 1*: medium, *Emotion 2*: despairing, *Intensity of Emotion 2*: strong, *Emotion 3*: N/A, *Intensity of Emotion 3*: N/A). The emotions were drawn from a pool of forty labels, and they are displayed in alphabetical order in Table 3.1. In addition, to ensure the quality of labeling, there were a total of seven judges involved.

Label Aggregation

Multiple judges can ensure the reliability of coding, but it requires label aggregation to eliminate labeling error and reflect consensus. The aggregation mechanism that we applied was based on the assumption that the judges were equally reliable.

We first converted the 3-emotion-item tuple-based coding (i.e., *Emotion 1*: sad, *Intensity of Emotion 1*: medium, *Emotion 2*: despairing, etc) to an 40-element array form $Intensity_{c,j}$, where each element $Intensity_{c,j}[e]$ represents the intensity of emotion e perceived by judge j in clip c. If an emotion did not exist in the list, the intensity of its corresponding element in the array should be zero. Otherwise, its intensity recorded in the tuple was copied directly to the array. That means, the intensity values in the array form became $\{0, 1, 2, 3\}$ where 0 represents non-existence, 1 is weak, 2 is medium and 3 is strong. With the array-representation, it became straightforward to aggregate by

$$Intensity_c[e] = round(\frac{1}{Numberof Judges} \sum_j Intensity_{c,j}[e]) \tag{3.1}$$

The four-point intensity scale enabled a more robust way to aggregate labels, instead of the traditional {0,1} labels. An momentary error labeling would likely to be ignored through aggregation. If there was only one or two judges perceived that a certain emotion appeared in a clip, the intensity they assigned would be diluted since there were a total of seven judges. Similarly, if a judge perceived a emotion strongly, it would be reflected more in the aggregated label, than the case if the judge perceived it weakly.

Throughout the work, we focused on predicting the appearance of emotions:

$$Appearance_c[e] = \begin{cases} 1 & if\ Intensity_c[e] = 1, 2, or\ 3 \\ 0 & if\ Intensity_c[e] = 0 \end{cases} \tag{3.2}$$

Distribution of Emotions

With the aggregation method, now we can provide an overview of how emotion labels distribute in the database. Table 3.2 shows the sorted list of emotions according to the number of appearances in clips. There were a total of ten clips did not have labels provided, so the total number of useful clips was actually 288. In addition, the positive and negative emotions are placed in two separate columns.

Table 3.2 shows an evidence of data imbalance, that out of 298 clips, most of the emotions appear in less than 1/5 of the clips, and more than half of the emotions even appear in less than 1/10 of the clips. The table implies that there was significant class imbalance between emotions, which is actually very common in naturalistic emotional databases since it is difficult to balance all these emotions from natural recordings. That said, if one wants to build a classifier to recognize some of the minority emotions, such as the *joyful* (c.f. *happy*) emotion, the class-imbalance problem could lead to improper modeling since a classifier may be optimized to favor the majority class.

Similarity between Emotions

As stated in related work, humans use a significant number of labels to describe emotions, where some of them are subtly similar. The similarity between some of the emotions may

Table 3.2: Appearance Frequency in Clips of Categorical Emotions (Out of 288 Clips)

Negative Emotion	Number of Appearances	Positive Emotion	Number of Appearances
sad	70	happy	66
angry	68	pleased	58
annoyed	64	interes	53
disapproving	39	content	40
hurt	37	confident	31
disappo	26	relaxed	31
worried	18	calm	31
resentful	14	excited	25
disgusted	13	affecti	22
despairing	12	amused	19
irritated	10	loving	17
anxious	10	serene	11
afraid	8	proud	9
nervous	7	surprised	8
guilty	1	joyful	8
ashamed	1	relieved	7
jealous	1	hopeful	6
embarrass	1	satisfied	5
bored	0	sympathetic	3
panicky	0	agreeable	0

create confusion on classifiers, which makes it challenging to have a system that matches the level of granularity people apply to distinguish emotions. In other words, a classifier may suffer from a big deal of confusion in the classification of similar emotions, e.g., *happy* and *pleased* emotions.

As a result, we may want to identify which emotions are beforehand in order to resolve the confusion imposed on the classifiers. However, an immediate question will be raised: "how do we claim that a pair, or a set of emotions are similar, or have shared meaning?" Of course, we can always judge it by ourselves, with our own knowledge and beliefs. But even better, we can approach this in an *objective* way, by leveraging judges' joined perception on the meanings of emotions.

The idea was that, if two emotions appeared (or were labeled) concurrently in clips, it's likely that they showed the same emotional image to the judges. That is, if the co-appearance is frequent, two emotions are likely to be similar. The *similarity* between two emotions $sim(e_1, e_2)$ can be formulated by Pearson's correlation coefficient:

Table 3.3: The Top Correlated Emotion Pairs

Emotion Pairs	Correlation Coefficient
angry, annoyed	0.686
happy, pleased	0.529
angry, disapproving	0.496
sad, hurt	0.484
happy, excited	0.477
annoyed, disapproving	0.471

$$sim(e_1, e_2) = PearsonCoeff(Appearance_M[:, e_1], Appearance_M[:, e_2]) \tag{3.3}$$

where $Appearance_M$ is a matrix where each column is the appearance array $Appearance_c$ for clip c, as defined in Equation 3.2. This idea was also adopted in the context of collaborative filtering [74].

Having the equation applied to the Belfast database, we obtained most correlated pairs of emotions and listed them in Table 3.3. We can see that the prototypical (i.e sad, happy, and angry) emotions have significant correlation with other emotions, for example angry is similar to annoyed, happy is similar to pleased, and angry is similar to disapproving emotion. The analysis gave us a hint to decode the confusion in experimental results and we will se this being applied in Section 3.5.

Classification Tasks

The categorical coding provided by the Belfast database allows us to experiment a variety of classification tasks. Here we summarize the tasks upfront and describe the rationales behind them.

The first one was to find a *generalized model*. We planned to build a model hat is generalized enough to recognize a variety of emotions (i.e., over ten emotions) simultaneously and is capable of predicting the intensity levels. Although this might sound a little far-reaching if we consider state of the art classification accuracy, the challenge didn't stop us from designing such a model. In particular given the fact that the Belfast database provided such a rich set of emotion labels. In the end, we finalized with a hierarchical model which contains multiple sub-models, where $subModel_e$ is responsible for predicting both the appearance and the intensity of a given emotion e. This means that, each $subModel_e$ predicts whether emotion e does not appear ($intensity = 0$), or appears with weak ($intensity = 1$), medium($intensity = 2$) or strong intensity ($intensity = 3$).

Intuitively, we can build a four-class classifier for each $subModel_e$. However, if we simplify the model to predict only the appearance of a given emotion, it is in fact already a challenging *two-way classification task to discriminate an emotion from the remainder* [81]. By the

norm that adding more classes (i.e., representing intensity levels) would usually worsen the performance, it may be wise to have another hierarchy in the sub-model, so that the sub-model first identifies the appearance and then predicts the intensity. Therefore, we continued in this task with two-class classifiers. In addition to the purpose of building a generalized model, the two-way classification task also allowed us to evaluate how well each of the forty emotions can identified well.

The second task was to build *multi-emotion classifiers*. For example, we built a classifier to discriminate *happy, sad, angry, and the remaining* emotions. This task allowed us to compare ourselves with state of the art performance.

Finally, we defined the third task for building an emotional-aware application. Since such application should be able to recognize how often negative emotion appears, we experimented a *positive-negative two-class classifier*.

3.4 Voice Analysis Library: The Feature Set

We created a voice analysis library to extract features that can best describe the vocal expression of affect. As a high level overview, the features were calculated in two stages. The library first extracted several *waveforms* from the speech signals, such as fundamental frequency, energy, etc. They are in fact the same as the low-level descriptors (LLDs) described in Interspeech Emotional Challenge 2009 [78]. Then, the library calculated some *statistical measures* to describe the dynamics of the waveforms, e.g., mean, variance, first-order maximum, etc. Altogether these measures formed the final feature set, which contained a total of approximately 1000 features. We didn't do feature selection as an intermediate step, so the features were fed directly into machine learning algorithms to perform the three classification tasks.

Table 3.4: Voice Features

Waveforms	Measures
Pitch	
F0 contour	Basic statistics (referring to Table 3.5) of zero and first-ordered F0 waveform; proportion of voiced sections
Stylized F0 contour	Proportion of accent and descent; basic statistics of accent and descent slopes; basic statistics of zero and first-ordered level waveform
Intensity	
Energy contour	Basic statistics of zero and first-ordered energy waveform; basic statistics of zero and first-ordered peaks found in energy waveform; EMS (Equation 3.4) and EDS (Equation 3.5)
Temporal aspects	
Speech rate	Basic statistics of zero and first-order speed waveform
Voice activity	Proportion of active and inactive sections

Voice quality	
Glottal waveforms	Basic statistics of timings of opening phase (OP), closing phase (CP), closed phase (C), open phased (O) and total cycle (TC) found in the glottal waveform; basic statistics of timing ratios of closing to opening phase ($rCPOP$), open phase to total cycle ($rOTC$), closed phase to total cycle ($rCTC$), opening to open phase ($rOPO$), and closing to open phase ($rCPO$) in the glottal waveform
Spectrogram	Basic statistics of zero and first-ordered energy waveforms in each Bark frequency band; basic statistics of zero and first-ordered peaks found in energy waveforms in each Bark frequency band; basic statistics of zero and first-ordered cumulative energy waveforms above each of the Bark-based cut-off frequency thresholds; basic statistics of zero and first-ordered cumulative energy waveforms above each of the Bark-based cut-off frequency thresholds

If we read Table 3.4 closely, the waveforms are grouped into several categories, which are *pitch, intensity, temporal aspects, and voice quality.* From now on we present the extraction of them sequentially.

Pitch

First of all, *F0* represents the rate at which vocal folds open and close across glottis. It is a physiological measure. *Pitch* on the other hand, is a psychological quantity. From the computation perspective, an F0 tracking algorithm is measuring the physiological quantiy, or to be fair, the physiological pitch. It describes how a listener perceive a sound. A sudden increase in pitch can often be perceived as high activation, such as *anger*, whereas low variance of pitch is often conceived as low energy, for example, *sadness* [43]. We made use of open source software Praat [3] and Prosogram [55] to extract both pitch waveform and stylized pitch waveform [55]. In particular, the stylized pitch were extracted based on a controlled cognitive study so that a stylized pitch can follow people's perception on pitch accent, descent and level. An example of a stylized pitch contour is showed in Figure 3.1. It is anticipated that making use of the stylized pitch that approaches people's perception closely on pitch, can help mimic people's perception of emotions. From the stylized waveform, the library also calculated a *level*-based stylized waveform by outputing an averaged pitch value from each stylized segment.

Note that, the unvoiced sections of a pitch contour, i.e., where pitch is zero, were first removed so that the voiced parts of the pitch contour were concatenated together as a continuous contour and then fed into the following calculations.

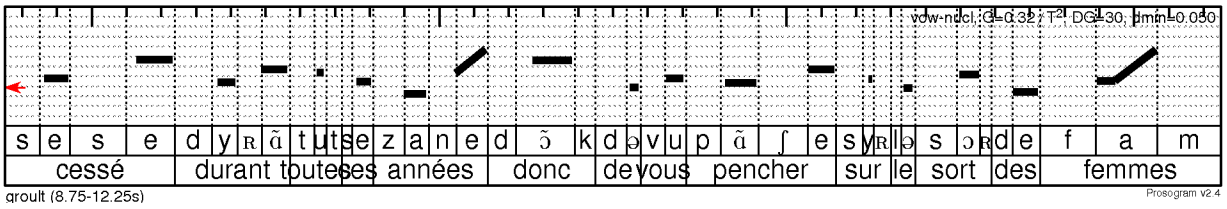


Figure 3.1: A Stylized Pitch Waveform

Table 3.5: Basic Set of Statistical Measures

Standard Measures	Robust Measures
Mean	Median
Variance	Mean absolute deviation
Maximum	95 percentile (95p)
Minimum	5 percentile (5p)
Range	95p - 5p

Statistical Measures

The library calculated measures based on a set of basic statistical measures, which are summarized in Table 3.5. It was our hope that by including multiple statistical measures, the final feature set will represent all possible dynamics which might be affected by emotions. In addition to the standard statistical measures, we also included the robust version of measures, so as to make the features less sensitive to the noise produced during waveform extraction (e.g., median vs. mean, 5 percentile vs. min etc).

In addition, we hoped to model that some emotions may have less momentary change in pitch (i.e., *monotone*), such as sadness. Therefore, we had the library calculating first-order perturbation [69]. First-order perturbation was calculated by taking difference between adjacent samples in zero-order waveform, where the zero-order waveform is simply the original waveform. In this way, it can describe the rapid change of pitch cycles from the current one to the next.

Intensity

The *intensity* reflects the effort to produce speech. Studies showed that *angry* utterance usually displays rapid rise of energy, and on the contrary *sad* speech usually is characterized by low intensity.

Based on the observation, our purpose became creating features that can describe the overall energy level and some momentary energy 'onset' and 'offset'. For the latter, the library first extracted a raw energy waveform by root-mean-square with a moving window.

Then, it identified peaks in the energy waveform, and then it calculated statistical measures to capture the dynamics in-between the peaks. Describing dynamics by peaks was based on the hypothesis that peaks are the major perceptual components in an energy envelope.

Also, Moore II proposed two measures EMS and EDS [59] to accommodate the fact that direct energy values are less robust to different recording conditions (i.e., microphone distance, recording level, etc) varied slightly from session to session. EMS (energy median statistics) and EDS (energy deviation statistics) are energy perturbation measures among the voiced sections of a contour. They were computed by taking all of the statistics listed in Table 3.5 from the voiced sections and then computing the standard deviation (EDS) and the median values (EMS) of the i^{th} statistic by

$$EMS = Median(STAT_i[E_v]), i = 1, ..., N \tag{3.4}$$

$$EDS = Std(STAT_i[E_v]), i = 1, ..., N \tag{3.5}$$

where $STAT_i$, is the i^{th} statistic computed on the voiced sections, E_v one of the V voiced sections broken down from the energy contour, and N is the total number of statistics computed on each voiced section.

Temporal Aspects

In *temporal aspects*, we included measures that can describe speech rate and voice activity (i.e., pauses). Some research showed that those two temporal properties may be affected by emotions [43]. For example, *sadness* often result in slower speech and more pauses.

In terms of implementation, we made use of Morgan's *mrate* implementation [61] to calculate speech rate. In addition, we adopted ETSI's extended front-end processing module ($ETSI\_AFE$) [24] to approximate the amount of pauses. The approximation was achieved by firstly using the front-end to calculate a sequence of voice activity flags. Then, the library collected the total duration of inactive periods as the amount of pauses.

Voice Quality

Studies reported that emotions may influence the *voice quality* of utterances [43]. For example, some voice becomes *sharp* or *jagged* while some voice sounds *soft*.

Glottal waveforms are useful to describe these sound characteristics [86]. As illustrated in Figure 3.2, a glottal (flow) waveform represents the time that the glottis is open (O) (with air flowing between vocal folds), and the time the glottis is closed (C) for each vibrational cycle. In addition, an open phase can be further broken down into opening (OP) and closing (CP) phases. If there is a sudden change in airflow (i.e., shorter open and close phases), it would produce more high frequency and the voice therefore sounds more *jagged*, other than *soft*. To capture it, the library calculated measures describing timings of the phases and the ratios of closing to opening phase ($rCPOP$), open phase to total cycle ($rOTC$), closed phase to total cycle ($rCTC$), opening to open phase ($rOPO$), and closing to open phase ($rCPO$). The extraction of glottal waveforms was based on Moore II's implementation [58].

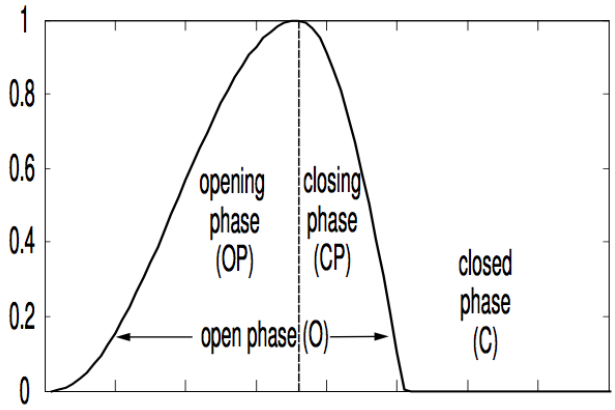


Figure 3.2: A Glottal Vibrational Cycle

We also included a *spectrogram* to describe the energy distribution across frequency bands. The reason was that the emphasis on certain frequency may be speaker dependent and may be used to reflect emotions [28]. In particular, we analyzed the frequency bands in a way that follows the nature of listening. Listening devotes "unequal" emphasis to different areas of the audible spectrum, i.e., critical-band processing. It was our hope that by following the nature of listening, we could better approach human decoding of emotions. In particular, we processed the frequency bands with Bark scales [98], from $Bark1$ to $Bark14$ (0Hz - 2320Hz).

Moreover, some work claimed that as the amount of high-frequency energy increases, the voice sounds *sharp* and less *soft* [43]. Therefore, we analyzed the amount of high-frequency energy in the spectrogram, by calculating the cumulative values in the spectrogram that appear above certain cut-off frequency thresholds. A set of frequency thresholds were chosen, including 510Hz ($> Bark5$), 920Hz($> Bark8$), 1480Hz($> Bark11$) and 2320Hz($> Bark14$).

Feature Normalization

After finalizing the feature set, we considered methods for feature normalization. Based on the fact that each individual may have her own pattern of vocal expression (for instance a male's high pitch is only as high as a female's normal pitch), we need to make sure that the features or patterns of different users are lying in the same range.

In particular, the Belfast database contains clips from over 100 speakers, which is a good dataset for us to experiment a *user-based normalization*. In implementation, we collected the range information for feature $f_c[i]$ on a *user* basis, which are maximum $max_{u(c)}[i]$ and minimum $min_{u(c)}[i]$. The notation c represents the clip from which the feature $f_c[i]$ was extracted, and $u(c)$, a function of c, represents the user (speaker) of clip c. In other words, the maximum and minimum values were gathered from all the clips of the same user. Then,

we applied the range information to normalize each feature:

$$f_c^N[i] = \frac{f_c[i] - min_{u(c)}[i]}{max_{u(c)}[i] - min_{u(c)}[i]} \tag{3.6}$$

In addition to the user-based normalization, we also proposed another coarse-grained normalization method: *gender-based normalization*. The reason was that, there were only about two to five clips from the same speaker, which may be not enough to represent a valid range of a user's feature. In this method, the range information $max_{g(c)}[i]$ and $min_{g(c)}[i]$ will be collected from all the clips from speakers having the same gender as $g(c)$ (gender of clip c), which should be more reliable in terms of data size:

$$f_c^N[i] = \frac{f_c[i] - min_{g(c)}[i]}{max_{g(c)}[i] - min_{g(c)}[i]} \tag{3.7}$$

3.5 Experimental Results

We performed several experiments to examine the tasks proposed in Section 3.3. The experiments were conducted in Matlab environment and Weka toolkit [36], and it worked in the following. We first fed the audio clips into the voice analysis library, to acquire a feature vector for each clip. Then, the feature vectors along with class labels were fed to Weka for classification. The model of choice was *J48 decision tree*, which has several advantages. The model usually performs reasonablely well, because it is capable of ignoring noisy and useless features, and makes no prior assumptions about the data. Finally, we could easily interpret what features work well by referencing the built trees. In the end, we applied 10-fold cross validation to analyze the performance.

Predicting the Appearance of each Emotion

As the first task, we experimented two-class classifiers for sub-models, where we looked at the classification between the appearance (class 1) and non-appearance (class 0) of a given emotion. In other words, it was a two-way classification task classifying an emotion and the remainder. Table 3.6 displays the performance, where the results are sorted by the data size of class 1. The sorted list allows us to see the trend of performance over the levels of class imbalance. For each classifier, we listed the F-measure for class 0 (F_{c0}) and class 1 (F_{c1}), and the weighted and unweighted average F-measures (F_{ua} and F_{wa}). However, the weighted average F-measure and the F-measure of class 0 cannot fully reflect the performance of the sub-models, since the classes were mostly imbalanced. For example, towards the bottom half of the table where the classes get more imbalanced ($ratio > 1/10$), if a classifier blindly predicts class 0 most of the time, it will still have the weighted average F-measure higher than 0.9. Nonetheless, the F-measure for class 1 and the unweighted-average F-measure depict the performance better. They can represent how well a classifier recognizes the minority class,

i.e., *the appearance of an emotion*, and we will use those measures to judge performance from now on.

If we look at Table 3.6 again by focusing on the F_{ua} and F_{c1} measures, we will see that only some emotions (or sub-models) had barely acceptable result. If we choose 0.3 as a threshold for F_{c1}, the following six emotions had F_{c1} larger than it: *sad, angry, annoyed, pleased, interested*, and *content* emotions. This roughly means that for 30 percent of the time that the sub-models can recognize the appearance of the particular six emotions correctly. In addition, out of the six emotions, only the *angry* emotion has F_{c1} over the 0.4 level. And yet, the *sad* emotion, one of the prototypical major emotions, was not even in the list.

In addition, there were about twenty emotions with F_{c1} equal to zero, indicating that it remains a challenge to build a generalized model recognizing a variety of emotions, or to do well in the an-emotion-and-the-remainder two-way classification task. We believed that the class imbalance was a big barrier. For example, some emotions only appeared in less than 15 clips, which led to a poor 1:20 imbalance data ratio. In addition, the challenge may simply be due to the inherent fact that some emotions are subtle in speech, such as *disgusted, proud*, and *despairing* emotions. information of vocal expression, With only the information of vocal expression, it was likely that classifiers would confuse the subtle emotions with others.

In summary, we learned that only the popular and major emotions can be recognized in a reasonable accuracy whereas for some subtle emotions the classifiers didn't work well.

Table 3.6: Two-way Classification Result for each sub-model (class 0: non-appearance, class 1: appearance)
* had F_{c1} greater than 0.3

Emotion	Data Size of Class 0/1	F_{c0}	F_{c1}	F_{wa}	F_{ua}
*sad	218/70	0.803	0.343	0.691	0.573
*angry	220/68	0.827	0.426	0.733	0.626
happy	222/66	0.797	0.227	0.667	0.512
*annoyed	224/64	0.853	0.359	0.743	0.606
*pleased	230/58	0.783	0.362	0.698	0.573
*interested	235/53	0.826	0.340	0.736	0.583
*content	248/40	0.899	0.300	0.816	0.600
disapproving	249/39	0.896	0.128	0.792	0.512
hurt	251/37	0.841	0.189	0.757	0.515
confident	257/31	0.911	0.097	0.823	0.504
relaxed	257/31	0.922	0.258	0.851	0.590
calm	257/31	0.891	0.258	0.823	0.575
disappointed	262/26	0.947	0.154	0.875	0.550
excited	263/25	0.928	0.040	0.851	0.484
affecting	266/22	0.959	0.136	0.896	0.547
amused	269/19	0.929	0.211	0.882	0.570

worried	270/18	0.933	0.056	0.878	0.495
loving	271/17	0.967	0.118	0.917	0.542
resentful	274/14	0.967	0.143	0.927	0.555
disgusted	275/13	0.971	0.000	0.927	0.485
despairing	276/12	0.964	0.000	0.924	0.482
serene	277/11	0.978	0.091	0.944	0.534
irritated	278/10	0.975	0.000	0.941	0.487
anxious	278/10	0.964	0.000	0.931	0.482
proud	279/9	0.996	0.000	0.965	0.498
afraid	280/8	0.993	0.000	0.965	0.497
surprised	280/8	0.989	0.000	0.962	0.495
joyful	280/8	0.993	0.000	0.965	0.497
nervous	281/7	0.996	0.000	0.972	0.498
relieved	281/7	0.993	0.000	0.969	0.497
hopeful	282/6	0.982	0.000	0.962	0.491
satisfied	283/5	1.000	0.000	0.983	0.500
sympathetic	285/3	1.000	0.000	0.990	0.500
guilty	287/1	1.000	0.000	0.997	0.500
ashamed	287/1	1.000	0.000	0.997	0.500
jealous	287/1	1.000	0.000	0.997	0.500
embarrass	287/1	1.000	0.000	0.997	0.500
bored	288/0	0.000	0.000	0.000	0.000
agreeable	288/0	0.000	0.000	0.000	0.000
panicky	288/0	0.000	0.000	0.000	0.000

Performance Improvement with Feature Normalization

We experimented feature normalization methods proposed in Section 3.4. The *user-based featue normalization* method delivered significant improvement over the previous result, but the gender-based normalization did not. The result relieved our previous worry about the data problem in user-based normalization. Data with only 2-5 clips per user were sufficient to deliver effective accuracy improvement with user-based normalization. We summarized the result of user-based normalization in Table 3.7, and we only included the meaningful F_{c1} and F_{ua} measures. To display the magnitude of improvement, we also copied the result from Table 3.6 and listed them in columns named as *'Before Norm.'*, i.e., before normalization. To save space, we didn't list the emotions that have zero F_{c1} in both before-normalization and after-normalization conditions.

A paired t-test showed that there was a significant increase in the F_{ua} scores from before-normalization ($M = 0.4855, SD = 0.1556$) to after-normalization ($M = 0.5016, SD = 0.1556$) conditions; $t(39) = 2.2795, p = 0.0282$. Similarly, there was a significant increase in

the F_{c1} scores from before-normalization ($M = 0.1084, SD = 0.1337$) to after-normalization ($M = 0.1366, SD = 0.1605$) conditions; $t(39) = 2.1694, p = 0.0362$.

Qualitatively, three additional emotions (*happy, hurt,* and *relaxed*) have F_{c1} greater than 0.3 with the feature normalization. Unfortunately, the F_{c1} of the *pleased* emotion dropped below than the threshold. There were a total of three emotions (*sad, angry,* and *annoyed*) having F_{c1} greater than 0.4, and the F_{c1} of *sad* emotion became greater than 0.5.

Note that in the remaining tasks, we applied the normalized feature set to evaluate their performance.

Table 3.7: Feature Normalization and the Improvement in Performance over the values from Table 3.6
* has F-Measure of Class 1 greater than 0.3 after normalization

	F-measure of Class 1		Unweighted Average F-measure	
Emotion	Before Norm.	After Norm.	Before Norm.	After Norm.
*sad	0.350	0.531	0.574	0.685
*angry	0.430	0.459	0.627	0.646
*happy	0.238	0.346	0.512	0.581
*annoyed	0.383	0.426	0.611	0.624
pleased	0.326	0.252	0.566	0.544
*interested	0.321	0.306	0.579	0.582
*content	0.312	0.338	0.603	0.618
disapproving	0.143	0.293	0.512	0.594
*hurt	0.167	0.341	0.512	0.616
confident	0.105	0.127	0.504	0.510
*relaxed	0.271	0.393	0.594	0.664
calm	0.239	0.226	0.570	0.567
disappointed	0.182	0.261	0.557	0.599
excited	0.044	0.150	0.482	0.543
affecting	0.167	0.100	0.555	0.517
amused	0.190	0.143	0.563	0.549
worried	0.054	0.176	0.495	0.562
loving	0.143	0.000	0.549	0.472
resentful	0.160	0.000	0.561	0.471
disgusted	0.000	0.083	0.481	0.521
serene	0.111	0.148	0.541	0.553
irritated	0.000	0.100	0.485	0.534
afraid	0.000	0.267	0.491	0.623

Merging Similar Categorical Emotion Labels

Up until now, the two-class classifiers for prototypical (happy, sad, and angry) emotions achieved F_{ua} in the range of $0.58 \sim 0.68$. This is approaching but still worse than the state-of-the-art accuracy [81]. However, as we have already hinted before, we believed the classification error was due to the inherent fact that some emotions have shared meanings with others in the remainder class.

Given the fact that these correlated emotions may confuse classifiers, we were pondering, "why don't we move the subtly different emotions to the same class as the target emotion, and then to build a better classifier on the *merged* emotion?" In other words, the data that confused the main emotion of interest can also be utilized or *recycled into the main class* to avoid confusion. We believed that this is a reasonable way since it is legitimate and practical to have classifiers discriminating a set of similar emotions (i.e., happy and pleased) from the remainder. Moreover, it helps *reduce class imbalance*, and allow algorithms to build more correct classifiers.

We built a merged version of sub-models for the three stereotypical emotions, and named the merged emotions as *angry+*, *happy+* and *sad+*. This was done by moving clips from class 0 to class 1 if the clips contain the chosen similar emotions. The selection of similar emotions was done by referencing the objective similarity measure proposed in Section 3.3.

Table 3.8 shows F_{c1} significantly increased by 0.1 for *angry+* and *happy+* emotions (i.e $angry : 0.459 \Rightarrow 0.557$ and $happy : 0.346 \Rightarrow 0.484$). Note that since the class distribution has changed, it actually has become a different problem. Making head-to-head comparisons may not be valid. However, the result still proved that merging similar emotions is a possible remedy for the problems of class imbalance and emotion labels with shared-meaning.

Given the performance of the two-class classifiers, it indicated very well that it is indeed challenging to realize such a generalized model. In particular, the two-way an-emotion-and-the-remainder classification task has its intrinsic barriers, such as data imbalance and confusing labels. Nonetheless, it was the experimentation of this task itself guided us to learn those intrinsic problems.

Classifying between Multiple Emotions

We proceeded to the second task of classifying between multiple emotions, in particular the prototypical emotions. In addition, we strategically reported the result in an incremental way, where we started with two-class (i.e., two emotions) classifiers, and then we incrementally reported classifiers with other emotions included. In this way it allowed us to understand which groups of emotions can be distinguished well (and yes there are). Moreover, it offered opportunities to report the subset of features that were effective in classifying prototypical emotions. Note that for the goal of feature evaluation, we purposely ignored the clip instances containing more than one of the target emotions, in order to minimize confusion posed on the classifiers.

Table 3.8: Results of Emotion (Sub-model) Merging: angry+, happy+, and sad+

	Before Merging			After Merging		
Emotion	Data Size of Class 0/1	F_{c1}	F_{ua}	Data Size of Class 0/1	F_{c1}	F_{ua}
angry	220/68	0.459	0.646			
annoyed	–	0.426	0.624	201/87	0.557	0.681
disapproving	–	0.293	0.594			
happy	222/66	0.346	0.581			
pleased	–	0.252	0.544	200/88	0.484	0.619
excited	–	0.150	0.543			
sad	218/70	0.531	0.685			
hurt	–	0.341	0.616	209/79	0.516	0.669
despairing	–	0	0.479			

Table 3.9: Performance of Classifying between Multiple Emotions

Emotion Set	Data Size of Classes	F-measures Classes	F_{ua}
angry, happy	68/66	0.739/0.723	0.731
angry, sad	57/59	0.713/0.718	0.716
happy, sad	65/69	0.617/0.689	0.653
sad, angry, happy	58/57/65	0.574/0.595/0.597	0.589
angry, happy, remainder	68/66/154	0.503/0.311/0.631	0.482
angry, sad, remainder	57/59/161	0.452/0.404/0.665	0.507
happy, sad, remainder	65/69/153	0.176/0.427/0.599	0.401
sad, angry, happy remainder	58/68/65/96	0.317/0.356/0.361/0.389	0.356
angry+, happy+, remainder	86/87/114	0.505/0.457/0.525	0.496
angry+, sad+, remainder	59/51/150	0.528/0.384/0.709	0.540
happy+, sad+, remainder	85/76/124	0.409/0.461/0.543	0.471
sad+, angry+, happy+ remainder	48/58/84/66	0.323/0.45/0.475/0.408	0.414
positive, negative	133/112	0.743/0.679	0.711

Classifying Between Prototypical Emotions

In Table 3.9 we can see that differentiating *angry* emotion from the other emotions can achieve about 0.71-0.73 of F_{ua}, but it obtained only 0.65 of F_{ua} distinguishing the *happy* emotion from the other emotion. The result somehow matched the trend reported in the previous section, that the *happy* emotion was modeled relatively poorly. In addition, we experimented a three-class classifier between the three emotions, and F_{ua} dropped to 0.589.

We also added a *remainder* class to build more realistic classifiers. This was done by adding the remainder clips not containing any of target emotions into an additional class. However, the result dropped significantly, where the 4-class classifier obtained F_{ua} only around 0.36. As what we learned in the previous section, the poor result should also be due to the data imbalance and the fact that, the emotions in the remainder class having shared meaning with the target emotions.

Therefore, we applied the *"emotion merging"* remedy introduced in the previous section, where we labeled the merged emotions as *sad+*, *angry+*, and *happy+*. From Table 3.9 we can see that this approach again helped slightly increase F_{ua} by an average of 0.05. In the end, we concluded that in the four-way classification task (*sad+, angry+, happy+*, and *remainder*), we achieved 0.414 of F_{ua}.

Feature Evaluation

The result in the previous section showed that classifying prototypical emotions could achieved reasonable result, so we decided to report effective features for these classifiers. Tables 3.10 to 3.13 list the features selected by the J48 decision tree algorithm. The tables indicate that the useful features actually spanned across the categories that we described in Section 3.4, except the features in temporal aspect, meaning the included features are useful. In particular, the high-frequency energy extracted from the spectrogram were mostly used. Also, the timing characteristics of the open and closures of the glottis were important. They altogether showed that the voice quality features are essential to building emotion classifiers. Moreover, pitch accent from the stylized pitch were applied frequently, implying that it is crucial to monitor the sudden increase in pitch, In addition, both standard and robust statistical measures were effective. Finally, the first order statistics were used frequently, meaning that the momentary change of the contours displays significant discriminating ability.

Between Positive and Negative Emotions

The final task was to classify between *positive* and *negative* emotions. To achieve that, the clips were assigned to either a positive or negative class based on whether positive or negative emotions appear. The definition of positiveness and negativeness are listed in Table 3.2. The clips containing both positive and negative emotions were ignored, with the same reason that we intended to avoid confusing instances. Table 3.9 shows that the classifier achieved 0.711 of F_{ua}.

3.6 Conclusion

To create a model that recognizes emotions in everyday lives, it becomes essential to adopt a naturalistic emotional database. Such a database contains recordings in everyday lives and a variety of emotion labels. Nonetheless, in the process of analyzing the data, we learned that there was a mismatch between the great characteristics of the database and the adaptability of machine learning methods. In particular, a naturalistic database often provide imbalance data and shared-meaning emotion labels.

With a feature set that contains a wealth of information to describe vocal expression, and a method of user-based feature normalization, we achieved0.7 unweighted average (UA) F-measure in a two-way classification task and 0.4 UA F-measure in a four-way classification task. We also proposed and experimented of a generalized model, which was designed to simultaneously classify a multitude of emotion labels. Although the generalized model did not work, the experiment process guided us to understand the challenging realism of modeling everyday emotions. As a future work, we are planning to improve the modeling in various aspects. For example, since naturalistic recordings may contain multiple emotions in each clip, we are interested to construct a more fine-grained labeling to identify the boundaries between the emotions, and therefore create a more fined-grained classifiers. Also, we plan to label the peaks of emotions within a clip, i.e., trajectory of intensity change of an emotion, so that we can track the development of emotions with temporal based machine learning methods.

As stated in Chapter 2, results derived from affective computing share a similar set of acoustic features in psychopathology research, including pitch, intensity, speech rate etc (i.e., prosodic features). So acquaintance in the emotion research helps us to proceed to the next chapters in the clinical mental health setting.

Table 3.10: Selected Features to Classify *Angry* and *Happy* Emotions

Waveform	Focused Part	Statistical Measure
stylized pitch	accent slope	95p−5p
stylized pitch	first-order level	mean
energy	first-order peaks	min
spectrogram	bark1 first-order peaks	min
spectrogram	bark5 first-order	range
spectrogram	bark6 first-order peaks	max
spectrogram	bark7 first-order peaks	range
spectrogram	bark9	mean
spectrogram	>bark5 first-order peaks	std
spectrogram	>bark8 peaks	med
glottal	rOTC	mean
glottal	rOPO	95p

Table 3.11: Selected Features to Classify *Angry* and *Sad* Emotions

Waveform	Focused Part	Statistical Measure
pitch	raw	mean absolute dev.
energy	EDS	5p
spectrogram	bark1 first-order peaks	mean absolute dev.
spectrogram	bark13	range
spectrogram	>bark5 first-order	std
spectrogram	>bark8 first-order peaks	mean absolute dev.
spectrogram	>bark11	5p
glottal	O	95p−5p

Table 3.12: Selected Features to Classify *Happy* and *Sad* Emotions

Waveform	Focused Part	Statistical Measure
stylized pitch	accent slope	mean absolute dev.
energy	raw	range
spectrogram	bark1	95p−5p
spectrogram	bark1 first-order	95p−5p
spectrogram	bark5 first-order peaks	95p−5p
spectrogram	bark9 peaks	5p
spectrogram	bark13 first-order peaks	range
spectrogram	>bark2	mean
spectrogram	>bark8	med
spectrogram	>bark11 first-order peaks	range
spectrogram	>bark14 first-order	95p−5p
spectrogram	>bark14 first-order peaks	5p

Table 3.13: Selected Features to Classify *Sad*, *Angry* and *Happy* Emotions

Waveform	Focused Part	Statistical Measure
pitch	raw	max
spectrogram	bark1	95p−5p
spectrogram	bark6 first-order peaks	95p−5p
spectrogram	bark8	95p
spectrogram	bark9 first-order	95p−5p
spectrogram	bark9 first-order peaks	max
spectrogram	bark10 first-order	range
spectrogram	bark12	max
spectrogram	bark13	mean
spectrogram	bark13	range
spectrogram	bark13 first-order peaks	mean
spectrogram	bark13 first-order peaks	range
spectrogram	bark13 first-order peaks	95p−5p
spectrogram	>bark5 peaks	std
spectrogram	>bark8 peaks	95p−p
spectrogram	>bark8 first-order peaks	max
spectrogram	>bark11 first-order	max
spectrogram	>bark14 first-order peaks	range
glottal	C	max
glottal	rCPOP	mean
glottal	rCTC	5p
glottal	rOPO	5p

Chapter 4

Phoneme Processing

4.1 Introduction

A phoneme is a basic element of a given language or dialect, which is the smallest segmental unit of sound employed to form meaningful contrasts between utterances [2]. Recognition of phonemes has been adapted to language identification, especially those exploiting the phonology difference between languages [97, 96]. Zissman [96] has shown that language modeling at the phoneme level is effective in identifying the alternations of spoken languages. Arguably the phoneme transition is also useful for recognizing different "speaking styles" (virtually different spoken languages) triggered by a variety of mental states. The chapter presents the a light-weight method to recognize phonemes. The method was applied for the recognition of speech rate. Previously we've described that the speech rate is an important feature to associate different mental states (e.g., fast speech in anger and slow speech in sadness). Counting the rate of phoneme transitions in an automatically generated transcript can be used to accurately approximate speech rates. However, generating phoneme sequences is expensive if we adapt an full-blown speech recognizer. We reduced computation by simplifying the prediction with only acoustic models and Gaussian smoothing filters, while still preserving reasonable speech rate estimation accuracy.

4.2 Estimating Rate of Speech

Speech dysfunction, such as slow, delayed or monotonous speech, are prominent features of patients suffering from severe depression, bipolar disorder or schizophrenia. From the perspective of audio signal processing, analyzing those speech features computationally delivers a mental health monitor. In the second application, we describe phoneme-based methods to measure the rate of speech (ROS), due to the fact that depressed patients often express slow and paused speech. By rate of speech, we mean the rate at which individual speech units are uttered. In this work we choose a "phoneme" as the individual speech unit. Adapting acoustic models trained by Sphinx III developed by CMU, the system predicts a phoneme

sequence for a given speech utterance and then approximates the rate of speech by counting the rate of phoneme transitions. In fact, this method can be easily adapted to estimate pause information by measuring the time periods where there there is no phoneme detected (no voice activity). For the purpose of efficient deployment on cell phones, we simplified the prediction method using acoustic models while still preserving reasonable speech rate estimation accuracy.

Several unsupervised methods exist for estimating rate of speech. One method relied entirely on the output of a speech recognizer and count the frequency of word transitions in the transcript [82]. However, it is arguable that estimating rate of speech does not require a full-blown recognizer which will generate more information than the speech rate itself. That is, it also recognizes the speech content. A speech recognizer consists of both acoustic model and language model. The acoustic model takes Mel-frequency Cepstrum Coefficients (MFCCS, the energy distribution of over the frequency spectrum from 40Hz to 12KHz) as input. Then it estimates the speech content by first mapping these MFCCs to phonemes using Hidden Markov Models. The mapping works by Viterbi alignment algorithm, which a dynamic programming algorithm performed in quadratic time. With the aid of language modeling, some error of the alignment will be corrected by the language model. A language model utilize both a dictionary (word-to-phonemes breakdown) and probabilistic n-gram models to decide whether a sequence of phonemes (i.e., words) are predicted reasonably. For example, if a word "good" was predicted, it is very likely to have "morning" predicted as the next word.

By inspecting the recognition process in an automatic speech recognizer, we believe that using the acoustic model itself should be sufficient to predict phoneme sequence and therefore estimating the rate of speech. As long as we can approximate the rate of phone transitions, i.e., speech rate, it's tolerable to have some erroneous phoneme prediction. In addition, we further simplified the mapping process used in acoustic model by avoiding the the quadratic complexity given by the alignment algorithm. We dropped the usage of "transition matrices" for smoothing in the Hidden Markov Models, and replaced the functionality with Gaussian smoothing filters. It is intuitive that the convolution with filters ($n \approx 10$) should be more efficient than the Viterbi algorithm.

Rather than relying on any component of a speech recognizer, Morgan et al. approached this problem directly on the signal level. Using signal processing methods they measured the variation of the energy envelope in raw speech signals [60] as an approximator for rate of speech. Another related work also computes rate of speech by estimating phone boundaries. However, they predicted the phonemes by means of Multi-layer Perceptron [92], rather than adapting the acoustic model of a recognizer.

The rest of the work presents as follows. It first explains the procedure for creating a simplified acoustic model, along with an matrix-based implementation so the performance can be optimized with linear algebra libraries. Then it describes an initial experimental result of phoneme prediction. With the addition of Gaussian filters, the results of the simplified acoustic model approach the ones by a full-blown speech recognizer.

4.3 Simplified Acoustic Model

Acoustic Model of Automatic Speech Recognizer (ASR)

We made use of an acoustic model trained by the CMU Sphinx-3 recognizer [90] to build the speech rate estimator. Now we briefly describe the structure of the original acoustic model, and we propose the simplified model based on this one in the next section. The acoustic model used in Sphinx-3 adapts an common structure that is universal to other implementation, e.g., HTK toolkit [95].

Sphinx-3 is based on sub-phonetic acoustic models. The basic sound in a language are classified into phonemes or phones. There are roughly 50 phones in English. Phones are refined into context-dependent triphones, i.e., phones occurring given the left and right phonetic contexts. The reason is that the same phone within different contexts can have widely different acoustic manifestations, requiring different acoustic models. Phones are also distinguished according to their position within the word: beginning (b), end (e), internal(i), or single (s).

Each triphone is modeled by a Hidden Markov Model. Typically 3-state HMMs are used, where each state has a statistical model for its underlying acoustic. Each state is modeled as a gaussian mixture. The first, second, and the third state of a HMM respectively represents the contextual phone on the left, the phone of interest, and the contextual phone on the right. However, if we have 50 base phones, with 4 position qualifiers and 3-state HMMs, we end up of a total of $50^3 * 4 * 3$ distinct HMM states. So HMM states are clustered into a much smaller number of groups where each group is called a "senone" (tied state), and all the states mapped into one senone share the same underlying statistical model. The number of senones to be maintained can be predetermined during the training stage. The acoustic feature vector has 39 elements, including 13-element Mel-frequency cepstral coefficients (MFCCs), and their first and second order derivatives. The feature vectors are computed in the rate of 100 vectors/second.

Simplified Acoustic Model Decoding

We claim that without using the full blown acoustic models with HMMs (triphones with position qualifiers), we can still approximate the speech rate. The reason is that the current task is to merely recognize the transitions from one phone to the other. As long as we can approximate the rate of phone transitions, i.e., speech rate, it's tolerable to have some erroneous phoneme prediction.

We propose a method that only computes the emission likelihood from the middle state of all 3 state HMMs, without computing the emission likelihood of the first state (the left phone) and third state (the right phone). In this way, given a sequence of feature vectors, the method predicts a phoneme sequence by maximum likelihood estimation "locally" at each frame. It only calculates the likelihoods using the center Gaussian mixture of each senone, and it chooses the most likely phoneme straightforwardly. It does not calculate the likelihood

Table 4.1: Snippet of a model definition file trained by Sphinx-3

center phone	left phone	right phone	position attr	left state id	base state id	right state id
AA	-	-	-	0	1	2
			⋮			
AA	AA	AA	s	145	155	170
AA	AA	AO	s	145	148	164
AA	AA	AW	s	145	155	170
AA	AA	AXR	s	145	148	165
AA	AA	B	s	146	155	168
AA	AA	D	b	146	151	168
AA	B	B	e	144	149	168
AA	B	B	i	144	149	168
AE	AA	B	b	184	201	232
AE	AA	CH	b	184	202	232
AE	AA	M	b	184	200	224
AE	AA	NG	b	184	200	221
AE	AE	T	b	184	201	232
AE	CH	NG	i	180	207	226

given the left and right mixtures, nor does it accommodate the transition probabilities to predict the most probable sequence "globally" using a lattice approach, i.e., the Viterbi algorithm. Details and formulation follow in the next section.

For example, Table 4.1 shows the snippet of a model definition file trained by Sphinx-3. Each row represents a 3-state HMM, labeled with a center phone of interest, a left phone, a right phone, and the corresponding state id's. Since the states were clustered into senones (tied states), multiple HMMs may share the same senone, so a state id may appear more than once. In this example, our method only computes the emission probabilities from state 1, 155, 148, 151, 149, 201, 202, 200, and 207, which correspond to phoneme AA and AE. The computation is reduced by ignoring the computation of contextual left and right states and the transition probabilities in HMMs. Finally, our method retrieves the phone with the highest probability.

Matrix Multiplication for Likelihood Calculation

A big matrix was pre-computed offline so that during phoneme estimation, the emission probabilities of a feature vector from all senones (states) can be computed online by a single operation of matrix multiplication and several exponential operations. The matrix $\mathcal{M}$ is of size $P * Q$ where P equals 2 * the size of feature vectors ($= d$) + 1 and Q equals the number

of mixture components (= K) * the number of considered senones. Each column corresponds to the parameters of a Gaussian mixture component given a senone. Below describes how the big matrix is constructed.

The Gaussian mixture of each senone was trained with a diagonal covariance matrix, so the likelihood of a feature vector $f = (f_1, f_2, ..., f_d)$ given the parameters of a mixture $\theta = (\mu, \sigma, \pi)$ (i.e., the means, variances and weight of each mixture components c) can be written as:

$$p(f|\theta) = \sum_{c=1}^{K} \pi_c \frac{1}{(2\pi)^{d/2} \prod_{i=1}^{d} \sigma_{c,i}} exp(-\sum_{i=1}^{d} \frac{(f_i - \mu_{c,i})^2}{2\sigma_{c,i}^2}) \tag{4.1}$$

The equation can be decomposed in a way to separate feature values from the multiplying coefficients, which can be pre-computed in order to speed up the likelihood calculation. By taking the logarithm on the likelihood of each mixture component c, we can re-write the equation as

$$\begin{aligned}
p(f|\theta) &= \sum_{c=1}^{K} \exp(\log(p(f|\theta_c))) \\
&= \sum_{c=1}^{K} \exp(\log(\pi_c \frac{1}{(2\pi)^{d/2} \prod_{i=1}^{d} \sigma_{c,i}} exp(-\sum_{i=1}^{d} \frac{(f_i - \mu_{c,i})^2}{2\sigma_{c,i}^2}))) \\
&= \sum_{c=1}^{K} \exp(\log(\pi_c) - \frac{d}{2}\log(2\pi) - \sum_{i=1}^{d} \log(\sigma_{c,i}) - \sum_{i=1}^{d} \frac{(f_i - \mu_{c,i})^2}{2\sigma_{c,i}^2}) \\
&= \sum_{c=1}^{K} \exp(\sum_{i=1}^{d} \underbrace{\frac{-1}{2\sigma_{c,i}^2}}_{len=d} f_i^2 + \sum_{i=1}^{d} \underbrace{\frac{\mu_{c,i}}{\sigma_{c,i}^2}}_{len=d} f_i + \underbrace{(\log(\pi_c) - \frac{d}{2}\log(2\pi) - \sum_{i=1}^{d} \log(\sigma_{c,i}) - \sum_{i=1}^{d} \frac{\mu_{c,i}^2}{2\sigma_{c,i}^2})}_{len=1})
\end{aligned} \tag{4.2}$$

From Equation 4.2, we can see that the coefficients with underbracing markings are used to calculate the inner product with the feature values (including the quadratic f_i^2 and the ordinary f_i term) plus a constant. These values can be pre-computed as a vector and saved into a big matrix $\mathcal{M}$. That is, the form a column (length= $2 * d + 1$) of the big matrix corresponding to a particular mixture component of a Gaussian mixture:

$$[(\frac{-1}{2\sigma_{c,i}^2})_{i=1...d}, (\frac{\mu_{c,i}}{\sigma_{c,i}^2})_{i=1...d}, (\log(\pi_c) - \frac{d}{2}\log(2\pi) - \sum_{i=1}^{d} \log(\sigma_{c,i}) - \sum_{i=1}^{d} \frac{\mu_{c,i}^2}{2\sigma_{c,i}^2})]^T$$

Given the matrix $\mathcal{M}$, we can now compute the emission probability $p(f|\theta_P)$ of feature f by a senone S (a Gaussian mixture with K components) in the following steps.

1. **Expand feature vector:** expand a feature vector $f = (f_1, f_2, ..., f_d)$ by including squared terms and a constant term to $f' = (f_1^2, f_2^2, ..., f_d^2, f_1, f_2, ..., f_d, 1)$. f' is a row vector of size P.

2. **Matrix multiplication**: do $f' * \mathcal{M}$ to calculate the likelihoods of the feature in all mixture components throughout all senones. The output is a row vector of size Q.

3. **Take exponentials and do summation**: for every K elements in the length-Q output vector, take element-wise exponentials and do summation to calculate the likelihood of the feature by each senone (Equation 4.2).

Finally, a phoneme $\hat{\mathcal{P}}$ is chosen at each frame by maximum likelihood criterium. Note there's a multiple-to-one mapping from senones to a phoneme (Table 4.1).

$$\begin{aligned} \hat{\mathcal{S}} &= \arg\max_{\mathcal{S}} p(f|\theta_S) \\ \hat{\mathcal{P}} &= map(\hat{\mathcal{S}}) \end{aligned} \tag{4.3}$$

4.4 Experimental Results

Implementation

We applied the SphinxTrain module to train an acoustic model with the ICSI meeting corpus [61]. The collection includes a total of approximately 72 hours of meetings with naturalistic conversations collected at the International Computer Science Institute in Berkeley during the years 2000-2002. We set the training parameters so that the trained model has about 2000 tied states, each with 32 component gaussian mixtures. Since we were only interested in the states (senones) in the middle of HMMs, only 749 states were considered to compute emission probabilities. The speech estimation routine was implemented with MATLAB. The trained model files are read into and processed in the MATLAB environment, including means, variances and weights of the Gaussian mixtures (senones) and the model definition file (Table 4.1).

For the model we adapted, P equals 2 * 39 + 1 = 79 and Q equals 32 * 749 = 23968. Matrix multiplication requires about 4M FLOPS (floating point operations), which can be done real-time in mobile devices nowadays. Fast matrix multiplication libraries can further reduce the computation overhead.

The ground truth is gathered by processing the transcription and dictionary file provided by the ICSI database. The true speech rate (rate of phoneme) is computed by expanding word-based transcripts into phoneme-based transcripts with the dictionary and then calculating the rate.

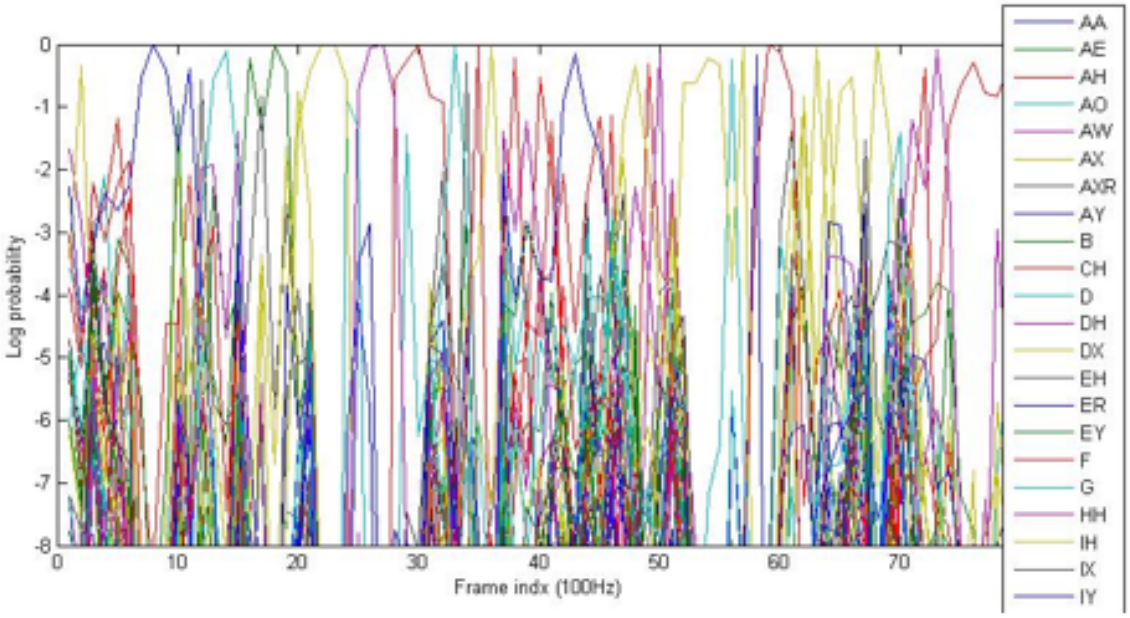


Figure 4.1: The log likelihood trajectories of a speech utterance given 44 phonemes (Gaussian mixtures)

Speech Rate Estimation by Raw Probability Estimation

Given a sequence of feature vectors computed from a speech utterance, the method first computed the most probable phoneme sequence, where the phoneme probability is calculated by taking the maximum of the likelihood estimation among all the senones mapped to the given phoneme. Then, it counted the number of transitions between different phonemes. The initial result was not satisfying because the phoneme rate was overestimated in most cases. Investigation shows that, the raw probability is noisy so there are more erroneous phoneme transitions.

Here we take a test speech utterance as an illustrative example. Figure 4.1 is a chart of the log likelihood predictions of the 44 phonemes. The chart looks noisy in a way that there are many momentary sharp peaks corresponding to unreliable high likelihood of phonemes. Table 4.2 shows the ground truth and the prediction of phoneme sequence for this example. In the row of predicted phone sequence, some of the correctly predicted phonemes are underscored. We can see that the prediction was not persistent and there were many momentary prediction errors. The predicted distinct phone count is 160 but the ground truth is only 34 based on the transcribed phone sequence.

Table 4.2: The prediction of phoneme sequence of a speech utterance

Transcript	TRANSCRIPT ONE FOUR FIVE ONE ONE FOUR SEVEN O
Transcript phoneme sequence	T-R-AE-N-S-K-R-IH-P-T W-AH-N F-AO-R F-AY-V W-AH-N W-AH-N F-AO-R S-EH-V-AX-N OW
Predicted phoneme sequence	M N K D P K T T T JH T Y R R HH AE EH AE AE AX N N N N S S S K K K K K R IX UH IH V P V P F T T T K AH W W AH HH M W W W W L W AA AH AH AH AX N AX N N EH AX AX R V P V F F F F F F AH R AO AO AO AO AO ER ER ER IY R R R R R AX CH V V V CH F F F AO AA AO AY AY AY AY AY AY AY EH AE AXR B UW T T V V SIL AX AH L L L W L L AA AA AA AH AH AH AH AY AH AE IH AX AX L D N DX N N N N M N Y M M D D D P T L K T N JH NG N D V V W W W W W W AA AH UH AH EH EH AX N N N N AA AXR SH N P P F F F F R AO AO AO ER EH ER R R R R R AX AX S S S S S S S S S T DH DH EH EY EH EH EH AH AH T V N N AH AX AX V N N N N N AX EH AW AW AW OW AW AW AW OW OW OW UW OW OW OW OW OW OW OW OW OW OW L L M OW OW D D D L T T T L T T N M M
Predicted distinct phoneme sequence	M N K D P K T JH T Y R HH AE EH AE AX N S K R IX UH IH V P V P F T K AH W AH HH M W L W AA AH AX N AX N EH AX R V P V F AH R AO ER IY R AX CH V CH F AO AA AO AY EH AE AXR B UW T V SIL AX AH L W L AA AH AY AH AE IH AX L D N DX N M N Y M D P T L K T N JH NG N D V W AA AH UH AH EH AX N AA AXR SH N P F R AO ER EH ER R AX S T DH EH EY EH AH T V N AH AX V N AX EH AW OW AW OW UW OW L M OW D L T L T N M

Smoothing by means of Gaussian Filter

We adapted Gaussian Filter to smooth out momentarily erroneous prediction. Gaussian filter is a low-pass filter whose impulse response is a Gaussian function. We designed a 1-D Gaussian filter with variance σ and length $8 * \sigma + 1$, as in Equation 4.4. The Gaussian filter is normalized so that the summation of the coefficients is 1.

$$y = exp(-\frac{x^2}{2\sigma^2}),\ x = -4\sigma, -4\sigma + 1, \cdots, 0, 1, 2, \cdots, 4\sigma \tag{4.4}$$

Using the same example, Figure 4.2 shows a much cleaner confidence measure (smoother and less sharp peaks) than Figure 4.1. σ was set to 2.

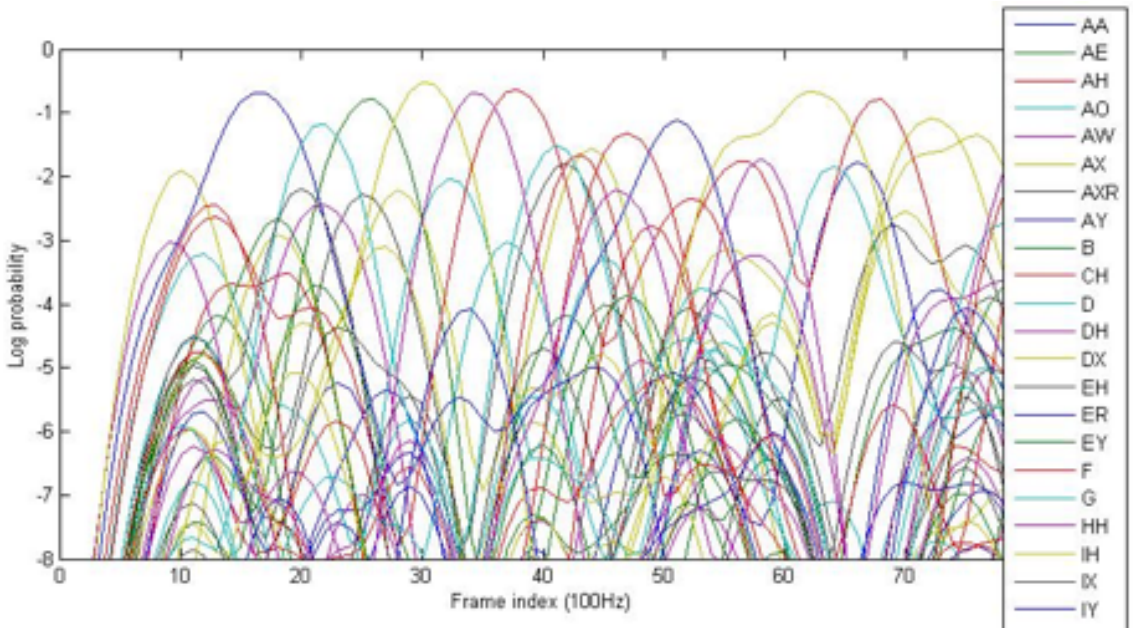


Figure 4.2: The Gaussian-filter smoothed log likelihood trajectories of a speech utterance given 44 phonemes.

Table 4.3 shows the new predicted sequence. The predicted distinct phoneme count reduced to 63 from 160 in the previous case. In the third row of the table below, we can see that the predicted phone sequence became cleaner. It matched with the transcript phone sequence better.

By visual inspection on the log likelihood in Figure 4.2, we suspected that the prediction could be improved further by eliminating the phone predictions with smaller confidence measure. This happens when there are the pauses in between words. So in this case the prediction to a certain phoneme is not necessary.

Thresholding to Eliminate Unconfident Predictions

By inspecting the confidence chart and comparing the likelihood measure of phonemes with the transcript phoneme sequence, we can visually conceptualize a threshold that rules out the incorrect and not-as-confident phoneme prediction. We hypothesized that confidence threshold stay between exp(-1.5) (=0.2231) and exp(-2.0) (=0.1353). In Table 4.4 we show the improved prediction with threshold set to exp(-1.7) (=0.182684). The count of distinct predicted phonemes dropped to 46, in comparison to 63 in the non-thresholded case.

Table 4.3: The prediction of phoneme sequence of a speech utterance with the aid of a Gaussian filter

Transcript	TRANSCRIPT ONE FOUR FIVE ONE ONE FOUR SEVEN O
Transcript phoneme sequence	T-R-AE-N-S-K-R-IH-P-T W-AH-N F-AO-R F-AY-V W-AH-N W-AH-N F-AO-R S-EH-V-AX-N OW
Predicted phoneme sequence	M M M N N N N N N N N T T T T T T T T R R R AE AE AE AE AE N N N N N S S S K K K K K R R UH IH P P P P T T T T T W W W W W W W W W W W W W AH AH AH AH AH N N N N AX AX AX AX V V V F F F F F F F AO AO AO AO AO AO ER ER ER ER R R R R R R V V V V F F F F F AY AY AY AY AY AY AY AY AY AY AXR AXR AXR UW T T V V V L L L L L L L AA AA AH AH AH AH AH AY AY AY AX AX AX AX N N N N N N N N M M M D D D D D D T N N N N NG NG V V V W W W W W W AH AH AH AH AH EH N N N N N N N N P P P F F F F F AO AO AO AO ER ER R R R R R R S S S S S S S S S DH DH DH EH EH EH EH EH EH AH AH AH N N AX AX AX N N N N N N N AW AW AW AW AW AW OW OW OW OW OW OW OW OW OW OW OW OW OW OW OW OW OW OW OW D D D L L L T T T T T T T T M M M M M M M M
Predicted distinct phoneme sequence	M N T R AE N S K R UH IH P T W AH N AX V F AO ER R V F AY AXR UW T V L AA AH AY AX N M D T N NG V W AH EH N P F AO ER R S DH EH AH N AX N AW OW D L T M

Evaluation

We applied the final method to a test set of 1000 speech utterances. Figure 4.3 shows the prediction-versus-ground-truth phoneme rate. A clear correlation displays in the figure.

We ran correlation analysis over the three methods we proposed: prediction with raw likelihood measure, smoothed likelihood measure, and smoothed likelihood measure with a threshold. Table 4.5 shows evolvement of the performance. It shows that the method achieves 0.69 correlation coefficient in the end.

Comparison with ASR

We further performed experiments to compare how well our simplified acoustic model estimate ROS in comparison to a full-blown speech recognizer. Sphinx-3 was used to test ASR performance, with language weight set to 23. Since a speech recognizer only generates

Table 4.4: The prediction of phoneme sequence of a speech utterance with the aid of a Gaussian filter and thresholds

Transcript	TRANSCRIPT ONE FOUR FIVE ONE ONE FOUR SEVEN O
Transcript phoneme sequence	T-R-AE-N-S-K-R-IH-P-T W-AH-N F-AO-R F-AY-V W-AH-N W-AH-N F-AO-R S-EH-V-AX-N OW
Predicted phoneme sequence	_ _ _ _ _ _ _ _ _ _ _ _ _ T T T T T T R R R AE AE AE AE AE N N N N N S S S K K K K K _ _ _ _ _ P P P _ T T T T _ _ W W W W W W W W W W AH AH AH AH AH N N N N AX AX AX _ _ V V F F F F F F F AO AO AO AO AO AO ER ER ER ER R R R R R R V V V V F F F F F AY AY AY AY AY AY AY AY AY AY _ AXR _ _ _ _ _ _ _ _ _ L L L L L L AA AA AH AH AH AH AH AY AY _ AX AX AX AX N N N N N N N _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ V V W W W W W W AH AH AH AH AH EH N N N N N N _ _ _ _ P F F F F F AO AO AO AO ER ER R R R R R R S S S S S S S S S S DH DH DH EH EH EH EH EH EH AH _ _ _ _ _ AX AX N N N N N N N AW AW AW AW AW AW OW OW OW OW OW OW OW OW OW OW OW OW OW OW OW OW OW OW OW _ D D D _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
Predicted distinct phoneme sequence	T R AE N S K P T W AH N AX V F AO ER R V F AY AXR L AA AH AY AX N V W AH EH N P F AO ER R S DH EH AH AX N AW OW D

Table 4.5: The evolvement of performance by Gaussian filters and thresholds

Setup	Mean squared error with linear regression	Std of squared eror with linear regression	correlation
Raw likelihood	11.31	21.10	0.227
Smoothed likelihood ($\sigma = 2$)	10.42	23.26	0.240
Smoothed likelihood ($\sigma = 2$) with threshold (exp(-1.7))	5.49	10.57	0.690

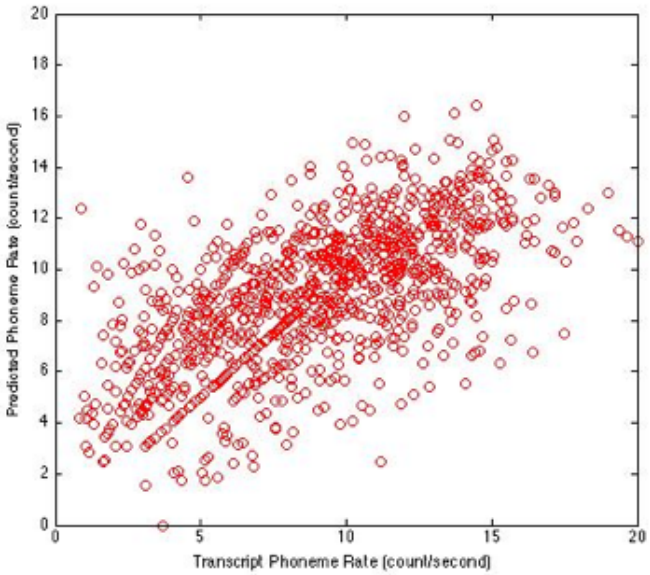


Figure 4.3: The correlation between the predicted speech rate (Y-axis) and the ground truth (X-axis).

Table 4.6: Performance comparison with a full-blown ASR

Setup	Mean squared error with linear regression	Std of squared eror with linear regression	correlation
Smoothed likelihood ($\sigma = 2$) with threshold (exp(-1.7))	5.49	10.57	0.690
ASR	3.80	9.58	0.78

word sequences, we leveraged a lexicon dictionary to expand the words into the phonemes. Table 4.6 shows that ASR has better ROS estimation than our method. Although we didn't measure running time, it's clear that our method runs faster than ASR.

4.5 Discussion

In this project, we showed that using simplified acoustic model without leveraging language models could achieve reasonable rate of speech estimation. Correlation coefficient of 0.69 was achieved using a speech dataset with natural conversation. The method avoids the original HMM decoding method by simply calculating emission probability of Gaussian mixtures. Gaussian smoothing and threshold was used to improve the estimation. For future work, we'd

like to evaluate the method with more dataset. Dataset with emotional speech utterances or collected from patients with mental illness would be a reasonable next step, since the project goal is to create a mental health monitor. In addition, we also want to compare the method with the related work based on the energy envelope (i.e., mrate [60]).

4.6 Conclusion

A phoneme is a basic element of a given language or dialect, which is the smallest segmental unit of sound employed to form meaningful contrasts between utterances. It is also an informative unit of which listeners can pay attention to its variation in order to distinguish abnormalities. The chapter presents a light-weight phoneme recognizer to estimate the rate of speech, by simplifying the utility of acoustic models for predicting the most likely phoneme sequence. The method is light-weight enough to be predict phoneme sequences on mobile phones comfortably, while maintaing certain level of accuracy.

Chapter 5

Voice Source Processing

The theory of speech production says that, the human speech is initiated by the opening and closing actions of vocal folds, which will generate a series of glottal vibrational cycles. By the resonance of the vocal tract and the shape of the mouth, the periodic pulses are re-shaped to the speech we perceive in the end, containing distinct vowels, consonants etc. The chapter focuses on the voice source, and shows that it can serve as a new dimension of evidence to monitor mental health. The hypothesis is that *the mental affects the physical*: mental states may physically and unconsciously affect the muscles of glottis, and alternates the shape of glottal vibrational cycles. The alternation is the focus of the chapter. An example is that the depressive illness often contributes to the retardation of physical activities so as the responsiveness of the glottal muscles. That is, glottal muscles may be slackened as a pathology, which can be captured using speech analysis to characterize the responsiveness of glottis opening. This chapter examines this hypothesis with two additional applications/datasets. The first dataset contains speech samples from patients with tumors in the neck, so the glottal muscle is pathologically affected. Experiments show that the voice source features improve the recognition of non-intelligible speech, as direct characteristics of speech pathology. The second dataset is about the psychological stress, which often manifests physical response in the autonomic nervous system, cf. heart rate. Experiments show that the glottal features improve the recognition of stress, indicating that the autonomic nervous system also physically affects the glottal muscles, a phenomenon capturable by voice source processing.

5.1 Introduction

Glottal vibrational cycles can serve as a promising feature for monitoring mental health. Moore et al. [59] showed that linear classifiers using a combination of these features can distinguish depressed and healthy subjects with 90% accuracy. This implies that the glottal activities in speech production can be greatly affected by mental illness, a good indicator of physical change induced by mental states. We hypothesize that the glottal features can improve stress detection as well. In analogy, mental stress often manifests physical response

in the autonomic nervous system (cf. heart rate) [13]. So glottal features, indicating physical change in glottal muscles, may also respond to the autonomic nervous system. Our experiments showed that the glottal features indeed improved the classification accuracy ($> 10\%$ relative improvement for recognizing stress increase vs. stress decrease). Furthermore, the chapter verifies the hypothesis with a dataset where speakers have direct speech pathology. These speakers have tumors in the neck and are undergoing chemico-radiation therapy. The glottal muscles are clearly affected, so is the intelligibility of the speech. Experiment shows that the glottal features improves the recognition of intelligibility, a direct characteristics of speech pathology (28% relative improvement).

5.2 Extracting the Glottal Waveforms

As illustrated in Figure 3.2, a glottal (flow) waveform represents the time that the glottis is open (O) (with air flowing between vocal folds), and the time the glottis is closed (C) for each vibrational cycle. In addition, an open phase can be further broken down into opening (OP) and closing (CP) phases. If there is sudden change in airflow (i.e., shorter open and close phases), it would produce more high frequency and the voice therefore sounds more *jagged*, other than *soft* (slower change of airflow). To capture it, the library calculated measures describing timings of the phases and the ratios of closing to opening phase ($rCPOP$), open phase to total cycle ($rOTC$), closed phase to total cycle ($rCTC$), opening to open phase ($rOPO$), and closing to open phase ($rCPO$). When the speech source is altered, say the muscle is slackened, the opening phase (OP) will last longer, so is the ratio with the closing phase ($rCPOP$). These durations and ratios can serve as good features to describe the alternation. The extraction of glottal waveforms was based on Fernandez's implementation [28][27].

Following the linear source-filter theory, the estimate of the glottal vibrational cycles $G(z)$ may be obtained by inverse filtering a stationary segment of the speech signal $S(z)$ with a vocal tract transfer function estimate $V(z)$ obtained from the segment. Let S(n), G(n), V(n) be respectively the z-transforms of the acoustic speech signal $s(n)$ and impulse responses $g(n)$ and $v(n)$. The speech production often involves the lip radiation effect, which is ignored in Equation 5.1 because it can be typically modeled and removed by a differentiator (single pole transfer function), typically a pre-emphasis filtering (e.g., highpass at 6 dB/octave).

$$\begin{aligned} S(z) &= G(z)V(z) \\ s(n) &= g(n) \bigotimes v(z) \end{aligned} \tag{5.1}$$

The theory displays difficulty in estimating the glottal waveform. There is an *interaction* effect between the resonance of vocal tract and the glottal vibrations. Due to these interactions, extracting glottal vibrational cycles from the output signal becomes more difficult. There is a need to estimate the vocal tract filter $V(z)$ so that we can use it to inverse filter

the speech signal for glottal vibrational cycles $G(z)$, but the question is how to estimate $V(z)$ in a clean way given the interaction effect? The algorithm resolves this by seeking regions where these components interact minimally. Specifically, the algorithm identifies the closed-phases (Figure 3.2) of glottal vibrational cycles as the first step, where the glottis is closed so there is minimal interaction. When the glottis is closed, no air is flowing through and only the vocal tract is effective in speech production within this short window, so the formant should be stationary. The estimation of formant is reliable as well. Note formant frequencies are the realization of vocal tract resonances, including the first formant F_1, the second formant F_2, the third F_3, the forth F_4 etc. Linguistic studies show that different vowels have distinct signature of combination of formant frequencies.

Closed Phase Identification

In fact, whether the formant is *stationary* is the property that the algorithm verifies to identify closed phases. The algorithm repetitively estimates the first formant frequency F_1 with a small sliding window. If the first formant estimates are stable (i.e., do not change in an amount larger than a threshold), the vocal tract estimate $V(z)$ is reliable enough such that we can use it for both (1) calculating the durations of closed-phases (C) and (2) inverse filtering the speech $S(z)$ to acquire the glottal vibrational cylces $G(z)$.

Estimating the formant frequencies is realized by linear predictive coding (LPC) analysis, a common, efficient practice used in the speech analysis domain. LPC generates a set of polynomial coefficients representing the poles of the vocal tract filter $V(z)$. The LPC analysis window is set to $N/4$, where N is the pitch period in samples. The order is set to $\min\{16, N/4\}$.

Let roots $\{z_k\}$ be the roots of the polynomial coefficients. For each complex-conjugate pair, the formant frequencies F_n equal $\frac{Fs}{2\pi} angle(z_k)$, $n = 1, 2, 3, 4, \ldots$. Note the associated bandwidths are $b_n = -log(|z_k|)/\pi$. The formant frequencies and the bandwidths are efficient approximation of the frequency response of vocal tract filter. These values characterize the "envelope" of the frequency response, illustrated in Figure 5.2. The algorithm is summarized in Figure 5.1.

Identification of Instances of Maximum Excitation

The next part of the algorithm identifies the instances of maximum excitation. The purpose is clear. The instances of maximum excitation are the turning points where the glottis starts to close: the amount of air flowing through starts to drop. A maximum excitation breaks an open phase (C; the contrast of a closed phase O) into two sub-phases, the opening phases (OP) and closing phases (CP).

The identification exploits the properties of the average group delay of minimum-phase signals to reliably locate the maximum excitations. Speech signals can be modeled as the impulse response of a minimum-phase system. A characteristic of such systems is that the average slope of the unwrapped phase response is zero, or, if the impulse response is shifted

Let $s_{seg}(n)$ be the speech segment between two consecutive excitation instants, N its length(i.e., the pitch period in samples), and F_s the sampling frequency.

- Formant Tracking

  1. Perform LPC analysis method on $s_{seg}(n)$ with a one-sample shift, a window length $N_w = N/4$ and an order $p_{seg} = min\{16, N_w - 3\}$.
  2. For each set of LPC coefficients
     a) Let $P(z)$ be the polynomial of coefficients with roots $\{z_k\}$. For each complex-conjugate pair, find the formant candidates $f_n = \frac{F_s}{2\pi} angle(z_k)$ and their associated bandwidth $b_n = -log(|z_k|)/\pi$
     b) Let f_1 be the smallest f_n for which $b_n \geq 5f_n$
  3. Let F_1 be the set of f_1, the track of first formants. Let $F_{1,med}$ be a median filtered version of F_1 with a 4-point window.
  4. For every value m of F_1 not exceeding an allowed threshold of 1050 Hz, let m^* be the closest time index not exceeding the threshold, and let $F_1(m) = F_{1,med}(m^*)$.

- Initial Identification of Stationary Region

  5. Given $F_1(m)$, define the formant modulation function $D(n_0) = \sum_{m=n_0}^{n_0+4} |F_1(m) - F_1(m-1)|, 1 \leq n_0 \leq N - N_w - 5)$ (a cumulative first difference over a 5-point window), and let $n_0^* = argmin_{n_0} D(n_0)$
  6. Let $[N_i, N_f] = [n_0^* - 1, \cdots, n_0^* + 4]$ be an initial stationary region. Let μ_F and σ_F be the sample mean and variance of the first formant over the interval.

- Growing the Stationary Region to the Right

  7. While $|F_1(N_f + 1) - \mu_F| < 2\sigma_F$
     a) let $N_f \leftarrow N_f + 1$
     b) Update μ_F and σ_F

- Growing the Stationary Region to the Left

  8. While $|F_1(N_f + 1) - \mu_F| < 2\sigma_F$, let $N_i \leftarrow N_i - 1$

Figure 5.1: Algorithm for identifying the closed-phase region of a glottal cycle [27]

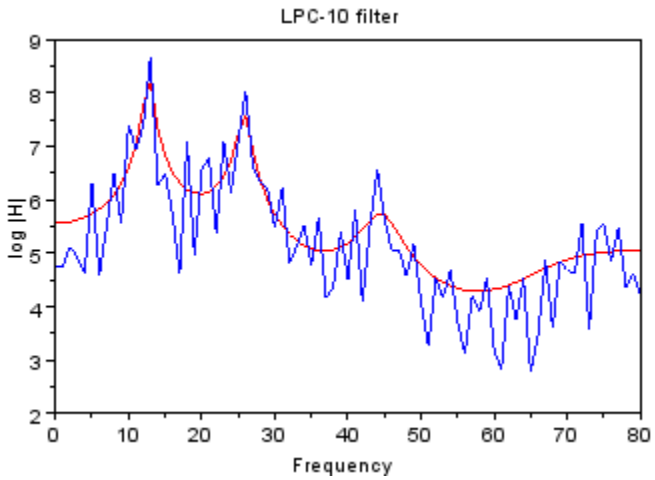


Figure 5.2: Illustration of a frequency response and its envelope, which can be characterized by the frequency locations and bandwidths of the peaks (formants).

in time, proportional to the time shift [94]. If an analysis window is centered around the excitation, the average slope of the phase response of the short-time signal should be close to zero. Otherwise, it exhibits a slope proportional to the offset of the excitation with respect to the center of the window. This suggests an algorithm which, by using a time window small enough to capture primarily one impulse, tracks the short-time frequency response and examines the average slope of the unwrapped phase response. The algorithm is summarized in 5.3.

The following sections describe two experiments in which glottal features are effective for detecting speech pathology, and consequently recognizing mental states that contribute to speech source alternation.

5.3 Application I: Classification of Intelligible vs. Non-intelligible Speech

The first application verifies the effectiveness of glottal features as indicators for speech pathology. The experiment utilizes the "NKI CCRT Speech Corpus" (NCSC) recorded at the Department of Head and Neck Oncology and Surgery of the Netherlands Cancer Institute, as described in [57]. The corpus contains recordings from 55 speakers (10 females and 45 females), who were undergoing concomitant chemo-radiation treatment (CCRT) due to the inoperable tumors of the head and neck. The speech source (i.e., glottal muscles) are likely to be affected by the tumors at the neck, and this application helps verify that the glottal features are effective in distinguishing the level of speech pathology. All speakers read a Dutch text of neutral content. Not all speakers were Dutch native speakers. Average speaker

Segment the speech into M disjoint voiced segments based on pitch. Let $f_0(m)$ denote the mean fundamental frequency of the mth segment, and let $s(n)^m$ be the mth voiced segment. For $m = 1, \ldots, M$:

1. Calculate the 10th-order LPC residual of $s(n)^m$, a Hanning analysis window of 25 msecs, and a frame rate of 100 frames/sec.
2. Find the short-time fast Fourier transform (STFT) of the residual form using a Hanning window of length $1.5/f_0(m)$ secs, zero-padded to the next integer power of 2. Shift the analysis window by one sample
3. For each frame n of the STFT:
   a) Unwrap the phase
   b) Using linear regression, find the best linear fit to the unwrapped phase. Let $\phi(n)$ be the slope of the fit for the n^{th} frame.
4. Smooth the phase function $\phi(n)$ with a Hanning window of 4 msecs, and remove the mean.
5. Assign the zero-crossing instances of the zero-mean smoothed phase slope function to the instants of maximum excitation.

Figure 5.3: Algorithm for identifying instances of maximum excitation [27]

age was 57. The original samples were segmented at the sentence boundaries, resulting a total of 1646 utterances.

Thirteen recently graduated or about to graduate speech pathologists evaluated the speech recordings on an "intelligibility" scale from 1 to 7. To establish a consensus from the individual intelligibility ratings, the evaluator weighted estimator (EWE) [34] was used. The EWE is a weighted mean of the ratings, with weights corresponding to the reliability of each rater, which is the cross-correlation of her/his rating with the mean rating (over all raters). The average rank correlation (Spearman's rho) of the individual ratings with the mean rating is 0.783. The EWE was calculated and discretized into binary class labels (intelligible, non-intelligible), dividing at the median of the distribution. Note that the class labels of the speech are not exactly balanced (725/921) since the median was taken from the ratings of the non-segmented original speech.

Baseline Feature Set

The evaluation strategy was to verify whether glottal features can achieve better performance (e.g. classification accuracy) than a "baseline" feature set. We adapted the one used in Interspeech 2012 Speaker Trait Challenge [80] as the baseline acoustic feature set, pri-

Table 5.1: Low-level descriptors in the baseline feature set

4 energy related LLD
Sum of auditory spectrum (loudness)
Sum of RASTA-style filtered auditory spectrum
RMS Energy
Zero-Crossing Rate
54 spectral LLD
RASTA-style auditory spectrum, bands 1-26 (0-8 kHz)
MFCC 1-14
Spectral energy 250-650 Hz, 1 k- kHz
Spectral Roll Off Point 0.25, 0.50, 0.75, 0.90
Spectral Flux, Entropy, Variance, Skewness,
Kurtosis, Slope, Psychoacoustic Sharpness, Harmonicity
6 voicing related LLD
F0 by SHS + Viterbi smoothing, Probability of voicing
logarithmic HNR, Jitter (local, delta), Shimmer (local)

marily due to the fact that the Interspeech Challenge provides such feature set alongs with the recognition performance evaluated directly on the NCSC dataset. With such reference result, the experiment will be convincing and valuable if our improved feature set performs better. In short, the baseline feature set unifies the acoustic feature sets used for the Interspeech 2010 Paralinguistic Challenge dealing with ground truth (non-perceived) speaker traits (age and gender) with the new acoustic features introduced for the Interspeech 2011 Speaker State (SSC) and Audio-Visual Emotion Challenges (AVEC) aiming at the assessment of perceived speaker states. The challenge uses TUM's open-source openSMILE feature extractor [25] and provide extracted feature sets on a per-utterance level. The feature set preserves the high-dimensional 2011 SSC feature set including energy, spectral and voicing related low-level descriptors (LLDs, in the form of signal waveforms, Table 5.1); a few LLDs are added including logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral sharpness, as in the AVEC 2011 set. The functionals summarizing the statistics over each converted LLD is listed in Table 5.2. Altogether, the 2012 Speaker Trait Challenge feature set contains 6125 features.

Experimental Results

Linear SVM was adapted for the binary classification task, with performance evaluated with 10-fold cross validation. The baseline feature set achieved 74.19% accuracy. With the baseline result provided, we were able to show that the glottal features achieved higher accuracy. The glottal feature set was extracted in the same manner as the baseline feature set, by projecting the waveform contours (i.e., low level descriptors; LLDs) to a feature

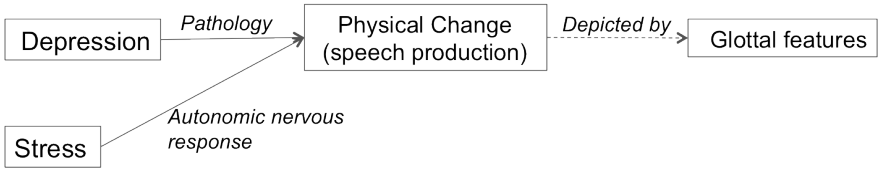
Table 5.2: Applied functionals in the baseline feature set

Functionals applied to LLD / Δ LLD
quartiles 13, 3 inter-quartile ranges
1 % percentile (≈ min), 99 % percentile (≈ max)
position of min / max
percentile range 1 % - 99%
arithmetic mean1, root quadratic mean
contour centroid, flatness
standard deviation, skewness, kurtosis
rel. duration LLD is above / below 25 / 50 / 75 / 90% range
rel. duration LLD is rising / falling
rel. duration LLD has positive / negative curvature
gain of linear prediction (LP), LP Coefficients 15
mean, max, min, std. dev. of segment length
Functionals applied to LLD only
mean of peak distances
standard deviation of peak distances
mean value of peaks
mean value of peaks arithmetic mean
mean / std.dev. of rising / falling slopes
mean / std.dev. of inter maxima distances
amplitude mean of maxima / minima
amplitude range of maxima
linear regression slope, offset, quadratic error
quadratic regression a, b, offset, quadratic error
percentage of non-zero frames

vector by functionals. In the case of glottal features, the LLDs are the contours of glottal timings across vibrational cylces (i.e., sequences of *O, C, OP, CP, rCPOP, rOTC, rCTC, rOPO*, and *rCPO*). Moreover, to each of these LLDs, the delta coefficients are additionally computed. The computed functionals are adapted from those used in Interspeech Challenge 2009 [78], much simpler than the version in Interspeech Challenge 2012 (Table 5.2). They are mean, standard deviation, kurtosis, skewness, minimum, maximum range, and rel. position. In the end, the glottal feature set achieved 81.60% accuracy (7.41% absolute improvement and 28.7% relative improvement).

The result implies that the glottal features are better indicators for speech pathology. Retrospectively, the result also serves as the basis for verifying the hypothesis of the chapter, the mental affects the physical. Now it shows the glottal features describe the alternation of speech production. The next application leverages a stress speech dataset, showing that mental stress will affect the glottal muscles physically through the autonomic nervous system,

Figure 5.4: Hypothesis of Stress Detection by Glottal Features



and that the glottal features help improve the detection of mental stress. Note mental stress often manifests physical response in the autonomic nervous system (cf. increase heart rates). The hypothesis of relationship is illustrated in Figure 5.4.

5.4 Application II: Classification of Speech Under Stress

We evaluated stress detection with a dataset named Speech Under Simulated and Actual Stress (SUSAS) [37], developed by John Hansen. It is the most common dataset found in the literature for stress detection tasks [38]. For our experiment, we made use of the recordings under actual stress, where each subject was asked to speak (and repeat) 35 distinct English words while riding one of two roller coaster rides. High stress and neutral speech utterances were marked depending on the position of a riding course. There are a total of 7 subjects (3 females and 4 males) involved, producing a total of 1900 utterances. Each utterance was segmented as a word, lasting about one second.

The Feature Sets

For comparison, a baseline feature set was developed, following a similar strategy used in the previous application. The experiments started by applying the baseline feature set to obtain the baseline performance. Then, an experimental feature set containing the glottal features was composed for the goal of improving the baseline performance. Table 5.3 shows the feature sets, where the glottal timings are included together with the baseline feature set to form the experimental feature set.

The feature set adapts the one used in Interspeech Emotion Recognition Challenge 2009 [78]. The rich feature set is composed of prosodic and spectral features which support emotion recognition with state-of-the-art accuracy.

Recognizing Stressed vs. Neutral Speech

This was a 2-way classification problem, where utterances with high stress were put to class 1 and the neutral utterances were assigned to class 2. Both baseline and experimental features

Table 5.3: The feature set, computed by applying functionals on LLD waveforms.

LLDs	functionals
(Δ)ZCR (Δ)RMS energy (Δ)F0 (Δ)HNR (Δ)MFCC 1-12 (Δ)Glottal timings x 9 [a]	mean, standard deviation, kurtosis, skewness, minimum, maximum range, rel. position

[a]The experimental feature set includes the additional glottal timings as part of the LLDs, whereas the rest of LLDs are used as the baseline feature set

Table 5.4: Comparison in the recognition of stressed vs. neutral utterances, by including additional glottal features (class size: 1200/701).

Feature Set	F-Measures	ROC Area	Accuracy
Baseline	0.867/0.770	0.763	83.18%
+Glottal	0.877/0.788	0.832	84.43%

were extracted from the utterances. The feature vectors were fed into SVM (regularized linear SVM, features scaled, 10-fold cross validation). Table 5.4 shows that experimental feature set (denoted as "+Glottal" because the experimental set includes the additional glottal features than the baseline set) outperformed the baseline one with 1%, reaching 84% of accuracy. Note the accuracy is much better than blindly guessing the majority class, which is of accuracy 63% because of the imbalanced data size 1200/701. Also, the area under the ROC (receiver operating characteristic) curve significantly increased from 0.763 to 0.832.

Recognizing Stress Increase vs. Stress Decrease in Speech

We hypothesized that stress detection can be further improved by user normalization. Because of user difference, the feature vectors in the previous task may be biased with offsets in different directions and scales in the feature space, ruining the classification. Nonetheless, if we look at the distance from a feature vector in neutral condition to another vector in stressed condition of the same user, we can focus on the within-user stress change (i.e., stress increase) and ignore the user difference.

Because of the nature of SUSAS, each user speaks the same set of words in both stress and neutral conditions. Therefore, we calculated the distance vector (by element-wise subtraction) between each pair of stress/neutral utterances of the same word by the same user.

Table 5.5: Comparison in the recognition of stress increase vs. stress decrease, by including glottal features (class size: 337/336)

Feature Set	F-Measures	ROC Area	Accuracy
Baseline	0.923/0.923	0.923	92.27%
+Glottal	0.936/0.936	0.936	93.60%

We also randomized the order of subtraction so some distance vectors represent the increase of stress (a stress vector minus a neutral vector) whereas other distance vectors represent the decrease of stress.

The task became a two-way classification, where distance vectors with stress increase were put in class 1 and the distance vectors with stress decrease were put to class -1. Note we partitioned for cross-validation in a way that the distance vectors by each user were placed in the same pool (i.e., training set or test set), so this is a user-independent classifier. The classifier is not trained with some data from a user that is to be evaluated in test set, i.e., stress pattern of the user was not seen before. We extracted both baseline and experimental feature vectors. The feature vectors were fed into SVM (regularized linear SVM, features scaled, 10-fold cross validation). Table 5.5 shows that the additional glottal features outperformed the baseline with 1.3%, reaching 93.6% of accuracy (blind guess should give accuracy of 50% because of the dataset is symmetric and balanced). The 1.3% increase is significant at the 92% accuracy level. Also, the ROC area increased from 0.923 to 0.936. This again demonstrates that, by adding glottal features it performs better in stress detection. Glottal features shows an promising way of reflecting physical response to stress in the human voice.

Readers may be questioning that this is not the real stress vs. neutral classification. Nonetheless, we argue that this result is more insightful for real world applications. The result shows that if the norm of a user's speaking characteristics is obtained, the system can accurately detect stress change (increase or decrease) by the displacement vector from the norm to the current feature vector. We can calculate the difference from the current feature vector to the next, and judge whether a user has stress level increased or decreased. In addition, the result is very promising (93% accuracy for a balanced dataset), and can be used to create a temporal model of stress detection.

5.5 Conclusion

The chapter describes the usage of speech source features for the detection of mental states. Inspired from the related work that the features describing glottal vibrational cycles are effective in detecting depression, we hypothesized that the alternation of speech source vi-

bration is a good indicator to mental states. Recognizing the mental states should follow the idea of "the mental affects the physical". For the related work example, depressive illness often manifests motor retardation, which may also affect the muscles triggering glottal vibration. Similarly, the mental stress often carries "fight-or-flight response" [7], which is triggered by the autonomic nervous system. Experiments described in the chapter show that the glottal features are effective detecting the stress change, implying the autonomic nervous system also alters the glottal muscles. In fact, our physical body often directs the trigger of muscles in may subconscious ways. Imagine that a person is experiencing a serious crying episode. He or she may occasionally breathe in very hard, which is a bodily response by the sympathetic nervous system to opening throw in order to increase air flow. In Chapter 6, two applications are presented along with the topic of bodily response. The detection of mental stress is revisited with more detailed analysis. In addition, we evaluated the change of voice characteristics for subjects undergoing sleep deprivation, a condition having impacts on the body in multiple aspects.

Chapter 6

Trigger by the Physical Body

6.1 Introduction

Monitoring the mental states can be achieved by directly probing users' thoughts, say analyzing the diary or the speech content of a user. Another way to do this, which is the main approach of the thesis, is to observe of the physical realization of mental states. That is, the streamline of research is to identify the correlation between the onset of some mental states and the triggers of bodily response. As Chapter 5 hinted, mental stress affects the autonomous nervous system, which will trigger many bodily response, including the change of heart rate, sweating, and even the vibration of glottis. This chapter explores this idea with additional applications, in particular the effects of sleep deprivation. Mental disorders often have significant impact on sleep. So sleep deprivation has become an important factor to diagnose the existence of mental disorder. The application evaluated the change of voice characteristics for subjects who went through sleep deprivation. Moreover, the detection of mental stress is revisited, by reviewing the difference between stress with a thrill factor and stress with a high mental work load. It also tries to answer a question about the optimal length of speech for accurate stress detection etc.

6.2 Application I: Sleep Deprivation

The literature suggests a critical role for sleep in our bodily functioning. Generally, sleep deprivation may result in aching muscles, headaches, increased sensitivity to cold, increased blood pressure, increased risk to depression, diabetes, etc [93]. Moreover, the literature also suggests a correlation between the amount of sleep and emotional functioning. Healthy adult participants whose sleep was restricted to 5 hour per night over one week reported a progressive increase in negative emotion [22]. A goal of the present study was to delineate one possible modifiable mechanism by which a critical, but understudied, feature of adolescent emotion difficulties might be maintained; namely, sleep deprivation.

As part of a larger 2-day study on affect and sleep deprivation in adolescents compared to adults, the current study focused on vocal expression of emotion on one of the nights of sleep deprivation [54]. A multi-method approach was used, and my work focused on the part of computerized acoustic properties of vocal expression. Based on past research [30], we hypothesized that vocal expression of positive emotions would decrease and vocal expression of negative emotions would increase after sleep deprivation, relative to when rested. The second aim was to determine if sleep deprivation affects vocal expression of emotion differently for adolescents relative to adults. We predicted that the hypothesized emotional effects of sleep deprivation would be greater for adolescents relative to adults.

The Study

Note the study was based a collaboration with a team in Department of Psychology, led by professor Allison Harvey and graduate student researcher Eleanor McGlinchey. The study was conducted and led directly by Harvey's team, and I contributed partly, namely the part of the analysis of the computerized acoustic properties of vocal expression. The details can be found in [54], but this section will mainly report the part where I had direct contribution. Describing the analysis requires the delineation of dataset so as the study procedure. Therefore, a brief description of the study procedure follows.

55 healthy participants completed the study. 38 adolescents (15 female) aged 11-15 years and 17 adults (9 female) aged 30-60 years participated the study. Individuals aged 16-30 years were excluded primary reasons to provide a clear neuro developmental difference and clear differentiation of sleep patterns between the adult and adolescent groups [54].

The sleep deprivation protocol occurred over 2 nights. On the first night, participants were asked to restrict their sleep to a maximum of approximately 6.5 hours at home. Participants came to the laboratory on the second night at 22:00. At 22:30, a baseline Stanford Sleepiness Scale (SSS) rating was completed, which is a 1-item measure of subjective sleepiness (with scale 1-7). In addition, the first Speak Freely Interview Procedure was administered. Participants were then continuously monitored throughout the night by trained laboratory staff. They were permitted to interact with the laboratory staff in order to ensure wakefulness, as well as to read, watch movies, and play board games. A small snack, such as fruit or crackers, was made available by the laboratory staff. No caffeine or other stimulants were allowed. Between 03:00 and 05:00, participants were given a 2-hour nap opportunity. After waking, participants had a breakfast consisting of fruit, crackers, yogurt, and cheese. At 06:30, the SSS and the second Speak Freely Interview were repeated.

In the Speak Freely Interview, participants were asked the following 4 questions by a trained research staff member, who requested they spend one minute answering each question. The questions for each time period were as follow. Recording conditions were kept consistent across all participants during all interviews.

- 22:30:
    1. How are you feeling right now?

2. What are you looking forward to tonight?
3. How do you expect you'll feel without sleep?
4. Is there anything you're not looking forward to?

- 06:30:

1. How are you feeling right now?
2. What are you looking forward to today?
3. How do you expect you'll feel the rest of this morning without sleep?
4. Is there anything you're not looking forward to?

Computerized Acoustic Properties of Vocal Expression

While a great deal of research has focused on acoustic properties as measures of emotion, minimal research has leveraged this method to investigate emotion in sleep deprivation. The vocal properties investigated were selected from Juslin and Scherer's [43] summary of properties that are correlated with emotion. A total of thirty features were extracted in the categories of fundamental frequency, jitter, intensity, shimmer, speech rate, pauses and high frequency energy. The features resemble the ones described in Section 3.4, but the details are described here for clarity.

Fundamental frequency (F0) is a measure of pitch and is represented by the rate (1/sec) at which the vocal folds open and close. The unit of measurement for F0 is cycles per second or hertz (Hz). A sudden increase in fundamental frequency is associated with high activation emotion such as anger, whereas a low rate in fundamental frequency is interpreted as low energy or sadness [43]. The dynamics of the fundamental frequency contour were calculated by several statistical measures: average (F0_avg), standard deviation (F0_std), minimum (F0_min), maximum (F0_max), and range (F0_range).

Jitter is pitch perturbation and is represented by small-scale rapid and random fluctuations of F0, meaning fluctuations of the opening and closing of the vocal folds from one vocal cycle to the next. Previous research suggests that jitter is an indicator of stressor-provoked anxiety [33]. Two methods were applied to calculate jitter in this study, (1) by calculating the average of the first-order difference sequence in F0 (F0_jitter_PF), and (2) by calculating the average of the difference sequence over the mean of running F0 values (rather than over the preceding F0 value) with different cycle lengths (F0_jitter_PQ_mean).

Intensity reflects the energy (in dB) in acoustic signal or loudness of speech. Previous research suggests that a rapid rise in intensity is associated with angry speech and sad speech is characterized by low intensity [43]. Several statistical measures were applied in order to describe the dynamics of intensity including Energy average (Energy_avg), standard deviation (Energy_std), minimum (Energy_min), maximum (Energy_max), and range (Energy_range). Intensity can also be analyzed by interpreting its distribution over frequency bands (i.e., spectrogram). Previous research suggests that an emphasis on loudness of psycho-acoustical barks in certain high frequency energy bands may be indicative

of emotional speech [28]. Specifically, the following energy values (in dB) in high frequency energy bands with bark scales were processed: energy_bark7 at 700-840Hz, energy_bark8 at 840-1000Hz, energy_bark9 at 1000-1170Hz, energy_bark10 at 1170-1370Hz, energy_bark11 at 1370-1600Hz, energy_bark12 at 1600-1850Hz, energy_bark13 at 1850-2150Hz, energy_bark14 at 2150-2500Hz, energy_bark15 at 2500-2900Hz, and energy_bark16 at 2900-3400Hz.

Shimmer is the loudness perturbations in speech and is measured by the small variations of energy amplitude in successive glottal cycles. Shimmer can serve as an indicator of underlying stress in human speech [33]. Two features were calculated to describe shimmer: 1) Loud_shimmer_PF which is the average of the first order difference sequence and 2) Loud_shimmer_PQ_mean which is the average of the difference sequence over the mean of running energy values (rather than over the preceding energy value) with different cycle lengths.

For the temporal aspects of speech, we included measures to describe speech rate and pauses. Previous research indicates that sadness often results in slower speech and more pauses [28]. Both speech rate and pauses were calculated by measuring the voiced sections (F0 $>$ 0) in speech. Speech rate was represented by the relative ratio of voiced versus unvoiced sections (ratio_voiced_over_unvoiced). Pauses were calculated and approximated by counting unvoiced sections (silence_voiced_count) and summing the total duration (in seconds) of silence (unvoiced sections, silence_duration).

As the amount of high-frequency energy increases, the voice sounds sharp and less soft [43], which can also be emotion-dependent. Therefore, we analyzed the amount of high-frequency energy in the spectrogram, by calculating the cumulative values (in dB) in the spectrogram that appear above two cut-off frequency thresholds: 500Hz (HF500) and 1000Hz (HF1000). In addition, the trend of high-frequency energy distribution (Slope1000) was calculated by the linear regression of the energy distribution in the frequency over 1000Hz.

All properties were extracted from the digital audio recordings via the MATLAB platform based on methods used by Moore, Clements, Peifer, and Weisser [59] and Fernandez and Picard [28]. Specifically we applied Moore's implementation to find intensity and Fernandez's implementation to find jitter, shimmer, speech rate, and high frequency energy. In addition, we used the Praat speech analysis software to extract fundamental frequency [3]. Praat is a computer program commonly used for acoustic analysis of vocal expression in clinical and research settings.

Experimental Results

Table 6.1 presents the mean values for each of the acoustic properties from 10:30 p.m. to 06:30 a.m. for the adolescent and adult participants. We conducted repeated measures ANOVAs for the 30 acoustic properties. For fundamental frequency (F0), there was a significant main effect of Time (between 10:30 p.m. and 06:30 a.m.) for F0 average, $F(1, 53) = 8.14, p < 0.01$, such that all participants expressed a decreased rate in F0 at 06:30 a.m. relative to 10:30 p.m. There was no main effect of Group (between adults and adolescents) $F(1, 53) < 1, ns$, nor a Group $\times$ Time interaction, $F(1, 53) < 1, ns$, for F0 average. Additionally, there were no

main effects of Group or Time and no Group × Time interactions for the standard deviation, minimum, maximum, and range of F0 (see Table 6.1).

There was a significant main effect of Time for both of the methods applied to calculate jitter. For the average of the first-order difference sequence in F0, $F(1, 53) = 4.12, p < 0.05$, such that all participants expressed an increase in jitter at 06:30 a.m. relative to 10:30 p.m. There was no main effect of Group $F(1, 53) = 1.65, ns$, nor a Group × Time interaction, $F(1, 53) < 1, ns$. Additionally, there was a main effect of Time for the average of the difference sequence over the mean of running F0 values with different cycle lengths, $F(1, 53) = 5.23, p < 0.05$, such that all participants expressed an increase in jitter at 06:30 a.m. relative to 10:30 p.m. There was no main effect of Group $F(1, 53) < 1, ns$, nor a Group × Time interaction, $F(1, 53) < 1, ns$.

For intensity, there were no main effects of Group or Time and no Group × Time interactions for the average, standard deviation, minimum, maximum, and range of energy (see Table 6.1). However, when intensity was measured in specific high frequency energy bands, there were significant main effects of Time in the following bark scales, such that all participants expressed decreases in psycho-acoustical barks at 06:30 a.m. relative to 10:30 p.m.: bark7 at 700-840Hz $F(1, 53) = 4.31, p < 0.05$, bark8 at 840-1000Hz $F(1, 53) = 5.33, p < 0.05$, bark9 at 1000-1170Hz $F(1, 53) = 9.89, p < 0.01$, bark10 at 1170-1370Hz $F(1, 53) = 11.62, p = 0.001$, bark11 at 1370-1600Hz $F(1, 53) = 12.97, p = 0.001$, bark12 at 1600-1850Hz $F(1, 53) = 16.81, p < 0.001$, bark13 at 1850-2150Hz $F(1, 53) = 13.40, p = 0.001$, bark14 at 2150-2500Hz $F(1, 53) = 8.15, p < 0.01$ and bark15 at 2500-2900Hz $F(1, 53) = 4.22, p < 0.05$. There were no main effects of Group nor any Group × Time interactions for these bark scales (see Table 6.1). Additionally, there were no significant main effects of Time or Group and no Group × Time interactions for bark16 at 2900-3400Hz (see Table 6.1). Note the magnitude of energy appearing in frequency bands (energy_bark7-16) is much smaller than the magnitude of the overall energy (energy_avg) given that each bark value is a decomposition of the total energy.

There was a significant main effect of Time for both of the methods applied to calculate shimmer. For the average of the first-order difference sequence in energy, $F(1, 53) = 11.83, p = 0.001$, such that all participants expressed an increase in shimmer at 06:30 a.m. relative to 10:30 p.m. There was no main effect of Group $F(1, 53) < 1, ns$, nor a Group × Time interaction, $F(1, 53) < 1, ns$. Additionally, there was a main effect of Time for the average of the difference sequence over the mean of running energy values with different cycle lengths, $F(1, 53) = 9.97, p < 0.01$, such that all participants expressed an increase in shimmer at 06:30 a.m. relative to 10:30 p.m. There was no main effect of Group $F(1, 53) < 1, ns$, nor a Group × Time interaction, $F(1, 53) < 1, ns$.

For the temporal aspects of speech, there were no main effects of Group or Time and no Group × Time interactions for speech rate (see Table 6.1). However, there was a main effect of Time for Pauses, $F(1, 53) = 12.58, p = 0.001$, such that all participants expressed a decrease in pauses at 06:30 a.m. relative to 10:30 p.m. There was no main effect of Group $F(1, 53) < 1, ns$, nor a Group × Time interaction, $F(1, 53) < 1, ns$. Additionally, there were no main effects of Group or Time and no Group × Time interactions for the total duration

of silence (see Table 6.1).

For high frequency energy in the spectrogram above 500Hz and 1000Hz, there were no main effects of Group or Time and no Group × Time interaction for 500Hz (see Table 6.1), but there was a main effect of Time for 1000Hz, $F(1, 53) = 7.91, p < 0.01$, such that all participants expressed a decrease in high frequency energy above 1000Hz at 06:30 a.m. relative to 10:30 p.m. There was no main effect of Group $F(1, 53) < 1, ns$, nor a Group × Time interaction, $F(1, 53) < 1, ns$. In addition, the spectral slope over 1000Hz became flatter in all participants at 06:30 a.m. relative to 10:30 p.m., $F(1, 53) = 26.56, p < 0.001$. There was no main effect of Group $F(1, 53) < 1, n$s, nor a Group × Time interaction, $F(1, 53) < 1, ns$.

Table 6.1: Mean values for acoustic properties in adolescents and adults (with standard deviations in parentheses).

	Adolescents		**Adults**	
	22:30	**06:30**	**22:30**	**06:30**
F0_avg (Hz)	148.72 (27.86)	139.20 (20.73)	144.46 (30.89)	134.19 (22.03)
F0_std (Hz)	51.08 (15.89)	43.72 (17.40)	51.10 (15.61)	52.48 (26.67)
F0_min (Hz)	98.18 (11.37)	95.16 (5.75)	93.62 (10.32)	90.86 (8.84)
F0_max (Hz)	237.11 (47.3)	213.86 (48.26)	229.22 (61.75)	231.48 (98.86)
F0_range (Hz)	138.93 (45.49)	118.70 (48.55)	135.60 (54.53)	140.61 (101.37)
F0_jitter_PF (Hz)	0.34 (0.07)	0.36 (0.07)	0.32 (0.07)	0.33 (0.06)
F0_jitter_PQ_mean (Hz)	0.20 (0.04)	0.21 (0.04)	0.19 (0.04)	0.20 (0.03)
Energy_avg (dB)	0.21 (0.01)	0.22 (0.01)	0.21 (0.01)	0.21 (0.01)
Energy_std (dB)	0.13 (0.02)	0.12 (0.02)	0.13 (0.02)	0.13 (0.02)
Energy_min (dB)	0.09 (0.02)	0.09 (0.02)	0.09 (0.02)	0.09 (0.02)
Energy_max (dB)	0.45 (0.04)	0.44 (0.04)	0.43 (0.03)	0.45 (0.03)
Energy_range (dB)	0.36 (0.05)	0.35 (0.05)	0.35 (0.05)	0.36 (0.04)
Energy_bark7 at 700-840Hz (dB)	$3.46*10^{-3}$ ($1.09*10^{-3}$)	$2.60*10^{-3}$ ($1.59*10^{-3}$)	$3.70*10^{-3}$ ($1.57*10^{-3}$)	$3.48*10^{-3}$ $1.97*10^{-3}$)
Energy_bark8 at 840-1000Hz (dB)	$2.66*10^{-3}$ ($1.37*10^{-3}$)	$1.83*10^{-3}$ ($1.75*10^{-3}$)	$3.20*10^{-3}$ ($1.59*10^{-3}$)	$2.85*10^{-3}$ ($2.55*10^{-3}$)
Energy_bark9 at 1000-1170Hz (dB)	$2.36*10^{-3}$ ($1.53*10^{-3}$)	$1.54*10^{-3}$ ($1.52*10^{-3}$)	$2.66*10^{-3}$ ($1.46*10^{-3}$)	$2.37*10^{-3}$ ($1.90*10^{-3}$)
Energy_bark10 at 1170-1370Hz (dB)	$1.87*10^{-3}$ ($1.14*10^{-3}$)	$1.16*10^{-3}$ ($0.98*10^{-3}$)	$1.98*10^{-3}$ ($1.16*10^{-3}$)	$1.76*10^{-3}$ ($1.41*10^{-3}$)
Energy_bark11 at 1370-1600Hz (dB)	$1.54*10^{-3}$ ($0.92*10^{-3}$)	$0.92*10^{-3}$ ($0.68*10^{-3}$)	$1.59*10^{-3}$ ($0.81*10^{-3}$)	$1.36*10^{-3}$ ($1.05*10^{-3}$)
Energy_bark12 at 1600-1850Hz (dB)	$1.28*10^{-3}$ ($0.77*10^{-3}$)	$0.75*10^{-3}$ ($0.59*10^{-3}$)	$1.26*10^{-3}$ ($0.70*10^{-3}$)	$0.96*10^{-3}$ ($0.67*10^{-3}$)
Energy_bark13	$1.18*10^{-3}$	$0.64*10^{-3}$	$0.87*10^{-3}$	$0.63*10^{-3}$

Table 6.1: Mean values for acoustic properties in adolescents and adults (with standard deviations in parentheses).

at 1850-2150Hz (dB)	($0.83*10^{-3}$)	($0.61*10^{-3}$)	($0.57*10^{-3}$)	($0.43*10^{-3}$)
Energy_bark14	$0.85*10^{-3}$	$0.45*10^{-3}$	$0.51*10^{-3}$	$0.38*10^{-3}$
at 2150-2500Hz (dB)	($0.68*10^{-3}$)	($0.56*10^{-3}$)	($0.38*10^{-3}$)	($0.31*10^{-3}$)
Energy_bark15	$0.46*10^{-3}$	$0.25*10^{-3}$	$0.32*10^{-3}$	$0.27*10^{-3}$
at 2500-2900Hz (dB)	($0.40*10^{-3}$)	($0.32*10^{-3}$)	($0.26*10^{-3}$)	($0.29*10^{-3}$)
Energy_bark16	$0.40*10^{-3}$	$0.22*10^{-3}$	$0.30*10^{-3}$	$0.26*10^{-3}$
at 2900-3400Hz (dB)	($0.46*10^{-3}$)	($0.29*10^{-3}$)	($0.25*10^{-3}$)	($0.25*10^{-3}$)
Loud_shimmer_PF	0.73 (0.16)	0.79 (0.16)	0.72 (0.14)	0.79 (0.12)
Loud_shimmer_PQ_mean	0.27 (0.05)	0.29 (0.04)	0.26 (0.04)	0.28 (0.04)
ratio_voiced_over_unvoice	$2.90*10^{-3}$ ($2.08*10^{-3}$)	$3.0*10^{-3}$ ($1.87*10^{-3}$)	$2.90*10^{-3}$ ($1.88*10^{-3}$)	$3.0*10^{-3}$ ($1.94*10^{-3}$)
silence_voiced_count	$9.60*10^{-3}$ ($1.57*10^{-3}$)	$8.40*10^{-3}$ ($2.45*10^{-3}$)	$9.81*10^{-3}$ ($1.65*10^{-3}$)	$8.63*10^{-3}$ ($1.35*10^{-3}$)
silence_duration (seconds)	0.70 (0.08)	0.70 (0.14)	0.71 (0.06)	0.77 (0.05)
HF500 (dB)	8.31 (8.34)	5.97 (5.78)	6.99 (4.36)	6.41 (6.75)
HF1000 (dB)	0.67 (0.46)	0.48 (0.28)	0.66 (0.32)	0.51 (0.21)
Slope1000	-0.82 (0.74)	-0.39 (0.46)	-0.91 (0.74)	-0.58 (0.51)

6.3 Discussion

The application reported the impact of sleep deprivation on emotions to vocal expression in adolescents relative to adults. Following the prediction, the computerized acoustic properties analysis added to the support for the hypothesis that sleep deprivation resulted in dramatic changes in pitch, energy, and vocal sharpness. In other words, vocal expression took on a lower pitch, became less intense, and energy levels decreased. Previous research has described decreases in pitch as being associated with sadness [43]. Additionally, high frequency energy has been associated with low physiological activation. Low activation appears to be associated with sadness and fatigue [72]. Finally, increased perturbations in pitch and loudness of speech (jitter and shimmer) have been interpreted as indicative of stress or anxiety [33]. We also note that there was a decrease in pauses at 06:30 relative to 22:30; however, there were no differences in the rate of speech or total duration of silence. Therefore, it is unlikely that the pitch, energy, and vocal sharpness findings can be explained by participants speaking more slowly due to fatigue. Overall, these results are consistent with previous studies indicating that adolescents and adults experience negative mood in relation to sleep deprivation [22].

The second hypothesis was that the predicted effects of sleep deprivation would be particularly pronounced in the adolescent group relative to the adult group. The computerized acoustic properties analysis did not support this hypothesis, but in [54] it reported that based on the computerized text analysis, the adolescent group expressed fewer positive emotion words than the adult group when sleep deprived.

Several caveats are important to consider. First, the relatively small sample size, particularly for the adult group, may have limited statistical power. Additionally, the small sample size of the adolescent group precluded analysis of pubertal status on the expression of emotion and there is evidence to suggest that the voice is going through changes during puberty, particularly among adolescent boys [32]. However, we believe that a strength of the current study was the within subjects design, which allowed comparison of vocal expression at two different time points within the same participant. Second, future work should use a high quality external microphone (i.e., VoiceTracker array microphone by AcousticMagic) in a sound attenuated room. It is possible that the built-in microphone used in the current study may lose some nuanced details. However, in the current investigation, recording conditions were kept consistent across all participants and the within-subject design allowed for the analysis of a change in acoustic properties. More discussion related to other types of analysis can be found in [54].

6.4 Application II: Simulated and Actual Stress

It is believed that stress plays a critical role in survival by increasing arousal through the activation of the fight-or-flight response in the presence of danger [7]. Several hormones were generated to facilitate immediate physical reactions associated with a preparation for violent muscular action, including acceleration of heart and lung function, and constriction of blood vessels in many parts of the body. Challenges people not used to are also perceived as danger by the body, which indirectly induces work-related stress and anxiety. There is a growing attention in this area to monitor mental stress [16] and reduce it with appropriate feedbacks [77] to promote work performance and physical health.

In this section, we revisit the sensing of stress in voice, by nonverbal voice measures as indicators of stressor on the body. Section 5.4 describes part of the analysis for such goal, but mainly for analyzing the effectiveness of glottal features. This section describes several additional analysis to answer several questions for realistically building a real life stress detector. The analysis includes: (1) the investigation of a realistic evaluation in the classification of stress vs. neutral speech, (2) the investigation of an user-normalization approach, and (3) the understanding of the optimal speech duration for the best classification result. In fact, the analysis is a progression of experiments to improve the classification accuracy. In the end, with the techniques combined, the accuracy of classifying stress vs. neutral speech is improved to up to 95%.

The dataset Speech Under Simulated and Actual Stress (SUSAS) [37] is used for answering the questions. For our experiment, we made use of two set of recordings under both

actual and simulated stress. Under the "actual stress" condition, each subject was asked to speak (and repeat) 35 distinct English words while riding one of two roller coaster rides. High stress and neutral speech utterances were marked depending on the position of a riding course. There are a total of 7 subjects (3 females and 4 males) involved, producing a total of 1900 utterances. Each utterance was segmented as a word, lasting about one second. Under the "simulated stress", 9 speakers were asked to perform high workload tasks (manipulating flight control tasks on a desktop computer) while reading out the same set of English words. High stress and neutral speech utterances were marked depending on whether a high workload or low workload task is undergoing. A total of 1261 utterances were collected and each utterance corresponds to a word.

Intuitively, the actual stress involves some thrill factor during roller coaster rides, which may involve screaming from time to time, so it is more prominent in vocal expression. On the other hand, the simulated stress is triggered from the high mental work load, which should be more subtle, harder to recognize in the human voice.

To be clear about the concept and consistent throughout the chapter, the actual stress will be referred as the *thrill stress* and the simulated stress will be referred as the *work load stress*. Nonetheless, the following sections will first focus on the data with thrill-stress stressors, evolving the classification method. In the end, the finalized method will be applied to the work load stress data to compare the result with thrill stress.

User-independent Classification of Stress vs. Neutral Speech

In Section 5.4, the binary classification task between stress and neutral speech was explored. The evaluation worked by cross-validation. However, after re-visiting the partitioning method, we consider the original method not realistic for deploying a real-world application, although the result was valid. The evaluation randomly partitioned the data as folds, so the utterances by a user was scattered across multiple partitions. This implies that during cross-validation, some stress/neutral data of a user is previously seen during the training phase, so the accuracy is boosted during the classification of the user's data. Realistically, when applying a model to new set of users, the model should be user independent and should not be trained with any data by the new user.

Therefore, a leave-one-subject-out cross validation was adapted. In each fold, a user was chosen for testing, using the model trained from the speech of the remaining users (the remaining 6 users in the thrill stress case). That said, the evaluation method matches well with the real-world testing scenario. The algorithm computed a feature vector for each speech utterance (an English word, about one second long) based on the feature set listed in Table 5.3 and utilized linear SVM for classification training/testing. In fact, comparing to linear SVM, the radial-basis SVM worked poorly for emotion recognition tasks [78].

A crucial step for applying linear SVM is feature normalization, which is suggested by the authors of libsvm [40]. Feature vectors should be scaled to range [0, 1] beforehand. The scaling factor obtained from the training data ($\mathcal{P}$) was kept as part of the model parameters, and will be applied to the test data ($\mathcal{Q}$) during test phase. The scaling method is detailed in

the following, a straightforward method and will be improved in the next step. If a feature value is subscripted as the i^{th} element of the k^{th} datum (feature vector),

$$minFeat_i = \min_{k \in \mathcal{P}} feat_{k,i} \tag{6.1}$$

$$maxFeat_i = \max_{k \in \mathcal{P}} feat_{k,i} \tag{6.2}$$

$$\hat{feat}_{k,i} = \frac{feat_{k,i} - minFeat_i}{maxFeat_i - minFeat_i}, k \in \mathcal{P} \cup \mathcal{Q} \tag{6.3}$$

As expected, because of the realistic cross-validation seeting the performance dropped, significantly from 83% to 62.79% in accuracy, from 0.763 to 0.626 in ROC area. Indeed, the accuracy became unacceptable, but this experiment setup is believed to be more realistic than the previous one. After investigation, two improvements evolved, with one focusing on the improvement of feature normalization method and one targeting on the question of "data instability". Details of the two improvements follow.

User-based Normalization

Based on the idea that, each user should have his/her own baseline/range of feature values across neutral and stress conditions, so it is necessary to accommodate that individually. Instead of scaling the training data "altogether", we scale the features of each user independently. That is, the feature vectors of each user are grouped together, and scaled respectively (i.e., with different scaling factors $minFeat_{i,u}$ and $maxFeat_{i,u}$ for each user u). If user u is assigned to the training set, feature k of user u should belong to training set $\mathcal{P}_u$:

$$minFeat_{i,u} = \min_{k \in \mathcal{P}_u} feat_{k,i} \tag{6.4}$$

$$maxFeat_{i,u} = \max_{k \in \mathcal{P}_u} feat_{k,i} \tag{6.5}$$

$$\hat{feat}_{k,i} = \frac{feat_{k,i} - minFeat_{i,u}}{maxFeat_{i,u} - minFeat_{i,u}}, k \in \mathcal{P}_u \tag{6.6}$$

Equations 6.4 to 6.6 reveal a problem of this method: the scaling factor for test user u' is undefined, since the scaling factors are only defined for the training data of the other users. To resolve this, we calculated the scaling factors of the test user u' from all feature vectors of the user. This partial solution in fact, requires collecting a significant amount of data from a given user (having sufficient coverage of the range) before the scaling becomes effective. In fact, Chapter 3 found that scaling factors collected from less than 5 data points are sufficient.

The result improved, achieving 72.73% accuracy and 0.787 ROC area.

The Length of Speech Samples

Due to the nature of the dataset, each feature vector was extracted from a speech utterance lasting only 1 second. It is in fact imaginable that detecting stress from one second of speech is deemed to be inaccurate. Therefore, we experimented whether lengthening the speech samples will improve the accuracy.

"Lengthening" speech samples was achieved by randomly concatenated recordings together. In other words, we randomly "synthesized" English sentences of length N by randomly concatenating N English words together. Moreover, consider the goal is to classify stress/neutral "sentences" by each user. Concatenating speech samples from different users does not make sense. Likewise, speech samples in different conditions (neutral/stress) should not be concatenated together either. Therefore, we concatenated recordings of the same user in the same condition.

Based on the concept, we can experiment the following question. Whether training utterances of length N will give the best performance when the test utterances are of length M? This requires generating synthesized sequence with increasing durations (2, 3, 4, 5, ..., K seconds) per user per condition. If a user has 38 neutral utterances of length 1 in the original dataset, the utterances can be combined separately without repetition into 8 utterances of length 5 (length 3 for the remainder case) where each original utterance is assigned to only one synthesized utterance.

Applying the same cross-validation and user-based normalization presented in the previous two steps, Figure 6.1 shows the accuracies in the combination of training data of length N and test data of length M ($2 < N, M < 30$). Figure 6.1 shows that, as long as the test data is of longer unit length, the accuracy will increase (up to 95%). This is a significant improvement, implying that the unit length of test data should be as long as 30 seconds. Nonetheless, when the unit length of training data increases, the accuracy stays invariant. This means that the linear SVMs can accommodate the variance of the training data (even when the training data is really short), but not the variance of a single, short test datum. If we can increase the unit length of the test datum, we can remove the variance factor and achieve better result.

Thrill vs. Work Load Stress

The same experiment was performed for the subset of data with work load stress, with result shown in Figure 6.2. In fact, Figure 6.2 follows the same trend as Figure 6.1: the longer the test speech the better. Nonetheless, the accuracy level for recognizing work load stress is not as high as detecting thrill stress. For example, the high value is 85%, lower than 95% in the thrill stress case. This follows the intuition that recognizing mental-load related stress through voice is more difficult than detecting thrill related stress that involves screaming from time to time.

This can be explained by feature analysis. The analysis is to look at what category of features are more important to the classification. A particular hypothesis is that the energy

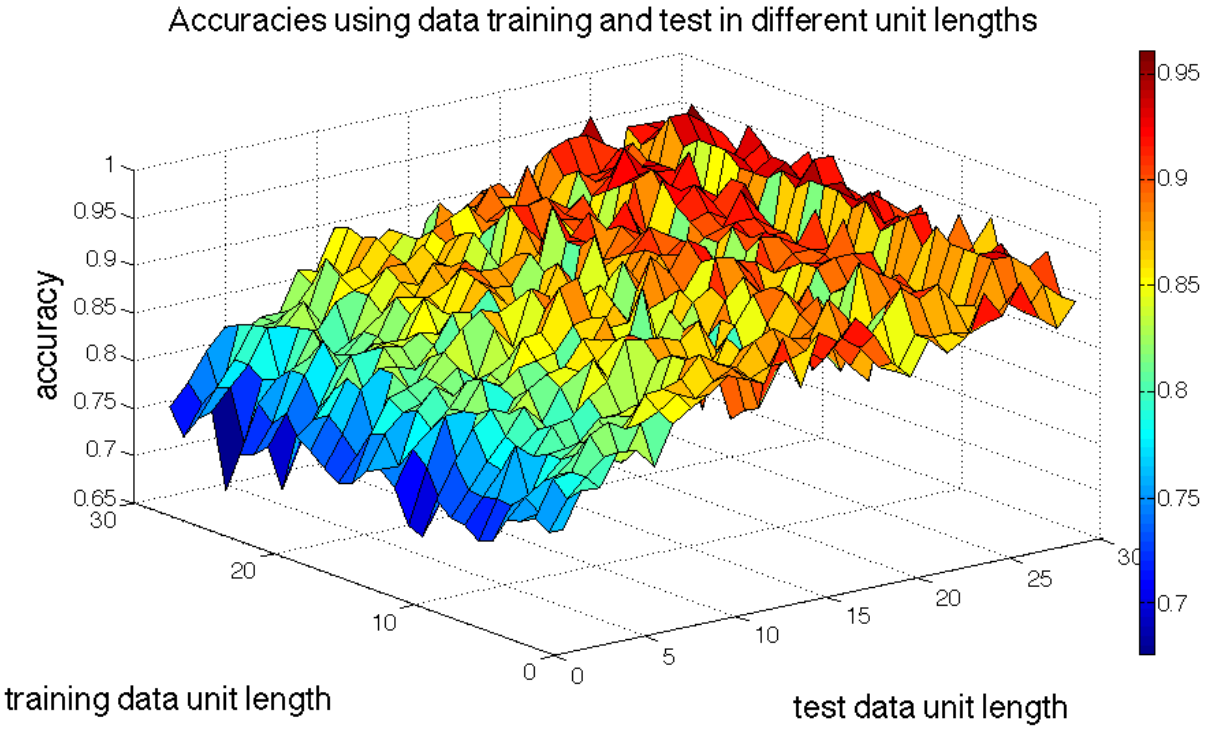


Figure 6.1: Accuracies with combination of training data of length N and test data of length M, $2 < N, M < 30$, with "thrill-stress" stressor.

feature in the thrill stress dataset should stand out more than in the work load dataset.

The Linear SVMs provide a way to evaluate the importance of features because an SVM assigns importance weights to its features for class prediction. The prediction is made by weighted linear combination of features, i.e., $y = \arg\max_c \sum_k w_{ck} x_k + b_c$. We can think of the weight w_{ck} as a vote assigned to a particular feature x_k. We constrained the variability of the features to lie between 0 and 1. The classifiers were given 384 different features as input (Table 5.3 without the glottal features), and, as the rank-order weight plot in Figure 6.3 shows, about 50-100 features have high weight after feature selection. Therefore, we reviewed the appearance of the features in the top 75 features for trained models in thrill and work load stress.

Figure 6.4 shows the distribution of each feature category in the top 75 features, chosen by the (a) thrill stress model and (b) work load stress model. It is clear that both models picked MFCCs, the energy distribution over the frequency bands, as the most important features ($> 60\%$), whereas the work load-stress model put more emphasis on it. This follows [43] that energy in high frequency associated well with physiological activation, which is higher during stress episodes. However, the distribution in Figure 6.4 could be distorted, weighted a bit towards MFCCs. This may due to the fact that over the 384 features, the MFCC category has 12 times more features than the other four categories (HNR, F0, ZCR and RMS-energy),

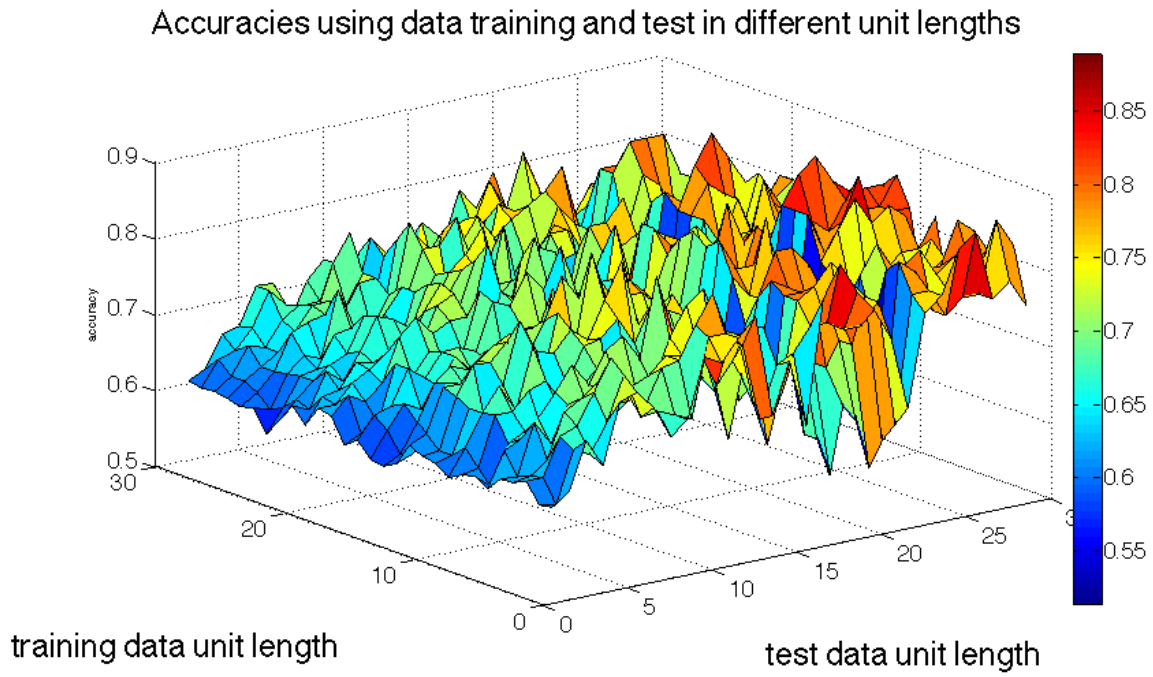


Figure 6.2: Accuracies with combination of training data of length N and test data of length M, $2 < N, M < 30$, with "work load-stress" stressor.

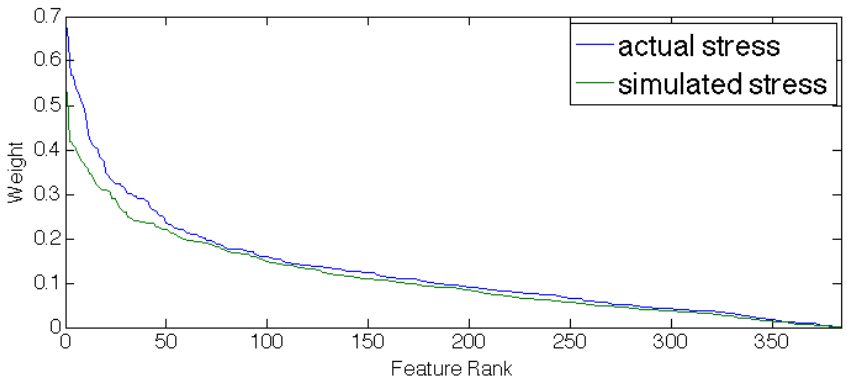
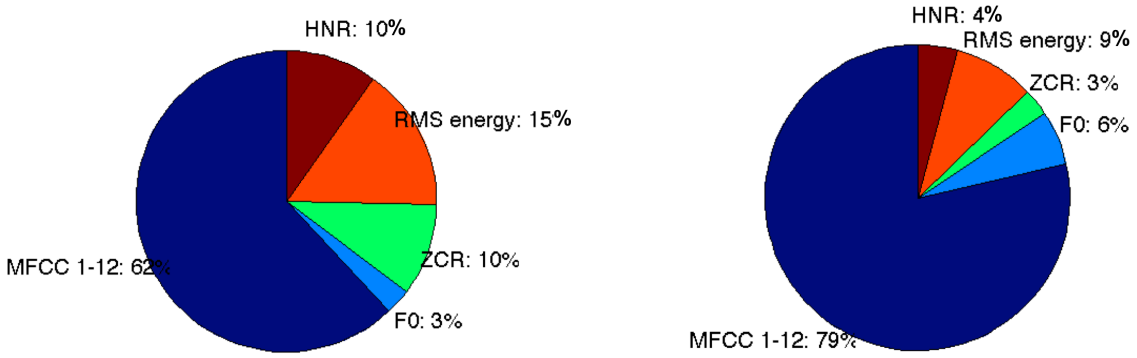


Figure 6.3: Weights of the Linear SVM's features for thrill stress and work load stress plotted in rank order show that,



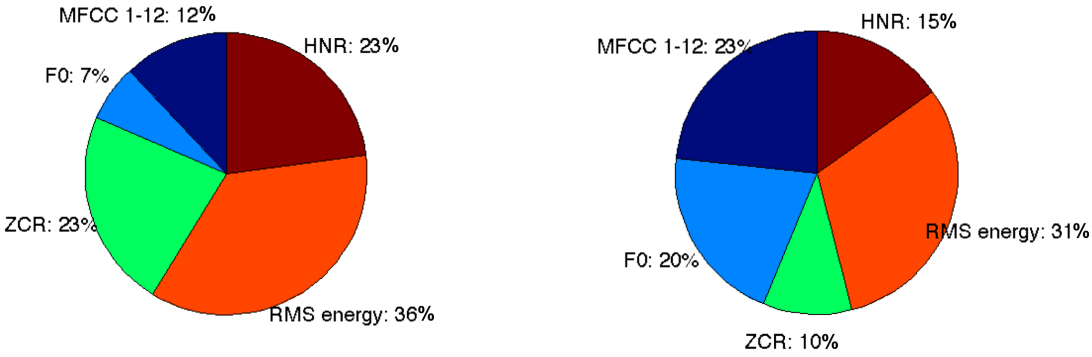
(a) (b)

Figure 6.4: Distribution of feature categories in the top 75 features chosen by linear SVM for (a) thrill stress and (b) work load stress.

because there are 12 LLDs in the MFCC category (MFCC 1-12) whereas there is only one in the other category (Table 5.3). Although the MFCCs are important so the linear SVMs pick the features repetitively as top features, the majority of the MFCC features may cause the other features to be lessly emphasized. Re-weighting the distribution by dividing by the total number of features per category leads to the normalized version in Figure 6.5. Note as a reference, if we calculate the distribution over the total set of 384 features, it will be evenly distributed, each category representing 20% of the total features. So Figure 6.5 magnifying RMS-energy implies that RMS-energy was considered important ($> 20\%$) in the top 75 features. Moreover, the RMS-energy features are treated more importantly for recognizing thrill stress (36%), than for recognizing work load stress (31%). This follows the original observation that screaming happens occasionally in thrill stress such that RMS-energy could be a more useful feature. In addition, the work load stress model emphasizes more on the MFCCs (23%) than the thrill stress model (12%) stress. Also note in Figure 6.5 the pitch (F0) featues were evaluated less important in comparison to the others.

6.5 Conclusion

This chapter describes two applications where the triggers from the physical body correlating with vocal expression are adapted to recognize mental states. The first one is sleep deprivation, which often places important role on emotional functioning. Sleep deprived causes increasing negative emotions, and in reverse mood disorders such as depression often triggers sleep deprivation. So we studied the vocal expression of emotion on sleep deprived subjects to understand the relationship. Sleep deprivation indicated decreases in pitch, bark energy (intensity) in certain high frequency bands, and vocal sharpness (reduction in high frequency



(a) (b)

Figure 6.5: "Normalized" Distribution of feature categories in the top 75 features chosen by linear SVM for (a) thrill stress and (b) work load stress.

bands $>$ 1000 Hz). The second application enables a detailed analysis on stress detection via voice. Accuracy as high as 95% was achieved, with the help of user-based feature normalization and test speech of duration $>$ 30 seconds. Nonetheless, the high accuracy was derived from a dataset with a thrill factor (roller coaster rides) where there's a prominent high RMS energy in stressed speech. Whereas the other dataset with high mental load displays more difficulty for detection. The accuracy was 85%. Although the RMS-energy was not as prominent, it shows that the energy distribution over frequency spectrum is important contributor to the accuracy. Putting together the work, Chapter 7 will conclude the thesis by providing an implementation of the speech analysis library to run efficiently on mobile phones.

Chapter 7

A Speech Analysis Library on Mobile Phones

The human voice encodes a wealth of information about emotion, mood, stress, and mental state. With mobile phones this information is potentially available to a host of applications and can enable richer, more appropriate, and more satisfying human-computer interaction. In this chapter we describe the AMMON (Affective and Mental health MONitor) library, a low footprint C library designed for widely available phones as an enabler of these applications. The library incorporates both core features for emotion recognition (from the Interspeech 2009 Emotion recognition challenge), and the most important features for mental health analysis (glottal timing features). To comfortably run the library on feature phones (the most widely-used class of phones today), we implemented the routines in fixed-point arithmetic, and minimized computational and memory footprint. On identical test data, emotion classification accuracy was indistinguishable from a state-of-the-art reference system running on a PC, achieving 75% accuracy on two-class emotion classification tasks. The library uses 30% of real-time on a 1GHz processor during emotion recognition and 70% during stress and mental health analysis.

7.1 Introduction

Were one to design an ideal device for affect/mental health monitoring by voice, it would probably look a lot like a cell phone. A small, handheld device that is regularly used for other voice-based tasks (i.e., calling others), and which helps to distinguish a particular user's voice from those around them (phones have a variety of noise-canceling and directional features built in). What is lacking for developers are the speech features needed for applications or better still, binary or real values that denote emotion or depression strengths - i.e., emotion classifier outputs.

While smartphones are gaining market share daily, "feature phones" are still the dominant

devices in the hands of users, and will be for some time to come[1]. Furthermore, the downside of the very powerful (1GHz) processors and large memory on smartphones is that they can run batteries into the ground much faster than older-model phones. The bright side is that clock speed can be curtailed (or the processor idled) so that power consumption scales approximately linearly with the amount of computation to be done. So to be feasible on feature phones and to be practical on smartphones, voice analysis must have a small computational footprint in both CPU time and memory. This is a primary goal in design of the AMMON library. The other goal is to ensure that analysis on the mobile library is as accurate as on a PC.

We have developed the AMMON library (Affective and Mental-health MONitor) to meet these goals. The library computes a rich set of prosodic and spectral features which support emotion recognition with state-of-the-art accuracy of around 70% based on the Interspeech 2009 emotion recognition reference dataset and feature set [79]. AMMON also includes features to describe glottal vibrational cycles, a promising feature for monitoring depression. Moore et al. [59] showed that linear classifiers using a combination of these features can distinguish depressed and healthy subjects with 90% accuracy. This implies that the glottal activities in speech production can be greatly affected by mental illness, a good indicator of physical change induced by mental states. We hypothesize that the glottal features can improve stress detection as well. In analogy, mental stress often manifests physical response in the autonomic nervous system (cf. heart rate) [13]. So glottal features, indicating physical change in glottal muscles, may also respond to the autonomic nervous system. Our experiments showed that the glottal features indeed improved the classification accuracy. AMMON was written in C and we developed it based on an existing mobile front-end (ETSI advanced extended front-end [24]).

Most feature phones today lack floating-point hardware. Feature phones have clock speeds in the 150 to 400 MHz range. The toolkit we describe is intended to run on these feature phones. So far we have demonstrated 30% of real-time performance on 1GHz ARM devices and 45-65% of real-time on 600 MHz ARM devices, which should be close to real-time on 300-400MHz ARM devices[2]. This should be acceptable for monitoring applications. e.g., for 200 MHz or slower devices, some blocks of the input can be dropped - the features are all block-based and therefore discarding data reduces data volume but not accuracy.

The rest of this chapter is structured as follows: Section 7.2 describes the related work. Section 7.3 presents the speech analysis library, including the voice feature set and the effort to improve efficiency. It includes the benchmarked performance running the library on mobile phones. Section 7.6 demonstrates the effectiveness of the features by applying them on an emotional speech dataset and a dataset of mental stress. The result matches the state-of-the-art result. Section 7.8 concludes the chapter and discusses future work.

[1]Globally it seems unlikely that smartphones will ever dominate the market in developing countries

[2]The toolkit is not yet fully optimized, and e.g., does not yet use ARM intrinsics, so this figure should decrease.

7.2 Related Work

In this section, we describe related work for discussing our contribution relative to these works.

Emotion Recognition

Automatic emotion recognition has a long history with speech processing [43]. An extremely useful landmark was the Interspeech Emotion Challenge 2009 [79]. This challenge included standard dataset of emotion-tagged speech, and a "baseline" implementation of feature analysis, known as openSMILE. Surprisingly, while some more sophisticated algorithms improved on the baseline system, the improvements were very small, and it is fair to say that the baseline implementations achieved state-of-the-art performance. Since the baseline code was publicly distributed, we were able to compare our own implementation against it. A second surprising result was that use of segmental features (phone-level features) did not improve on "suprasegmental" primitive features (MFCCs, pitch, dynamics, energy). This may change in the future, but for now it means that state-of-the-art emotion recognition is much simpler than phonetic analysis. Expressed in terms of speech recognition components, that means that fully-accurate emotion analysis requires only the front-end of a speech recognizer and not the (memory and compute-intensive) acoustic model or later stages.

As a quick reference, the state-of-the-art recognition accuracy is about 70% for five-way classification of emotions (happy, sad, fear, anger and neutral) in a standard database with actors expressing emotions [70]. On the other hand, for the Interspeech challenge, naturalistic transcripts were recorded and hand annotated. Accuracy was only 70% for two-way classification [79].

Speech Patterns in Depression

In a remarkable study, Moore et al. [59] showed that feature analysis can separate a control group of healthy subjects from a group of depressed patients with 90% accuracy. It relied most strongly on *glottal* features which are not part of most low-level speech analysis systems. We included these features in AMMON to support mental health analysis. As shown in Chapter 5, this lead to improvement in the accuracy of stress detection and recognition of cases involved with speech pathology.

Voice Analysis Library on Mobile Phones

There has been a lot of activity lately on toolkits for mobile applications, including speech analysis and machine learning. SoundSense applied voice analysis to infer activities happening around a user, including driving, listening to music, and speaking [50]. SoundSense extracted a set of low-computation features and fed them to the J48 decision tree algorithm running locally on the phones. The features included zero-crossing rates, low energy frame

rates, and other spectral features. By comparison, AMMON extracts affective features, including pitch and information about glottal vibrational cycles. It supports linear classification in real-time since the Interspeech challenge showed there to be little advantage in use of other classifiers for emotion recognition. EmotionSense is an emotion recognition library on mobile phones for psychological studies [70]. EmotionSense does not infer emotions locally on the phones, but it ships the computation to the cloud. This imposes significant penalties in terms of privacy, need for access to the network, centralized server costs etc.

Choudhury et al. developed the Sociometer [14], a framework that infers colocation and conversation networks from voice data. The work focused primarily on the social ties by analyzing the turn-taking by energy and voiced/non-voiced features in face-to-face conversations. Our work instead provides rich and multidimensional analysis of emotion during conversation which can support a variety of social applications.

7.3 Speech Analysis Library

In this section, we provide an overview of the AMMON architecture. We describe each architectural component in turn, as those illustrated in Figure 7.1.

Preprocessing

Sound processing starts with segmenting the audio stream from the microphone into frames with fixed duration (25 ms) and fixed stepping duration (10 ms). Not all frames are considered for further processing. The module performs *voice activity detection* for the non-speech frame dropping.

Feature Extraction

The selection of features is critical for building a robust classifier. We built a feature set based on the features defined in Interspeech challenge. It includes static feature vectors derived by projecting *low-level descriptors* (LLDs, in the form of signal waveforms) such as pitch and energy by descriptive statistical *functionals* such as lower order moments (mean, standard deviation etc). The static feature vectors were effective, which is probably justified by the supra-segmental nature of occurring with respect to the emotional content in speech [81].

Table 7.1 lists the LLDs in the categories of prosody, voice quality and spectral domains: zero-crossing rate (ZCR), root-mean-square (RMS) frame energy, pitch (F0), harmonics-to-noise ratio (HNR), mel-frequency cepstral coefficients (MFCC) 1-12. Moreover, to each of these LLDs, the delta coefficients are additionally computed. The features in Table 7.1 are the same as in Table 5.3 but repeated here for clarification.

In addition to the standardized set defined in the Interspeech challenge (16 LLDs), we include glottal timings in the LLDs, which had great success in measuring mental health [59]. As illustrated in Figure 3.2, a glottal (flow) vibrational cycle is characterized by the time

that the glottis is open (O) (with air flowing between vocal folds), and the time the glottis is closed (C). In addition, an open phase can be further broken down into opening (OP) and closing (CP) phases. If there is a sudden change in airflow (i.e., shorter open and close phases), it produces more high frequency and the voice therefore sounds more *jagged*, other than *soft*. To capture it, AMMON calculates the above 4 durations of each cycle and 5 ratios of the closing to the opening phase ($rCPOP$), the open phase to the total cycle ($rOTC$), the closed phase to the total cycle ($rCTC$), the opening to the open phase ($rOPO$), and the closing to the open phase ($rCPO$). In summary, there were a total of 9 glottal timing-based LLDs included.

Then, AMMON segments the LLDs into windows, meaningful units for the modeling of feature vectors. A window can either be a turn or a fixed duration. Finally, it calculates 9 functionals from each window, including mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position and range. In the end, a feature vector contains $25 * 2 * 9 = 450$ attributes.

Affect and Mental Health Recognition

AMMON uses linear Support Vector Machines (SVM) to recognize emotions based on the feature vectors (projecting LLDs by functionals). Linear SVM is currently a dominantly used mechanism for recognition emotions. In addition, doing prediction with a linear SVM is rather efficient, which is suitable to run on the phones. Training models is more expensive, but this can be done off-line (not on the phones).

Implementation

We implemented AMMON in C, which can be deployed to both feature phones (e.g., Symbian) and smart phones (e.g., Android). In the work we developed AMMON with Android NDK, where we can turn off the floating-point support in compile time to test the scenarios of feature phones. The Android platform has a dominant market share and is likely to lead the market in the near future. In addition, it supports implementation in both Java and C, which is convenient for re-using existing signal processing libraries written in C. For pre-processing, we leveraged the existing function of voice activity detection in ETSI front-end library [24].

AMMON for Emotion Analysis

We developed AMMON by extending an ETSI (European Tele-communications Standards Institute) front-end feature extraction library [24]. The original purpose of the front-end was for local extraction of features on phones for remote speech recognition. Nonetheless, the front-end was useful for AMMON because (1) The ETSI front-end was already extracting some of the LLDs, such as energy, F0 and MFCC. We can re-use the code. (2) The front-end was equipped with noise-reduction routines, designed especially for the case of background

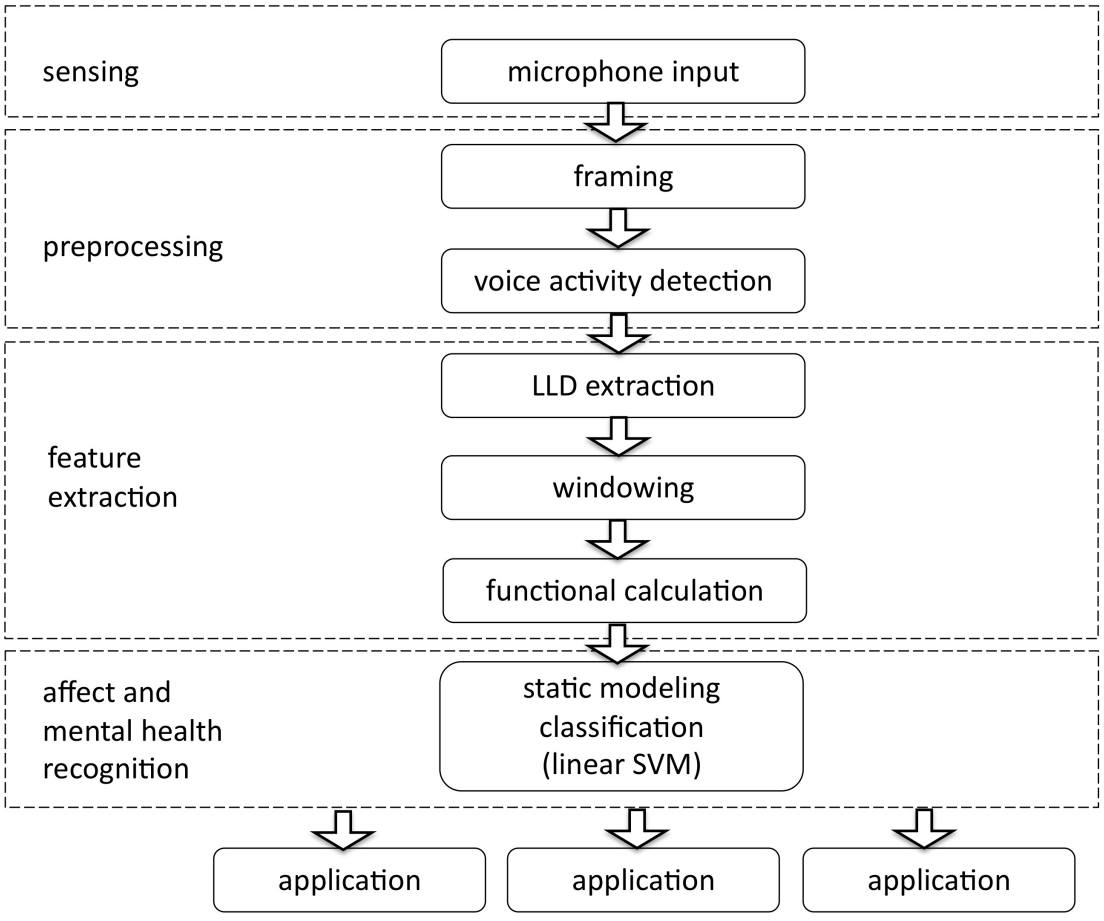


Figure 7.1: The AMMON Architecture

Table 7.1: The AMMON feature set, computed by applying functionals on LLD waveforms.

LLDs	functionals
(Δ)ZCR (Δ)RMS energy (Δ)F0 (Δ)HNR (Δ)MFCC 1-12 (Δ)Glottal timings x 9 [a]	mean, standard deviation, kurtosis, skewness, minimum, maximum range, rel. position

[a] AMMON includes glottal timings for mental health anlysis, whereas the rest of LLDs are sufficient for emotion analysis.

noise while using mobile phones. It will make the features more reliable. (3) The library had routines for voice activity detection, which can be used for frame admission control. Non-speech frames will not be considered for further processing. (4) The ETSI library was implemented purely with fixed-point arithmetics, ensuring the library to run efficiently on feature phones without floating-point hardware.

We ported the ETSI front-end to the Android platform with Android NDK. , which is a GNU C/C++ based compiler and tool-chains that can generate native ARM binaries. An Android application built in Java can evoke AMMON with JNI (Java native interface) by passing a raw file of voice recording, and AMMON will return the affective information, e.g., emotion classes and the confidence.

After porting the front-end to Android, we implemented routines for the remaining LLDs (ZCR, HNR and glottal timings), using fixed-point arithmetic in particular. Zero-crossing rate (ZCR) was straightforward. Harmonic-to-noise ratio was implemented by the calling autocorrelation function (ACF) provided by the ETSI library ($HNR = 10 \log ACF(T_0)/(ACF(0) - ACF(T_0))$, $T_0 = fs/F_0$). Finally, in terms glottal timings, it was computationally more expensive to extract them than other LLDs. Therefore, we re-designed the algorithm with details described in Section 7.4.

Implementing Functionals

Making a reliable estimate arguably requires as much data processing as possible. This means that, we have to calculate functionals over a large window of LLDs. Given the limited memory available on feature phones, it is not practical to buffer full conversational turns. AMMON should calculate the functionals over time without having to save the value at every sample. Therefore, we implemented an online, buffer-free algorithm to calculate the functionals (pseudo code can be found in [46]).

Given a new sample of an LLD, only the mean and the first to forth moments are updated, implying constant space per LLD. Then, it can calculate an up-to-date functional with the moments. For example, we can calculate variance with the second moment. We can also obtain kurtosis with the forth and the second moments. In terms of computation, each update and computation of functionals takes only constant time.

7.4 Extracting Glottal Timings

It is computationally more expensive to extract glottal timings than the other LLDs. So we implemented the routine with special care, including algorithimic improvement and code optimization. Following the algorithm described in Section 5.2, we analyzed the bottleneck by profiling. The most dominant part is formant tracking, which requires for every sample, estimating LPC (linear predictive coding) polynomials and *solving roots* of each polynomial to determine formant frequencies. The rationale for computing formants is as follows: in general, the production of speech sounds can be described as the interaction of *glottal ex-*

citation with the resonance of a *vocal tract*. Due to these interactions, estimating glottal vibrational cycles from the output signal becomes more difficult, so we seek regions where these components interact minimally. This part helps identify the closed-phases (C) of glottal vibrational cycles. When the glottis is closed, vocal tract is the only mechanism in effect in speech production. So formant frequencies should be stationary within short windows. [3]

Solving roots of polynomials is expensive, which involves eigensolving the companion matrix of a polynomial. Even worse, the root solving is evoked frequently, in windows advancing in every sample. But we can leverage the property in a way to avoid constant eigensolving or "finding" roots from scratch. We can "track" roots instead. The idea is as follows. Because the sequential LPC polynomials are computed from adjacent windows that share a majority of speech samples, these LPC polynomials – and their roots – should not change a great deal between any two adjacent windows. Thus, we applied Newton-Raphson iteration to track roots of the current polynomial starting from the roots of the previous one. The Newton iteration is much cheaper. However, it does not guarantee to find all the roots.

This leads to an algorithm that we have to strike a balance between the Newton's method and eigensolving. The former one is faster but does not guarantee a result. The latter is slower but can be counted on to find correct answers. If Newton's method fails, we resort to the eigensolver, which always finds a correct answer but much more expensive. Moreover, we applied several techniques to increase the probability of success in root tracking with Newton's iteration , but did it within the time budget that the Newton iteration gained over the eigensolver. e.g., subdivision between polynomials or kicking roots off the real or imaginary axis, as described in the following.

- Roots in sequential windows are significantly different from each other, and that Newton's method with two previous roots may incorrectly converge to the same root, i.e., a root is not found. Therefore, we can try to solve for the roots of a *linear interpolation* of the previous window's polynomial and the current window's polynomial, then use those intermediate roots to offer a better initial iteration point for subsequent iterations. This *subdivision* procedure can be invoked recursively to further improve the probability of successful root-tracking at the cost of greater computation time. The success of Newton's method is worthwhile for spending additional trials with the Newton's iteration. We know if we successfully find roots, the much more expensive eigensolver will not be invoked.

- Since all polynomials we deal with have entirely real coefficients, the interpolated intermediate polynomials are also real. By perturbing the polynomial of the previous frame with a small amount of complex coefficient, the interpolated polynomials become

[3] It is not necessary to do root finding in order to extract an inverse transfer function in order to remove the vocal tract effect. Transforming the LPC polynomial to reflection coefficients is an alternative. Similar to formants, the reflection coefficients interpolate and smooth well. The transformation requires very little computation, and has often been done in fixed point arithmetic. Once the coefficients are computed they can be used directly to form the inverse filter [53]. Nonetheless, formants were studied and proven effective in the detection of affective states. [5]

complex so the intermediate roots will be complex and less likely to collapse during Newton's method.

- We verify that the tracked roots are correct by multiplying the root binomials (i.e., $(x - a - bi)(x - a + bi)$) together to obtain the original polynomial. In many cases, this can be much faster to compute.
- If a polynomial has a multiple root, its derivative also shares that root. Otherwise, the found double roots are incorrect; a root is not found. We can use this as an early-stage verification criterium. The criterium of the previous item requires all roots for verification.
- When moving from a function with a double root to one with a root pair, we need our iteration to produce two distinct roots from the two identical roots in the previous window. To deal with these issues, we order the roots by their magnitude, and add a small real and imaginary value at each step to the roots at odd indices, and subtract that same value at each step from the roots at even indices. This has the effect of "kicking" root pairs off the real or imaginary axis, as well as pushing the roots in root pairs or double roots in opposite directions, allowing them to converge to different values over additional iterations.
- For polynomials of order=2, 3, and 4, we applied the closed form solution.

It should be noted that these techniques have tunable parameters - we can choose the maximum number of Newton's method iterations allowed, the small value to add or subtract at each step to deal with splitting double roots, and the number of subdivisions allowed. Allowing more iterations or subdivisions increases the probability of successful root tracking, but takes additional CPU time. There is a point at which it is more efficient to simply revert back to computing roots with an eigensolver. Implementors should find appropriate tuning parameters for their problem domain.

We implemented the Newton method ourselves, but for eigensolving, we applied CLAPACK (f2c'ed version of LAPACK, linear algebra package) [15]. However, the package was written in floating point. It is our future work to replace it with a fixed-point eigensolver, making AMMON truly applicable to feature phones (the remaining modules were implemented in fixed-point).

In addition to solving roots of the polynomials, the estimation of polynomials is also required to run in every sample. It involves using autocorrelation to construct *Toeplitz* matrices out of adjacent windows that share a majority of samples. We implemented the autocorrelation method in a way that the Toeplitz matrix is revised incrementally with each sample shift. This reduces the running time from quadratic to linear time.

The other bottleneck is the Fast Fourier Transform, which is evoked in every sample to calculate the phase change and locate the maximum excitation (the boundary between opening (OP) and closing (CP) phases). We optimized the part with a piece of ARM optimized assembly code.

7.5 Performance Evaluation

We evaluated the implementation in terms of its computational efficiency. And we break down the evaluation based on emotion recognition and mental health analysis.

Emotion Analysis: Compare with openSMILE

First, we compared AMMON with the open source toolkit openSMILE used in the Interspeech challenge. For emotion recognition, we excluded the computation of glottal timings. Since AMMON has voice activity detection and noise suppression modules whereas openSMILE does not, we also intentionally turned them off for fair comparison.

We compared them on phones with and without floating-point support. As such, we can understand whether AMMON can run emotion analysis comfortably on feature phones. As a benchmark, we made use of an emotional speech database (details in Section 7.6). There were 298 clips in the dataset, each with 10-60 seconds long. The benchmark was run on a Google Nexus One phone (1GHz Snapdragon CPU with floating-point hardware), where the floating-point was turned off to simulate the case of feature phones.

Table 7.2 shows that when the floating-point support was turned on (through compiler flags), AMMON ran comparably with openSMILE. OpenSMILE ran only slightly faster (17% of real time (xRT)) than AMMON (18% xRT), which supposedly was spending extra effort in fixed-point arithmetic. However, when the floating-point support was turned off, the fixed-point implementation paid off. OpenSMILE ran much slower (53% xRT), whereas AMMON stays the same (18% xRT). This implies that AMMON is more efficient than openSMILE on feature phones.

AMMON has additional voice activity detection and noise suppression modules. Finally, we turned on the modules of voice activity detection and noise suppression. AMMON ran in a total of 29% of real time. We also benchmarked the performance on two slower phones with 600 MHz CPU (Motorola Droid with TI OMAP 3430 CPU and HTC Aria with Qualcomm MSM7227 CPU). AMMON ran in a total of 45% of real time on Motorola Droid and 64 % of real time on HTC Aria. That said, AMMON should run emotion analysis in real time on 300-400 MHz feature phones [4].

Mental Health Analysis: Extract Additional Glottal Timings

Since the root solving module was not revised to fixed-point yet, we turned on the floating point support for the extraction of glottal features. The modification in Section 7.4 significantly improved the performance of glottal extraction, as illustrated in Figure 7.2. For root solving, we managed to reduce its running time by 68% (reduced to 1/3). Table 7.5 shows the breakdown of improvement by the order of polynomials. Table 7.5 shows that the additional improvement of the algorithm improves the reduced ratio from 48% to 68%. It is worthwhile

[4]Extrapolation based on by CPU frequency scaling may not hold due to factors such as slower and smaller memory systems, so benchmark will be made on more phones as a future work.

Table 7.2: Computational efficiency of AMMON. The running time are displayed in the percentage of real time (xRT) on a 1GHz phone.

toolkit	floating point ON	OFF
openSMILE	0.17 xRT	0.53 xRT
AMMON w/o Glottal Timings, VAD, and Noise Supr.	0.18 xRT	0.18 xRT

noting that in the yet improved version, the success rate of Newton's method dropped as the order the polynomial increases (< 0.5), due to the fact that roots are more likely to collapse with others as there are more in the space. By subdivision (with the overhead of additional trials using Newton's iteration), the success rate increased significantly to more than 0.8. The total running time improved consequently. N/A in Table 7.5 means that a closed form solution was used for polynomials of order 2-4.

Using assembly code for FFT reduced its running time by 85%. Incremental revision of the Toeplitz also reduced its running time in two orders of magnitude , showing that roots tracking with Newton method can efficiently replace constant root finding with eigensolvers.

As a whole, the new glottal extraction algorithm ran from 105% of real time to 41% of real time, a 61% decrease. This adds up the AMMON computation time for mental health analysis to 70% of real time (was 133% of real time). That said, doing mental health analysis on phones are more expensive. AMMON can run mental health analysis on smart phones in real time, but about 2 times slower than real time over the feature phones. Nonetheless, Chapter 5 and [59] showed that the glottal features were indeed valuable, although it is computationally expensive. It significantly increased recognition accuracy for mental health analysis.

7.6 Feature Evaluation

In this section, we demonstrate the effectiveness of AMMON in recognizing emotions. We show that using the feature set extracted in AMMON, it recognizes emotions in state-of-the-art accuracy.

Emotion Recognition

To evaluate the features in emotion analysis, we could have chosen the FAU Aibo dataset used in Emotional challenge 2009, where the recognition accuracy is available as a baseline for comparison. Nonetheless, given the goal to recognizing emotions in everyday conversations, the Aibo dataset is not entirely suitable. First, the Aibo dataset is in German, not in English. It is known that emotional expressions vary across languages and cultures, so a model

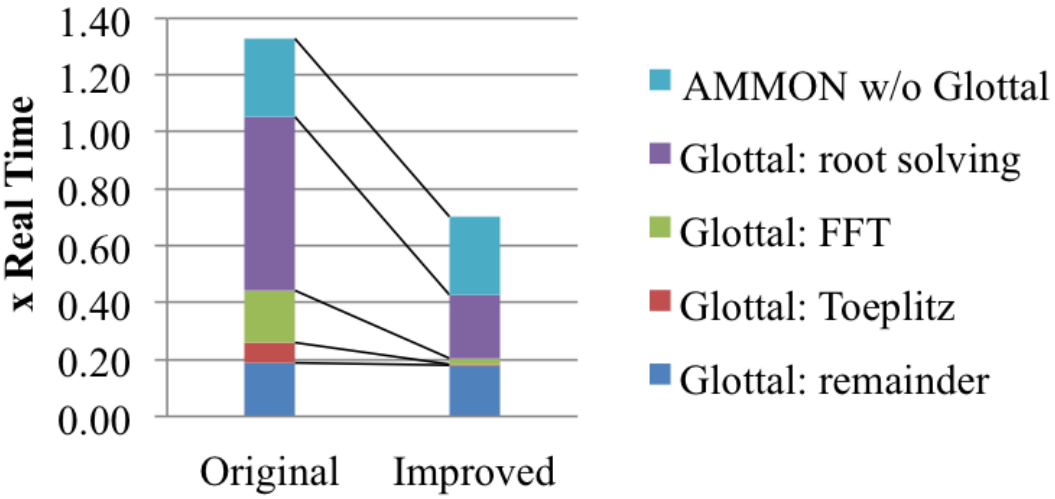


Figure 7.2: The breakdown of AMMON running time. The improvement of glottal extraction makes AMMON run 70% of real time on a 1GHz smartphone.

trained in German may not be applicable to conversations in English. In addition, emotions happened in the database were mostly non-prototypical and subtle (empathy), making it insufficient to support most of the applications that require information of prototypical emotions (i.e., sad, happy, etc).

Therefore, we chose the Belfast Naturalistic Database [23]. The dataset is in English, covering a wide range of emotional states that occur in everyday interactions, as well as prototypical examples of emotion such as full-blown anger (Table 3.1). The Belfast database consists 298 audiovisual clips from 125 speakers (31 males and 94 females). These clips were collected from a variety of television programs and studio-recorded conversations. Clips range from 10 to 60 seconds in length.

The Belfast database were labeled by multiple raters. Each clip was labeled by 3 most visible emotions (Table 3.1) and the intensity (weak, medium and strong). We aggregated the labels in terms of voting and strength (Section 3.3).

Recognizing Positive v.s. Negative Emotional Clips

We performed a 2-way classification task to separate clips with positive emotions from those with negative emotions. A clip is considered positive if none of the aggregated label has negative valence, and vice versa for negative clip. For the case that some clips were labeled with both positive and negative valence, we excluded them. The task is potentially useful for most applications, that the information whether users are in positive or negative mood is of interest.

Table 7.3: Performance Comparison in the Recognition of Positive v.s. Negative Emotional Clips. We also list the F-measures for both clasess (data size of classes: 112/133).

Feature Set	F-Measures	ROC Area	Accuracy
openSMILE	0.778/0.727	0.753	75.51%
AMMON	0.776/0.73	0.752	75.51%

The evaluation worked by comparing the performance of AMMON with that of the openSMILE toolkit. First, we applied AMMON to extract a feature vector for each clip. Note glottal timings were not extracted since this is for emotion analysis. Then, we fed the feature vectors to SVM, a widely used method in emotion recognition (regularized linear SVM, C=0.06, features scaled, 5-fold cross validation). We applied the same procedure to openSMILE: extracting feature vectors and performing classification. In the end, we classified between 112 positive valence clips (class 1) and 133 negative valence clips (class 2). Table 7.3 shows that AMMON had a comparable result to openSMILE, achieving 75% of accuracy and 0.75 ROC area. The accuracy is at the same level as the result of the Interspeech challenge, at 70% level classifying 2 emotions in a naturalistic database. The experiment implies that AMMON can support emotion analysis with the same level of accuracy as the PC reference system, i.e., openSMILE.

Identification of Prototypical Emotions

We proceeded to the next task, identifying prototypical emotions from the others. This task is useful for applications that require spotting emotions in history. We choose three emotions as the target: anger, sadness and happiness.

This was a 2-way classification problem, where clips with the target emotion were put in class 1 and the remainder clips were put to class 2. Nonetheless, this led to imbalanced partition, about 1:3 ratio between classes (Table 7.4). Identifying prototypical emotions in this setup became essentially more difficult than the one described in the previous section. We did a similar experiment, applying both openSMILE and AMMON to the clips and applied linear SVM (regularized SVM 5 fold cross-validation, C=0.5, features scaled).

Table 7.4 lists the recognition result. AMMON performed comparably to openSMILE. Since it is a more difficult task, the recognition result was not as good as the first task. Both toolkits achieved around 75% accuracy and 0.66 ROC area in the cases of anger and happiness. Happiness is usually considered more difficult to differentiate. Our result showed the same trend that both toolkits decreased to 68% accuracy.

Table 7.4: Performance Comparison in the Identification of Prototypical Emotions. We did 2-way classification for identifying anger from the remainder clips. The classes are imbalanced, with 87/200 instances. In addition, the same setup was repeated for identifying sadness and happiness.

	Anger v.s. Remainder (87/200)		
Feature Set	F-Measures	ROC Area	Accuracy
openSMILE	0.532/0.823	0.669	74.31%
AMMON	0.517/0.826	0.662	74.56%
	Sadness v.s. Remainder (78/209)		
Feature Set	F-Measures	ROC Area	Accuracy
openSMILE	0.493/0.84	0.655	75.69%
AMMON	0.515/0.849	0.669	77.00%
	Happiness v.s. Remainder (87/199)		
Feature Set	F-Measures	ROC Area	Accuracy
openSMILE	0.497/0.77	0.636	68.40%
AMMON	0.408/0.776	0.588	67.48%

7.7 Future Work

For any of these classifiers to be used in practice (Section 3.5, 6.4 and 7.6), it would be essential to have an "I don't know" category. When viewing these problems as a detection task, sometimes it is necessary to avoid false-positive. This is a future work, which can be modeled with a confidence measure, or a rejection threshold.

7.8 Conclusion

The pervasiveness of mobile phones opens up an opportunity for improving our psychological well-being, and it scales from individuals to the mass public. Emotion monitor can raise individual awareness and contribute behavior change. A mental health tracker can detect early stage problems, measure health trend of the public, and promote public health. Therefore, In this chapter we describe AMMON, an affective and mental health monitor. AMMON was designed to work on feature phones, so that most people can have access to this service. We were able to prove that the features extracted by AMMON were as effective as those by reference systems on PC. AMMON can recognize emotions in state-of-the-art

accuracy. In addition, we are investigating ways to replace the floating-point eigensolver library with a fixed-point version. However, re-inventing a fixed-point eigensolving library is not trivial. We are also considering the Jenkins-Traub algorithm to replace the companion matrix/eigensolving method for root solving.

Table 7.5: The improvement of running time (reduced by 68%) using Newton methods for root solving, with breakdown by polynomial orders. The "Newtons' success" row represents the percentage of polynomials at which the Newton's method successfully found the roots, so eigensolver was not required. The "Newton's Iters" row represents the number of times the Newton's iteration was called. The number is higher in the improved method because subdivision (of polynomials between the polynomials of the previous frame and the current frame) was used, and the success rate was improved. N/A means a closed form solution was used.

Polynomial Order	2	3	4	5	6	7	8	9	10	11	12	Total
	Stats											
Counts	978	2834	5135	4607	2213	998	1022	774	1078	650	2187	
	Original Performance											
Eigensolver (ms)	5.49	82.99	229.88	295.40	215.30	148.84	167.35	160.88	301.22	226.01	886.55	2823.35
	With Newton's Method but no additional improvements											
Newton's Iters	978	2834	5135	4607	2213	998	1022	774	1078	650	2187	
Newton's Success	0.76	0.84	0.66	0.75	0.62	0.63	0.50	0.50	0.42	0.49	0.51	
Improved (ms)	3.93	22.23	98.44	96.19	98.39	53.60	88.38	88.84	199.79	117.53	458.32	1417.04
Reduced Ratio	0.30	0.68	0.54	0.66	0.53	0.60	0.45	0.46	0.34	0.45	0.49	0.48
	With Newton's Method AND additional improvements											
Newton's Iters	N/A	N/A	N/A	10515	6795	3073	3969	3074	4990	2712	8387	
Newton's Success	N/A	N/A	N/A	0.92	0.87	0.88	0.84	0.82	0.81	0.85	0.85	
Improved (ms)	1.19	16.59	31.95	74.21	64.70	35.86	66.55	55.55	109.85	62.77	262.19	884.85
Reduced Ratio	0.78	0.80	0.86	0.75	0.70	0.76	0.60	0.65	0.64	0.72	0.70	0.68

Chapter 8

Conclusion

My research in human-computer interaction focuses on "user-centric sensing", which applies innovative sensing techniques to infer the state of a user. Using sensors to gather real-time information about people's activities in everyday situations can enable a new host of applications and new user experiences. My research has focused on applications and techniques for user-centric sensing in a range of application domains, from entertainment [9] and personal exercise [8] to critical healthcare problems [10, 11].

User-centric sensing is enabled by the dramatic success and proliferation of programmable mobile devices that make a wide range of sensor data and contextual information accessible to applications. Inferring real-time user state from mobile (sensor) data such as speech, accelerometer data, images, GPS, calendar, email, etc., opens up tremendous research opportunities in pursuing user-centric sensing. As devices are equipped with ever more powerful computing capabilities, applications can continuously monitor data, aggregate it, recognize patterns of interest, and create accurate user models, changing ordinary phones into "cognitive phones".

The human voice encodes a wealth of affective information, as well as indicators of early stage mental illness. Coupled with the pervasiveness of mobile phones, the human voice becomes the most accessible and unobtrusive means to monitor mental health of the general public. We believe that continuous capturing voice in this way can provide per-patient baseline data and enable qualitatively better diagnosis. This serves as a starting point for cognitive phones, phones with cognitive perception that look after us to promote mental health.

The thesis summarizes my investigation on the speech analysis methodologies towards unobtrusive mental health monitoring, involving the recognition of emotions, stress, associated abnormal speaking styles, etc. The research process involves multidisciplinary understanding and applications of psychology, machine learning, speech processing, and human computer interaction. The research is not trivial, but it is our hope that the research community in computer science will pay more attention to the mental health area, in which there is significant problems but highly undertreated.

Again, I strongly believe that machine perception will be the key enabling technology

for the next big wave of applications. By providing an accurate user model, it will enable researchers and developers to create a new host of intelligent applications that act according to the state of a user. By accurate user modeling I mean not merely activity recognition but also social networking, physical, cognitive and affective states. I will continue the focus on the healthcare area, leveraging sensors on mobile devices in particular, for user-centric sensing. This is a research area about "big data on mobile phones", where multi-sensor data stream are flowing to mobile phones 24/7 and there are millions of phones. In analogy, this can can be many orders of magnitude larger than the web data we see nowadays, which is discrete not continuous. Finally, by incorporating user-centric sensing into the design of user feedbacks based on behavior change theories and cognitive behavioral therapy (e.g., Appendix A), I can create a personal health assistant to address health problems.

Bibliography

[1] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision*. American Psychiatric Association, 2000.

[2] International Phonetic Association. "Handbook of the International Phonetic Association: a guide to the use of the international phonetic alphabet". In: Cambridge University Press, 1999. Chap. Phonetic description and the IPA chart.

[3] Paul Boersma and David Weenink. "Praat, a system for doing phonetics by computer". In: *Glot International* 5(9/10) (2001), pp. 341–345.

[4] Gloria J. Borden, Katherine S. Harris, and Lawrence J. Raphael. *Speech Science Primer: Physiology, Acoustics, and Perception of Speech*. Lippincott Williams & Wilkins; Fourth edition, 2002.

[5] Elif Bozkurt, Engin Erzin, Çiğdem Eroğlu Erdem, and A. Tanju Erdem. "Formant position based weighted spectral features for emotion recognition". In: *Speech Communication* 53 (2011), 11861197.

[6] V. Brèjard, A. Bonnet, and J.-L. Pedinielli. "Depressive Symptoms and Emotion in Adolescence: A Developmental-functionalist View". In: *European Psychiatry* 24 (2009), S620.

[7] Walter Bradford Cannon. *Bodily changes in pain, hunger, fear, and rage*. New York: Appleton-Century-Crofts, 1929.

[8] Keng-hao Chang, Mike Chen, and John Canny. "Tracking Free-Weight Exercises". In: *Ubiquitous Computing (Ubicomp)*. 2007.

[9] Keng-hao Chang, Jeffrey Hightower, and Branislav Kveton. "Inferring Identity Using Accelerometers in Television Remote Controls". In: *Proceedings of the International Conference on Pervasive Computing (Pervasive)*. 2009.

[10] Keng-hao Chang, Shih yen Liu, Hao hua Chu, Jane Hsu, Cheryl Chen, Tung yun Lin, Chieh yu Chen, and Polly Huang. "Diet-Aware dining table: observing dietary behaviors over tabletop surface". In: *Proceedings of the International Conference on Pervasive Computing (Pervasive)*. 2006.

[11] Keng-hao Chang, Drew Fisher, John F. Canny, and Bjoern Hartmann. "How's my Mood and Stress? An Efficient Speech Analysis Library for Unobtrusive Monitoringon Mobile Phones". In: *Proceedings of the 6th International ICST Conference on Body Area Networks (BodyNets, Invited paper)*. 2011.

[12] Nitesh V. Chawla, Nathalie Japkowicz, and Aleksander Koltcz. "Editorial: special issue on learning from imbalanced data sets". In: *ACM SIGKDD Explorations Newsletter* 6 (2004). Issue 1, pp. 1–6.

[13] Jongyoon Choi and Ricardo Gutierrez-Osuna. "Using Heart Rate Monitors to Detect Mental Stress". In: *IEEE Body Sensor Networks*. 2009.

[14] Tanzeem Choudhury and Alex Pentland. "The Sociometer: A Wearable Device for Understanding Human Networks". In: *Computer Supported Cooperative Work - Workshop on Ad hoc Communications and Collaboration in Ubiquitous Computing Environments*. 2002.

[15] *CLAPACK (f2c'ed version of LAPACK)*. `http://www.netlib.org/clapack/`, retreived in May 2011.

[16] Sheldon Cohen, Ronald C. Kessler, and Lynn Underwood Gordon. "Measuring stress: A guide for health and social scientists". In: New York: Oxford., 1995. Chap. Strategies for Measuring Stress in Studies of Psychiatric and Physical Disorders.

[17] Sunny Consolvo, David W. McDonald, Tammy Toscos, Mike Chen, Jon E. Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, Ian Smith, and James A. Landay. "Activity Sensing in the Wild: A Field Trial of UbiFit Garden". In: *Human Factors in Computing Systems*. 2008.

[18] Randolph R. Cornelius. *The Science of Emotion*. New Jersey: Prentice Hall., 1996.

[19] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G. Taylor. "Emotion recognition in human-computer interaction". In: *IEEE Signal Processing Magazine* 18(1) (2001), pp. 32–80.

[20] Charles Darwin. *The Expression of the Emotions in Man and Animals*. London: Harper-Collins, 1872.

[21] S. Das, A. Halder, P. Bhowmik, A. Chakraborty, A. Konar, and A. K. Nagar. "Voice and Facial Expression Based Classification of Emotion Using Linear Support Vector Machine". In: *2009 Second International Conference on Developments in eSystems Engineering*. 2009.

[22] David F. Dinges, Frances Pack, Katherine Williams, Kelly A. Gillen, John W. Powell, Geoffrey E. Ott, Caitlin Aptowicz, and Allan I.Pack. "Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4-5 hours per night." In: *Sleep* 20(4) (1997), pp. 267–77.

[23] Ellen Douglas-Cowie, Roddy Cowie, and M. Schroeder. "The description of naturally occurring emotional speech". In: *15th ICPhS*. 2003.

[24] *ES 202 212 Extended advanced front-end feature extraction algorithm V1.1.4*. Tech. rep. Source code retrievable through secure login on `http://www.etsi.org`, May 2011. ETSI, 2005.

[25] Florian Eyben, Martin Wöllmer, and Björn Schuller. "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor". In: *ACM Multimedia (MM)*. 2010.

[26] Gunnar Fant. *Acoustic theory of speech production*. The Hague: Mouton, 1960.

[27] Raul Fernandez. "A Computational Model for the Automatic Recognition of Affect in Speech". PhD thesis. MIT, 2004.

[28] Raul Fernandez and Ros W. Picard. "Classical and Novel Discriminant Features for Affect Recognition from Speech". In: *Interspeech 2005 - Eurospeech 9th European Conference on Speech Communication and Technology*. 2005.

[29] Alistair J. Flint, Sandra E. Black, Irene Campbell-Taylor, Gillian F. Gailey, and Carey Levinton. "Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression". In: *J Psychiatr Res* 27 (1993), pp. 309–319.

[30] PETER L. FRANZEN, GREG J. SIEGLE, and DANIEL J. BUYSSE. "Relationships between affect, vigilance, and sleepiness following sleep deprivation". In: *Sleep Research* 17(1) (2008), pp. 34–41.

[31] Liqin Fu, Xia Mao, and Lijiang Chen. "Speaker independent emotion recognition based on SVM/HMMS fusion system". In: *Audio, Language and Image Processing, 2008. ICALIP 2008*. 2008.

[32] Michael Fuchs, Matthias Froehlich, Bettina Hentschel, Ingo W. Stuermer, Eberhard Kruse, and Danie Knauft. "Predicting mutational change in the speaking voice of boys". In: *J Voice* 21 (2007), pp. 169–78.

[33] Barbara F. Fuller, Yoshiyuki Horii, and Douglas A. Conner. "Validity and reliability of nonverbal voice measures as indicators of stressor-provoked anxiety". In: *Res Nurs Health* 15 (1992), pp. 378–389.

[34] Michael Grimm and Kristian Kroschel. "Evaluation of natural emotions using self assessment manikins". In: *ASRU, IEEE*. 2005.

[35] Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan. "The Vera am Mittag German Audio-Visual Emotional Speech Database". In: *the IEEE International Conference on Multimedia and Expo (ICME)*. 2008.

[36] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. "The WEKA Data Mining Software: An Update". In: *SIGKDD Explorations* 11(1) (2009), pp. 10–18.

[37] John H. L. Hansen. *SUSAS*. Linguistic Data Consortium, Philadelphia. 1999.

[38] John H. L. Hansen and Sanjay Patil. *Speech under stress: Analysis, modeling and recognition*. Vol. 4343. SpringerLink, 2007, pp. 108–137.

[39] Jennifer Healey, Justin Seger, and Rosalind Picard. "Quantifying Driver Stress: Developing a System for Collecting and Processing Bio-Metric Signals in Natural Situations". In: *the Rocky Mountian Bio-Engineering Symposium*. 1999.

[40] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. *A Practical Guide to Support Vector Classification*. `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`, accessed April 2012. 2010.

[41] *INTERSPEECH 2012 Speaker Trait Challenge*. `http://emotion-research.net/sigs/speech-sig/is12-speaker-trait-challenge` accessed in March 2012. 2012.

[42] A Jackson, J Cavanagh, and J Scott. "A systematic review of manic and depressive prodromes." In: *J Affect Disord.* 74(3) (2003), pp. 209–217.

[43] Patrick N. Juslin and Klaus R. Scherer. "Vocal expression of affect". In: Oxford University Press, 2005. Chap. 3, pp. 65–135.

[44] Dacher Keltner and Paul Ekman. "Handbook of emotions". In: ed. by Michael Lewis and Jeannette M. Haviland-Jones. Guilford Publications, Inc, 2000. Chap. Facial Expression of Emotion, p. 151.

[45] J.A. Kientz, S. Boring, G.D. Abowd, and G.R. Hayes. "Abaris: Evaluating Automated Capture Applied to Structured Autism Interventions". In: *Ubicomp 2005: The 7th International Conference on Ubiquitous Computing*. 2005.

[46] Donald E. Knuth. "The Art of Computer Programming". In: Boston: Addison-Wesley, 1998, p. 232.

[47] Barry Kort, Rob Reilly, and Rosalind W. Picard. "An Affective Model of Interplay Between Emotions and Learning: Reengineering Educational Pedagogy Building a Learning Companion". In: *ICALT-2001 (International Conference on Advanced Learning Technologies)*. 2001.

[48] St. Kuny and H. H. Stassen. "Speaking Behavior and Voice Sound Characteristics in Depressive Patients During Recovery". In: *Journal of Psychiatric Research* 27(3) (1993), pp. 289–307.

[49] Mathias R. Lemke and A.C. Hesse. "Psychomotor Symptoms in Depression". In: *Letter to Am J Psychiatry* 155 (1998), pp. 709–709.

[50] Hong Lu, Wei Pan, Nicholas D. Lane, Tanzeem Choudhury, and Andrew T. Campbell. "SoundSense: Scalable Sound Sensing for People-Centric Sensing Applications on Mobile Phones". In: *Proc. of 7th ACM Conference on Mobile Systems, Applications, and Services (MobiSys '09)*. 2009.

[51] Anmol Madan, Ron Caneel, and Alex "Sandy" Pentland. "Voices of Attraction". In: *Augmented Cognition, (AugCog) HCI*. 2005.

[52] Hari Krishna Maganti, Stefan Scherer, and Günther Palm. "A Novel Feature for Emotion Recognition in Voice Based Applications". In: *Affective Computing and Intelligent Interaction*. 2007.

[53] J. Makhoul. "Linear Prediction: A tutorial review." In: *IEEE* 63(4) (1975), pp. 561–580.

[54] Eleanor L. McGlinchey, Lisa S. Talbot, Keng hao Chang, Katherine A. Kaplan, Ronald E. Dahl, and Allison G. Harvey. "The Effect of Sleep Deprivation on Vocal Expression of Emotion in Adolescents and Adults". In: *Sleep* 34(9) (2011), pp. 1233–41.

[55] Piet Mertens. "The Prosogram : Semi-Automatic Transcription of Prosody based on a Tonal Perception Models". In: *Speech Prosody*. Ed. by B. Bel and I. Marlien. 2004, pp. 23–26.

[56] Angeliki Metallinou, Sungbok Lee, and Shrikanth Narayanan. "Audio-Visual Emotion Recognition Using Gaussian Mixture Models for Face and Voice". In: *2008 Tenth IEEE International Symposium on Multimedia*. 2008.

[57] Lisette van der Molen, Maya van Rossum, Annemieke Ackerstaff, Ludi Smeele, Coen Rasch, and Frans Hilgers. "Pretreatment organ function in patients with advanced head and neck cancer: clinical outcome measures and patients' views". In: *BMC Ear Nose Throat Disorders* 9(10) (2009).

[58] Elliot Moore and Mark Clements. "Algorithm for Automatic Glottal Waveform Estimation without the Reliance on Precise Glottal Closure Information". In: *IEEE International Conference onAcoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04)*. 2004.

[59] Elliot Moore, Mark Clements, John Peifer, and Lydia Weisser. "Comparing objective feature statistics of speech for classifying clinical depression". In: *IEMBS*. 2004.

[60] Nelson Morgan, Eric Folser, and Nikki Mirghafori. "Speech Recognition Using On-line Estimation of Speaking Rate". In: *Eurospeech*. 1997.

[61] Nelson Morgan, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Adam Janin, Thilo Pfau, Elizabeth Shriberg, and Andreas Stolcke. "The Meeting Project at ICSI". In: *ICASSP*. 2003.

[62] Margaret Morris, Stephen S. Intille, and Jennifer S. Beaudin. "Embedded Assessment: Overcoming Barriers to Early Detection with Pervasive Computing". In: *Pervasive Computing*. 2005, pp. 333–346.

[63] Margaret E Morris, Qusai Kathawala, Todd K Leen, Ethan E Gorenstein, Farzin Guilak, Michael Labhard, and William Deleeuw. "Mobile Therapy: Case Study Evaluations of a Cell Phone Application for Emotional Self-Awareness". In: *J Med Internet Res*. 2010.

[64] Christopher J.L. Murray and Alan D. Lopez, eds. *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries and risk factors in 1990 and projected to 2020.* Harvard University Press, 1996.

[65] A. Nilsonne. "Speech characteristics as indicators of depressive illness". In: *Acta Psychiatrica Scandinavica* 77 (1988), pp. 253–263.

[66] World Health Organization. *Depression; what is depression?* Available at `http://www.who.int/mental_health/management/depression/definition/en/`. Accessed on Oct 2009.

[67] Alexander Pak and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).* 2010.

[68] Rosalind Picard. *Affective Computing.* Tech. rep. 321. MIT, 1995.

[69] NB Pinto and IR Titze. "Unification of perturbation measures in speech signals". In: *The Journal of the Acoustical Society of America* 87(3) (1990), pp. 1278–89.

[70] Kiran K. Rachuri, Peter J. Rentfrow, Mirco Musolesi, Chris Longworth, Cecilia Mascolo, and Andrius Aucinas. "EmotionSense: A Mobile Phones based Adaptive Platform for Experimental Social Psychology Research". In: *Ubicomp.* 2010.

[71] Divya Ramachandran, John Canny, Prabhu Dutta Das, and Edward Cutrell. "Mobilizing health workers in rural India". In: *the 28th international conference on Human factors in computing systems.* 2010.

[72] James A. Russell. "A circumplex model of affect." In: *Journal of Personality and Social Psychology* 39 (1980), pp. 1161–78.

[73] Peter Salovey and John D. Mayer. "Emotional intelligence". In: *Imagination, Cognition, and Personality* 9 (1990), pp. 185–211.

[74] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. "Item-based Collaborative Filtering Recommendation Algorithms". In: *WWW10, the Tenth International World Wide Web Conference.* 2001.

[75] Klaus Scherer. "The Neuropsychology of Emotion". In: ed. by Joan C. Borod. New York: Oxford University Press, 2000. Chap. Psychological models of emotion, 137?162.

[76] Didier Schrijvers, Wouter Hulstijn, and Bernard G.C. Sabbe. "Psychomotor symptoms in depression: a diagnostic, pathophysiological and therapeutic tool." In: *J Affect Disord* 109 (2008), pp. 1–20.

[77] Stephen M. Schueller. "Preferences for positive psychology exercises". In: *Journal of Positive Psychology* 5(3) (2010), pp. 192–203.

[78] Bjoern Schuller, Stefan Steidl, and Anton Batliner. *Emotion Challenge, Interspeech 2009.* http://emotion-research.net/sigs/speech-sig/emotion-challenge, retrieved in May 2011. 2009.

[79] Bjoern Schuller, Stefan Steidl, Anton Batliner, and Filip Jurcicek. *The INTERSPEECH 2009 Emotion Challenge: Results and Lessons Learnt.* SLTC Newsletter. 2009.

[80] Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, Gelareh Mohammadi, and Benjamin Weiss and. *INTERSPEECH 2012 Speaker Trait Challenge*. `http://emotion-research.net/sigs/speech-sig/IS2012-Speaker-Trait-Challenge.pdf`, accessed in March 2012. 2012.

[81] Björn Schuller, Anton Batliner, Dino Seppi, Stefan Steidl, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Noam Amir, Loic Kessous, and Vered Aharonson. "The Relevance of Feature Type for the Automatic Classi?cation of Emotional User States: Low Level Descriptors and Functionals". In: *Interspeech*. 2007.

[82] Matthew A. Siegler and Richard M. Stern. "On the Effects of Speech Rate in Large Vocabulary Continuous Speech Recognition". In: *ICASSP*. 1995.

[83] John A. A. Sillince. "A model of social, emotional and symbolic aspects of computer-mediated communication within organizations". In: *Computer Supported Cooperative Work (CSCW)* 4 (1996), pp. 1–31.

[84] Leslie Sim and Janice Zeman. "Emotion Awareness and Identification Skills in Adolescent Girls with Bulimia Nervosa". In: *Journal of Clinical Child and Adolescent Psychology* 3 (2004), pp. 760–771.

[85] Christina Sobin and Harold A. Sackeim. "Psychomotor symptoms of depression". In: *Am J Psychiatry* 154 (1997), pp. 4–17.

[86] Rui Sun, Elliot Moore, and Juan F. Torres. "Investigating glottal parameters for differentiating emotional categories with similar prosodics". In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2009.

[87] E. Szabadi, C. M. Bradshaw, and J. A. Besson. "Elongation of pause-time in speech: a simple, objective measure of motor retardation in depression." In: *Br J Psychiatry* 129 (1976), pp. 592–59.

[88] Paula T. Trzepacz and Robert W. Baker. *The Psychiatric Mental Status Examination*. Oxford University Press, 1993.

[89] Gerald Ulrich and K. Harms. "A video analysis of the non-verbal behaviour of depressed patients before and after treatment." In: *J Affect Disord.* 9 (1985), pp. 63–67.

[90] Carnegie Mellon University. *CMU Sphinx-3*. `http://cmusphinx.sourceforge.net/`. accessed in March 2012.

[91] Mehmet S. Unluturk, Kaya Oguz, and Coskun Atay. "Emotion recognition using neural networks". In: *the 10th WSEAS international conference on Neural networks*. 2009.

[92] Jan P. Verhasselt and Jean-Pierre Martens. "A Fast And Reliable Rate Of Speech Detector". In: *ICSLP*. 1996.

[93] Joseph C. Wu, J. Christian Gillin, Monte S. Buchsbaum, Tamara Hershey, J. Chad Johnson, and Jr. William E. Bunney. "Effect of sleep deprivation on brain metabolism of depressed patients." In: *Am J Psychiatry* 149 (1992), pp. 538–543.

[94] B. Yegnanarayana and R. L.H.M. Smits. "A robust method for determining instants of major excitations in voiced speech". In: *ICASSP*. 1995.

[95] S. J. Young. *The HTK Hidden Markov Model Toolkit: Design and Philosophy*. Tech. rep. University of Cambridge: Department of Engineering, Cambridge, UK, 1994.

[96] Marc A. Zissman. "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech". In: *IEEE Transactions on Speech and Audio Processing* 4.1 (1996), pp. 31–44.

[97] V. W. Zue, Timothy J. Hazen, and Timothy J. Hazen. "Automatic Language Identification Using a Segment-Based Approach". In: *Eurospeech*. 1993, pp. 1303–1306.

[98] Eberhard Zwicker. "Subdivision of the audible frequency range into critical bands". In: *The Journal of the Acoustical Society of America,* 33 (1961), p. 248.

Appendix A

Application Mockups

Emotional intelligence is defined as the ability to *recognize* moment-to-moment emotional experience *and* the ability to *manage* the emotions appropriately [73]. Emotionally intelligent individuals can recognize and respond to their own emotions, in order to manage stress and challenges. They can also better express these emotions to others, recognize others' emotional reactions, display trust, and produce empathic responses. However, people with mental disorders such as depression and PTSD (post-traumatic stress disorder) often display low emotional awareness [6][84]. They often lose their ability to recognize and manage the onset of harmful emotions, and may therefore fail to participate in preventing the onset of major mood episode. The frequent uncomfortable emotional experiences make them feel as though their emotions are unpredictable, out-of-control, and hard to identify.

With strategic feedback, it is anticipated that this will help users build their ability to retrospect about the emotions and developing coping skills. The application uses a mobile phone to constantly record the conversation from the user. After acquiring the voice data, the application will be able to identify the onset of emotions, and prompt appropriate feedback to users on the phone display. To be more specific, here we provide some scenarios that illustrate the lack of emotional intelligence in both a therapeutic and social situation. In each scenario, we also provide some design mockups to show the philosophy of the feedback mechanism.

- (Emotional flatness) Ethan is a kind, reliable, and amiable college student, but his emotional flatness inspired his friends to nickname him "Johnny 5", after a Hollywood fictional robot. However, he keeps it the same and never sees it as a problem. Recently he is about to graduate. He works so hard to find a job, but keeps failing. He is getting low but he doesn't even know.

  Proposal: With the feedback shown in Figure A.1, Ethan may learn that he is experiencing constant negative emotions. Feedback would allow Ethan to seek help early on and regain his confidence.

- (Social problem) A couple constantly argues, their relationship continues this way

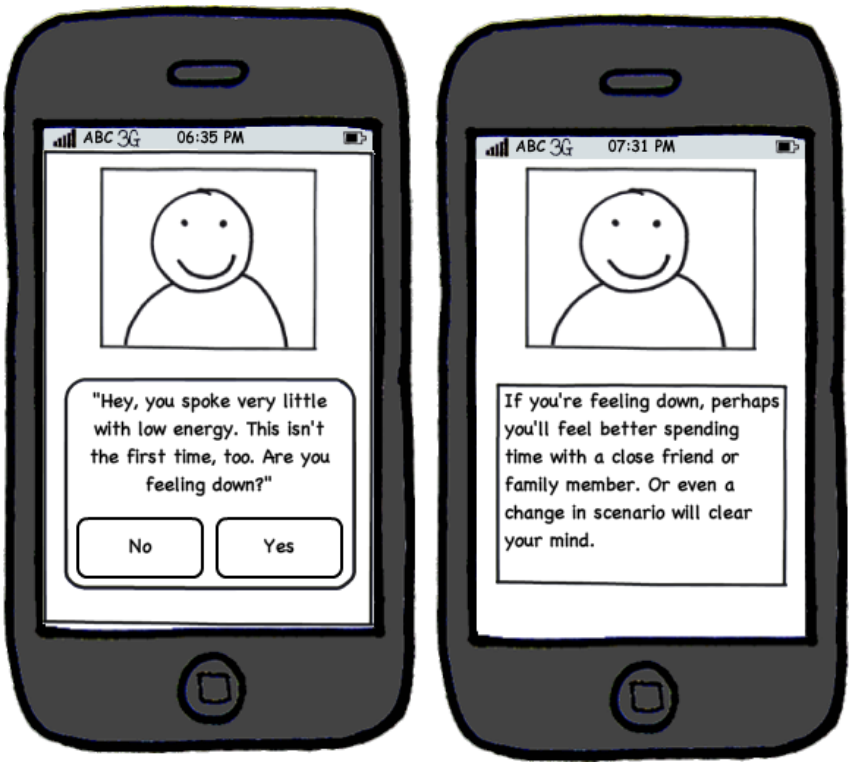


Figure A.1: Mockup for emotional flatness

for an extended period of time. However, the arguing grows worse. With emotional awareness, the couple may monitor both the frequency and intensity of their arguments, and may therefore attempt to adjust themselves in order to improve controlling their emotions.

Proposal: If the couples' arguments are growing more and more severe, an emotional awareness alert could offset future arguments. An alert illustrated in Figure A.2 can encourage each user to take his/her mind off each other, reflect, and even take a walk to clear the mind. These events may prevent each user from saying regretful words caused by spur-of-the-moment emotions.

- (Therapy) Alice's therapist wants to know how often she has negative emotions. However, Alice's accounts are often exaggerated. The therapist hopes for a way to accurately track her emotions on a regular basis so that she can properly diagnose Alice's dispositions.

  Proposal: Since Alice's therapist is unsatisfied with "about a few times a week" when asked how often she has negative emotions, the therapist could connect to the phone and download feedback of instances of negative emotions in Figure A.3 for a more

Figure A.2: Mockup for social problem

accurate picture . This would allow the therapist to diagnose and treat Alice better and accurately.

As shown, the application will act as a short break for people's everyday lives. Whether the emotions are anger, sadness, anxiety, etc., the system will document and alert the user of certain instances. This will clear the user's mind, make them reflect and question themselves, or even spur motivation and help. Overall, relationships, mental health, and physical health can be improved to make life more enjoyable.

Research of Feedback Mechanism

The feedback mechanism is the crucial components in building emotional awareness in the user. We are investigating two variables in the research of feedback mechanism: how well users will respond to our visual feedback design and the tone differences in the text-component.

For the visual feedback mechanism, there are two choices under consideration. The first allows the users to select an influential celebrity, historical figure, or someone they respect as the avatar or face of their mobile device. This intends to make the feedback more friendly

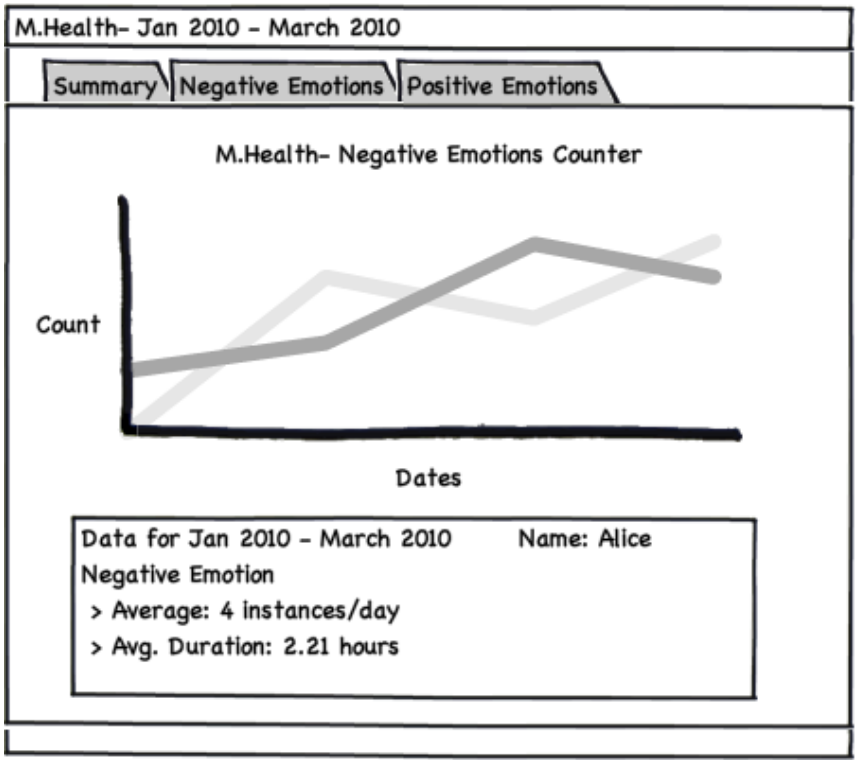


Figure A.3: Mockup for therapy

as well as influential. Whenever an instance of anger, sadness, or anxiety occurs, the mobile device now— will alert the user similar to that of a regular text-message, display the digital avatar/face complemented with a text message, which could include instructional suggestions, indirect mimicking, or a short story that expands the previous choice, as illustrated in Figure A.1, A.2, and A.3. However, the other possibility is to have no avatar or face at all, but retain everything else; this is geared towards those who prefer logical feedback or feel too mature for an avatar on their mobile device.device.

For the tone differences in the feedback design, we will explore the effectiveness of text that instructs the user what to do ("You should take a walk for some fresh air"), another that suggests what the user could try ("I would take a walk for some fresh air"), or a short, rich story. The short, rich story choice would work well with an avatar; a simple example is when the user experiences frustration or anger and he is alerted to Michael Jordan's story of being rejected from his high school basketball team, and how he persisted to win the NBA championship years later.