DETECTION OF CLINICAL DEPRESSION IN ADOLESCENTS' USING ACOUSTIC SPEECH ANALYSIS

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

By

Low Lu-Shih, Alex

B.Eng. Electronics

School of Electrical and Computer Engineering

Science, Engineering and Technology Portfolio

RMIT University

May 2011

ABSTRACT

Clinical depression is a major risk factor in suicides and is associated with high mortality rates, therefore making it one of the leading causes of death worldwide every year. Symptoms of depression often first appear during adolescence at a time when the voice is changing, in both males and females, suggesting that specific studies of these phenomena in adolescent populations are warranted. The properties of acoustic speech have previously been investigated as possible cues for depression in adults. However, these studies were restricted to small populations of patients and the speech recordings were made during patient's clinical interviews or fixed-text reading sessions.

A collaborative effort with the Oregon research institute (ORI), USA allowed the development of a new speech corpus consisting of a large sample size of 139 adolescents (46 males and 93 females) that were divided into two groups (68 clinically depressed and 71 controls). The speech recordings were made during naturalistic interactions between adolescents and parents.

Instead of covering a plethora of acoustic features in the investigation, this study takes the knowledge based from speech science and groups the acoustic features into five categories that relate to the physiological and perceptual areas of the speech production mechanism. These five acoustic feature categories consisted of the prosodic, cepstral, spectral, glottal and Teager energy operator (TEO) based features. The effectiveness in applying these acoustic feature categories in detecting adolescent's depression was measured. The salient feature categories were determined by testing the feature categories and their combinations within a binary classification framework.

In consistency with previous studies, it was observed that:

- there are strong gender related differences in classification accuracy;
- the glottal features provide an important enhancement of the classification accuracy when combined with other types of features;

An important new contribution provided by this thesis was to observe that the TEO based features significantly outperformed prosodic, cepstral, spectral, glottal features and their combinations.

An investigation into the possible reasons of such strong performance of the TEO features pointed into the importance of nonlinear mechanisms associated with the glottal flow formation as possible cues for depression.

LIST OF PUBLICATIONS

Journal publications:

- LOW, L.-S. A., MADDAGE, N. C., LECH, M., SHEEBER, L., and ALLEN, N., "Detection of clinical depression in adolescents' speech during family interactions," *IEEE Transactions on Biomedical Engineering,* vol. 58, pp. 574-586, 2011.

Conference publications:

- LOW, L.-S.A., MADDAGE, N. C., LECH, M., and ALLEN, N., "Mel frequency cepstral feature and Gaussian mixtures for modeling clinical depression in Adolescents," *Proceedings of the 8th IEEE International Conference on Cognitive Informatics,* pp. 346-350, 2009.

- LOW, L.-S. A., MADDAGE, N. C., LECH, M., SHEEBER, L., and ALLEN, N., "Content based clinical depression detection in adolescents*," in 17th European. Conf in Signal, Speech, and Image Processing (EUSIPCO 09'),* pp. 2362-2365, 2009.

- LOW, L.-S. A., MADDAGE, N. C., LECH, M., SHEEBER, L., and ALLEN, N., "Finding an Optimal length of utterance for the detection of depression symptoms in speech of Adolescents," presented at the *Proceedings of the 2009 International Symposium on Bioelectronics and Bioinformatics*, Melbourne, Vic., 2009.

- LOW, L.-S. A., MADDAGE, N. C., LECH, M., SHEEBER, L., and ALLEN, N., "Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents," *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 5154-5157, 2010.

- MADDAGE, N. C.; SENARATNE, R.; LOW, L.-S. A.; LECH, M.; ALLEN, N., "Video-based detection of the clinical depression in adolescents," *EMBC: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vol. 1-20,* pp. 3723-3726, 2009.

ACKNOWLEDGEMENTS

Another stage of my life is reaching to a close, and it is one that I would never forget. Not only has the light at and the end of the tunnel finally shone so brightly, rainbows of accomplishments have started to seep in. The coming to the end of this long and tough journey would not have been possible without people who have helped me accomplished this work directly or indirectly along the way. Therefore, I would like to take this opportunity to express my utmost gratitude to them.

First of all I would like to express my sincere thanks to my main supervisor Dr. Margaret Lech, who gave me the opportunity to work on this project and for having the faith in giving me the freedom to explore my own ideas. Her enthusiasm, support and guidance in the project are very much appreciated.

I also like to thank my second supervisor Dr. Namunu Maddage whose valuable advice started me in the right direction in the research and prevented me from going off track, as well as his time spent in teaching me the HTK toolbox setup.

This research would not have been possible without the support of Dr. Lisa Sheeber who provided us with the database to work on. Aside from providing the necessary statistical analyses on the demographic data, her constructive comments and suggestions improved the quality of our publications and therefore my sincere thanks also goes out to her.

Special thanks have to go out to Professor Nicholas Allen for his many invaluable discussions. His constant availability in providing his expertise in psychology, even with his busy schedule, has always been much appreciated. It always surprises me that in any of the meetings that we have, his advice and ideas always seemed to lead me on with a

gazillion more ideas. His contributions have been of great importance to this dissertation and will not be forgotten.

Finally I would like to express my gratitude to the dearest persons in my life. To Sarah-Anne Chan, who I have known for almost half of my life, thank you for spending your precious time in proof reading the grammar and vocabulary in my publications. I know your reading interests are in poetry and literature, and this probably bored you a lot. So a big super duper thanks to you. ☺

I am grateful for all the love and support my family has given to me. Thanks to my uncle Beng, and my brother Richard for always giving words of encouragement and support. To uncle Hee, my main motivator and is someone whom I really look up to. Thank you for all your advice, encouragement and support. You are the person who always has believed in me and changed my life around for the better. I will always be indebted to you. Your quote has always motivated me throughout my studies: "*The beginning of every problem solving task is like riding a bicycle up a hill, one has to push really hard to reach the top of it, but once all the hard work is done, your feet can be lifted off the pedal and the bicycle will glide down*".

Last but not least, to my mom for being such a selfless and supportive person. Her encouragement always spurred me on to achieve better. I am always grateful for her struggle in bringing the kids up and providing the financial support which allowed me to pursue my degree. Home cooked food was always provided at the dinner table. Everything she did was always for the kids and I am eternally grateful for that. From a son's promise to his mom: it is now my turn to take care and provide for you mom!

A new chapter in my life is now about to begin. Taking my knowledge gained from this journey, along with my uncle Hee's quote, I am well prepared for any life challenges ahead.

DEDICATION

To my parents and uncles, for always believing in me and turning my life around 360° for the better.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES AND ILLUSTRATIONS

Chapter One:

INTRODUCTION

"He is apparently unable to move and express himself freely. This very circumstance, that the answers come slowly, even on matters of indifference, shows that in this patient we have not to deal with a fear of expressing himself but with some general obstacle to the utterance of speech. Indeed, not only speech but all action of the will is extremely difficult to him . . . . The disturbance must be essentially confined to the accomplishment of voluntary movements. This constraint is by far the most obvious clinical feature of the disease and compared with this, the sad, oppressed mood has but little prominence."

– Emil Kraepelin

1.1 Preview

Speech is a natural form of communication for human beings and has been recognized as one of the potential sources in providing cues for depression. Clinical depression[1] has been related to severe affect disturbances of a person's feelings and moods, and can be conceptualized in the failure to regulate emotions [108]. For example, in one of the letters that was sent to us from a parent of a depressed adolescent, it was commented that *"I have always believed that I could pretty accurately identify the episodes and severity of depression just from the sound of her voice.... As a mother, I think I am familiar with*

---

[1] The term clinical depression and depression is used interchangeably in this study.

every nuance of the sounds my children utter. I can hear the depression"[2]. This reveals supporting evidence that when someone is depressed or could be depressed but pretends to be well; listeners attend to the tone of the voice rather than the linguistic content to make social judgments of the person's behavioral traits. By paying particular attention to ones tone and word emphasis, the voice of a person during social interactions may expose important information that can express a person's emotional state, moods, attitude and personality. It makes a large difference where the emphasis or stress is placed on certain words. For example, consider the same sentence spoken in various ways by altering the pitch and tonal patterns on how the words are stressed in the following sentence "I would appreciate if you pick up around the house more." By paying close attention to the emphasis placed on the word "appreciate", the tone of the voice could indicate a person being hostile, sarcastic, bitter, pleased, angry, painful, etc. (i.e., a loud voice tone on the word "appreciate" may indicate a person being angry or sarcastic). According to psychologists, the ways these emotions are expressed through a person's voice are one of the communication channels (other channels are face, body and gestures) that clinicians look out for in the search of possible tell-tale signs in depressive symptoms.

The evolution of speech information processing and speech recognition technologies in computers has made it possible to develop objective measures that can measure these speech cues. This is based upon the assumption that the emotional state of a person suffering from a depressive disorder affects the acoustic qualities of their speech, and therefore depression could be detected through an analysis of perceived

---

[2] The authors obtained a written permission from the parent to publish the citation.

changes in the acoustical properties. Therefore, this research delves into the realm of clinical depression in adolescents to investigate the acoustic properties of speech as potential indicators of depression.

This introductory chapter has several objectives which are to define the problem that arises with adolescents' clinical depression and to discuss how the idea and perspective of this research is new. The last section of this chapter gives an overview of the components for the rest of this dissertation.

1.2 Problem Description

Since the 1970s, the increased prevalence of clinical depression in adolescents has been linked to a range of serious outcomes, particularly an increase in the number of suicide attempts and deaths [62], [87]. Depression, which occurs in 4-8% of adolescents at any given point in time [121], is one of the most prevalent health problems affecting adolescents. Adolescents are usually defined as aged 13-20 years old. Indeed, many adolescents experience depression that is severe enough to warrant treatment, and around 20% of young people (adolescents) will have experienced clinically significant depressive symptoms by the time they reach adulthood [87], [115]. Depression in young people often goes unrecognized and untreated, making the treatment more difficult at the later stages of life [115]. It has been recognized that depression is a chronic or recurring disorder, and when untreated, the effects can be devastating. Stephen Fry, in his BBC documentary "*The Secret Life of the Manic Depressive*" provides a fascinating account of his troubled adolescent life, and it was only later during adulthood, when he was diagnosed as a sufferer of depression. The World Health Organization (WHO) [121]

shows that mortality rates due to depression increase exponentially with age, accounting for a total of 850,000 lives out of about the 121 million people affected by depression worldwide every year. It is also predicted that by the year 2020, depression will be responsible for the second greatest burden of disease internationally (following cardiac disease), making it a prominent public health concern. The key to suicide prevention is early intervention for depression. Therefore, the early detection of depression, especially depressive episodes occurring during adolescence, is of primary importance. It can minimize disturbance of typical functioning and development of social and academic skills.

Unfortunately, depressed adolescents often are not aware of their state and therefore are not able to express verbally their feelings. The diagnosis of depression in adolescents relies on observations of behavioral patterns, and interviews with parents and teachers. This process is time consuming and the illness is usually recognized in the advanced stages. The current diagnosis is qualitative and largely based on the personal skills, experience and intuitive judgment of a mental health practitioner. The number of highly skilled professionals is limited [107], and their availability is restricted to major towns and health centers. In the United States, the current work force of skilled clinicians at present is below the national requirement of 30,000 [60]. The development of objective and quantitative measures of depressive symptoms in speech will help clinicians by providing an objective adjunct to current diagnostic techniques. Specifically, an automatic, computer-based analysis of speech, indicating the probability of depression, will provide an important objective indicator that can be used as a mass-screening device, followed by more detailed (and more resource intensive) interview-based clinical

diagnosis of depression. It will give an immediate quantitative assessment of the potential mental state of a patient, and thus help to determine if a person showing certain emotional problems should seek professional help and further evaluation.

1.3 Thesis Aims

Most of the studies to date concentrate on deriving acoustic correlates of depression and only a few studies [37], [38], [82], [84], [90] [91], [97] use these correlates in the diagnosis of depression. No specific studies addressing acoustic parameters of speech as indicators of depression in adolescents have been published. This is especially important given the fact that adolescence is associated with a dramatic increase in the incidence of depressive symptoms and disorders, suggesting that this is a critical life phase for early detection and intervention with both depressed and at risk teenagers [70]. Both males and females are affected by voice changes during adolescence [48]. Female voices go down only by a couple of tones and therefore, in girls, the change is hardly noticeable. Boys however experience quite a dramatic change in tone. Their voices deepen and might drop by as much as a whole octave. During adolescence, the male cartilage supporting vocal cords grows larger and thicker. At the same time, the vocal cords grow 60% longer and become thicker. When they vibrate, they do so at a lower frequency than before. During adolescence, facial features also change quite considerably. As the facial bones grow, they create bigger spaces within the face. Larger cavities in the sinuses, nose and back of the throat give the voice more room to resonate. During this transitional period lasting for a number of years, the speech characteristics of adolescents differ significantly from the speech characteristics of adults. This indicates that the effects of depression on speech of

adolescents could also significantly differ from the effects of depression on speech of adults, and this remains a critical issue for future research.

Firstly, this study aimed for the first time, to provide an extensive investigation into the acoustic correlates in the detection of depression in adolescents.

Secondly, early studies of acoustic correlates in speech were limited to small databases (very few participants and short audio recordings) and manual or semi-manual processing of speech recordings (i.e., the number of depressed participants investigated in previous studies - Darby and Hollien (1977) - 13 subjects; Nilsonne *et al.* (1987) - 16 subjects; Kuny and Stassen (1993) - 30 subjects; Ellgring and Scherer (1996) - 16 subjects; France *et al.* (2000) – 42 subjects; Alpert *et al.* (2001) - 22 subjects; Ozdas *et al.*(2004) – 10 subjects; and Moore *et al.* (2008) – 15 subjects). To make issues more complicated, there have been discrepancies in the results presented from one study to another. For example, Nilsonne *et al.* (1987) found that the standard deviation rate of change in fundamental frequency (F0) and the mean absolute rate of change in F0 both correlated to clinical depression whereas France *et al.* (1987) reported that in females, F0 to be ineffective discriminators. Moreover, the majority of studies worked with data collected during clinical interviews from hospital patients currently undergoing depression episodes or people seeking help from practitioners. Therefore, the aim of this thesis was to provide further research validating the proposed measures with larger sample sizes and using naturalistic every day speech. It was assumed that algorithms that can accurately detect depression from the naturalistic conversational speech are more likely to be applicable to the screening for depression symptoms and early detection of depression.

Thirdly, apart from implementing traditional acoustic features such as prosodic, cepstral , spectral and glottal measurements as described in many previous studies [88], [65], [5], [33], [38], [91], [84], the acoustic feature derived from the Teager energy operator (TEO) called the TEO critical band based autocorrelation envelope (TEO-CB-Auto-Env) [123], that measures the number of additional harmonics due to the non-linear airflow in the vocal tract was tested as a possible indicator in discriminating speech of depressed and non-depressed adolescents. The reason for this was because there have been some psychological reports documenting that life stresses is an important component in the cause of depression (or subtype of depression) [43] and recent studies in speech analysis have shown that the non-linear feature of TEO to be successful in stress and emotion recognition [123]. However, the complex relationship between emotional stress and clinical depression still remains somewhat unclear [68].

Fourthly, we wanted to examine the role of some previously reported phenomena in our data, such as the well established gender differences in depressive symptoms during early adolescence [89], and the importance of glottal features in the detection of clinical depression [84].

Fifthly, we wanted to determine the optimal duration of the speech samples to be analyzed, due to the fact that in past research, there has been variations in the duration of the analysis of speech samples i.e. 20 seconds in [38], [84] and 30 seconds in [91].

1.4 Thesis Scope

This dissertation provides an initial attempt to investigate the acoustic correlates of depression in speech of adolescents. The work presented here uses a framework that

includes the processes of feature extraction, modeling, and classification. Speech recordings were used to extract a set of characteristic features and build statistical models representing depressed and non-depressed participants. The scores obtained through an automatic classification indicated the correlation level between the characteristic features and a given class.

This study chooses two types of classifier so that the classification results were not bias to either classifier. Besides implementing the Gaussian mixture models in the modeling and classification process, we compared it with a SVM approach that was modified to increase the computational efficiency in handling the large training dataset used in this study. The modified SVM approach replaces a single SVM with a parallel configuration of SVMs and a global optimization algorithm based on simulated annealing was used to determine the weight associated with each SVM.

A simple filter approach was implemented to select the top feature coefficients based on a ranking score on the TEO-based feature category that improved the overall accuracies in distinguishing speech of depressed and control adolescents.

1.5 Thesis Contributions

The thesis main contributions can be summarized as follows:

- Provided a comprehensive study of acoustic correlates of depression in naturalistic speech of adolescents recorded during family interactions. The communication of speech during these family interactions is important as this form of social communication represented naturalistic speech where expressive

behavior and emotional states can naturally occur. Prior to our study, no specific studies addressing acoustic parameters of speech as indicators of depression in adolescents had been published. In addition, the database used in this study was of a larger sample size which allowed further validation of acoustic features proposed by other studies.

- Investigated non-linear feature based on the Teager energy operator (TEO) which have been documented to be effective in stress classification but has never to the authors knowledge been performed in clinical depression detection in speech. This study found TEO-based feature to also appear to be powerful discriminants of depression in speech.

- Implemented feature categories proposed by other studies and provided consistent results that supported the importance of glottal features in the detection of clinical depression in speech.

- Examined gender differences in adolescent's depression and found that modeling genders separately was more effective in distinguishing between speech of depressed and non-depressed adolescents than with the modeling of both genders combined together. This shows the classification accuracies in detecting depression in speech of adolescents depended on the type of gender.

- Investigated the optimal overall length of the speech samples for the purpose of classification. It was determined that 1 min of concatenated voiced utterances maximized the subject-based accuracy in detecting depression in adolescents.

1.6 Thesis Structure

The remaining parts of the thesis are organized as follows:

Chapter 2 begins with an introduction to sound waves and human speech communication. A background review of speech production theory is then given, explaining the process of the speech production system and its anatomical structures in the production of speech sounds. This is followed by a discrete-time representation in modeling the speech production system. Finally, acoustic features derived from the discrete-time model that are closely tied to the anatomy of the physiological and perceptual phenomena in the speech production mechanism are discussed.

Chapter 3 provides a background review of two types of pattern recognition techniques used in this study relating to speech that deals in algorithms that construct useful information and patterns from derived acoustic features of the speech signal.

Chapter 4 gives a literature review of previous work relating to understanding the effects of emotional arousal on the physiology of the speech production mechanisms. This is followed by a discussion of past research efforts that make use of speech as indicators of emotions, psychological stresses and depression.

Chapter 5 describes the collection and formulation of the database, together with the preparation of the speech corpus used in this research investigation.

Chapter 6 talks about the framework in modeling speech contents of depressed and non-depressed participants along with the methods and algorithms implemented in the investigation.

Chapter 7 presents the evaluation methods, the experiments conducted and the results obtained.

Chapter 8 summarizes and discusses the research in this dissertation, followed by interpreting the major findings from experimental results obtained in *Chapter 7*. The chapter is then concluded with future work.

Chapter Two:

SPEECH PROCESSING FUNDAMENTALS

"The mediocre teacher tells. The good teacher explains. The superior teacher demonstrates. The great teacher inspires."

– William Arthur Ward

2.1 Preview

In order to develop man-machine interfaces to understand speech, it is necessary to learn about the mechanisms by which speech is produced and perceived. While it is beyond the scope of this dissertation to provide an in-depth detail on the theory of speech processing, this chapter is still designed to provide the essential details to equip the readers with the basic knowledge in applying digital signal processing techniques to the analysis of speech communication so that the experiments explained in the later chapters can be replicated.

2.2 Sound Waves

Before discussing about the human speech production process and how the nature of speech sounds are generated and perceived, it is very important to understand the basic principles of how sound waves work in general. This knowledge will form the foundation in the field of speech science.

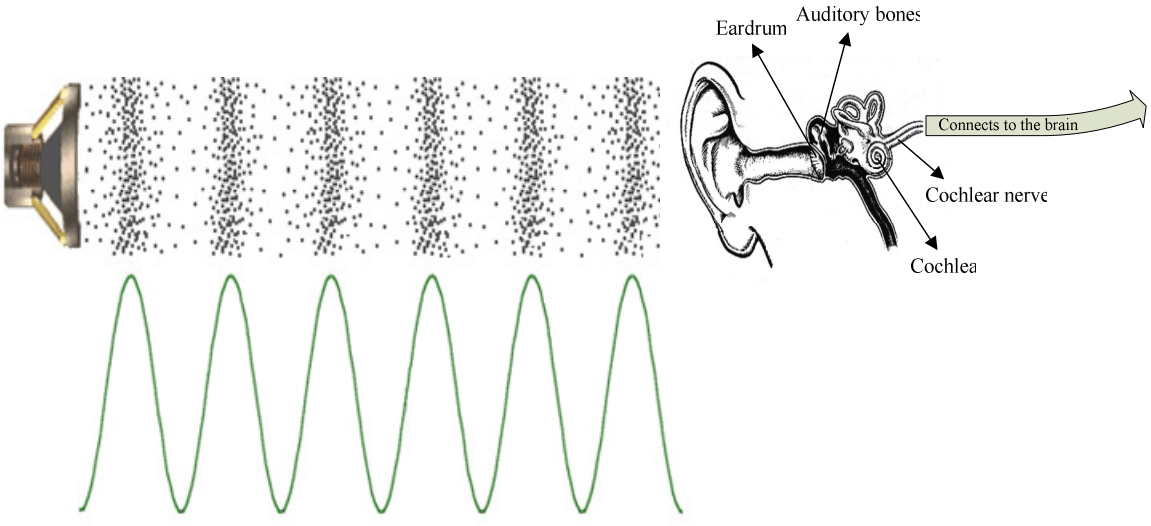

Figure 2.1: Representation of a sinusoidal waveform being converted into a sound wave.

Sound is a sequence of waves that propagates as variations of pressure in a medium such as air. They are created by the vibration of an object, which causes the surrounding air to vibrate. The vibrating air then causes the human eardrum to vibrate, which the brain interprets as sound. The diagram in Fig. 2.1 depicts an illustration of a sinusoidal waveform converted into a sound wave with the black dots corresponding to the molecules of air. A sound wave has the same characteristic as any type of waveform in an electronic signal in that it has wavelength, frequency, velocity and amplitude. The maximum amplitude of the sinusoidal waveform (high voltage) corresponds to the black dots being fully concentrated together. This concentration represents molecules of air being packed densely together. The minimum amplitude of the waveform corresponds to the black dots (air molecules) being spread more thinly. As the loudspeaker vibrates, it causes the surrounding molecules to vibrate in a particular pattern represented by the waveform. As the vibrating air hits the eardrum, it causes the membrane to vibrate in the

same pattern. As the eardrum moves, it sends a signal to the cochlea via three small auditory bones. Liquid in the cochlea stimulates the nerve endings which in turn send messages to the brain. The brain then interprets these messages from the nerves in what we call sound.

2.3 Speech Communication (the Speech Chain)

"Communication" here is defined by Wilson (1975) [120] as occurring "whenever the behavior of one individual (the sender) influences the behavior of another (the receiver)". Now let us consider the case of human speech communication in regards to the propagation of sound waves. Speech is used as a communication tool consisting of a chain of events in conveying information from a speaker's brain to a listener's brain. This chain of events is called the *speech chain* and is illustrated in Fig. 2.2. Like sound waves are created as described in the previous section, a speaker must produce a speech signal in the form of a sound pressure wave that travels from the speaker's mouth to the listener's ears. A speech waveform in itself is an acoustic sound pressure wave that stems from positions and movements of anatomical structures which make up the human speech production model. Speech signals are composed of a sequence of sounds that serve as a symbolic representation for a thought that the speaker wishes to relay to the listener. The arrangement of these sounds is governed by the rules of a language. The study of these rules by which speech sounds are assembled in a language is called *linguistics*. The study of the classification of speech sounds is called *phonetics*. A detailed discussion on linguistic and phonetics can be found in the selected references [30], [95].

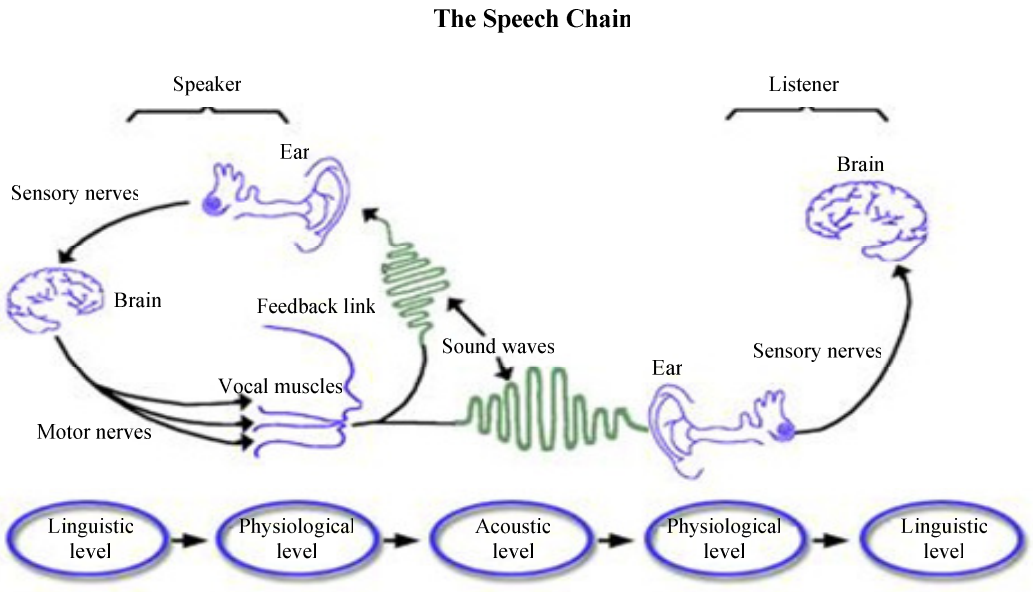

Figure 2.2: the speech chain illustrating different forms of a spoken message in its progress from the brain of a speaker to the brain of a listener (after Denes & Pinson, 1993).

The process of the speech chain originates from an idea or thought from the speaker's brain and consists of 5 states that connect the speaker to the listener in the following order:

1) *Linguistic level (speaker side)*: transmission of the speaker's ideas is converted by the selection and ordering of suitable words and sentences based on learned grammatical rules associated with a language.
2) *Physiological level (speaker side)*: The speaker then conveys this message on a physical level through a series of neural and muscular activity, with the generation and transmission of an acoustic sound pressure wave.
3) *Acoustic level*: The acoustic wave transmitted from the speaker is received by the listener's auditory system. The process from step 1 to 2 is then reversed.

4) *Physiological level (listener side):* On the listener's side, events start on a physiological level with the incoming sound waves activating the neural activity in the hearing and perceptual mechanisms. The sound waves are also feedback to the speaker's own ear, allowing the speaker to continuously monitor and make adjustments to the message intended to be transmitted.
5) *Linguistic level (listener side):* The speech chain is completed when the listener recognizes the words and sentences transmitted by the speaker.

2.4 Process of the Speech Production System

2.4.1 Anatomy

In this section, the study of the production of speech sounds from an anatomical point of view of the speech production system will be briefly discussed. As mentioned in *Section 2.3*, the speech waveform is an acoustic sound pressure wave that originates from voluntary movements of anatomical structures in the human speech production system.

Fig. 2.3 shows the classical schematic diagram of the human speech production mechanism. The main anatomical components of the system are composed structurally of the lungs, trachea (windpipe), larynx (voice box), pharynx (part of the throat), oral cavity (mouth) and nasal cavity (nose). The finer anatomical structures are the vocal folds, velum, tongue, teeth and lips which act as articulators in producing various speech sounds depending on the positions they are at.

There are four main processes (shown highlighted by circles in Fig. 2.3) of the speech production mechanism that produce the physical production of speech sounds.

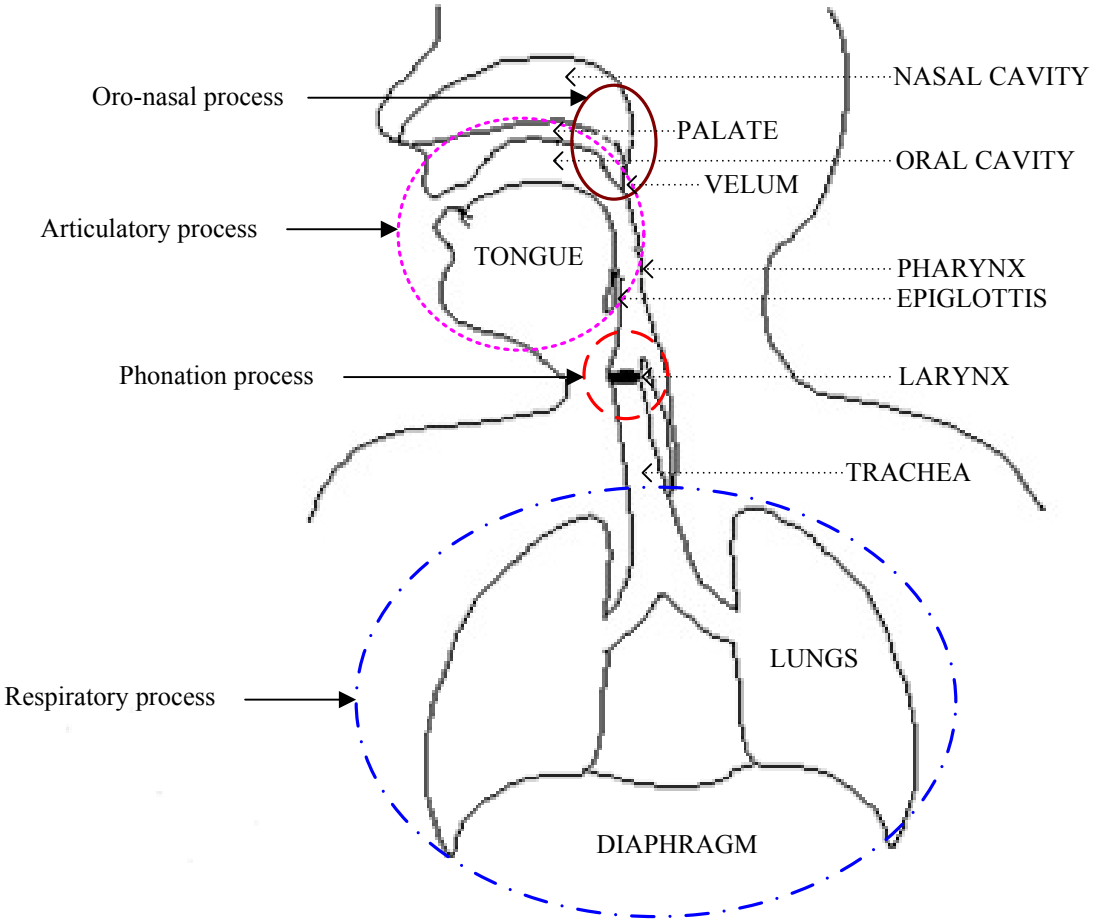

Figure 2.3: Schematic diagram to show subdivisions of the human speech production mechanism[3]

The processes include: respiratory (initiation for setting airstream in motion), phonation, articulation and oro-nasal.

The respiratory process starts from the lungs and is the moment when air from the lungs is expelled up the trachea to the vocal folds.

The phonation process occurs at the larynx where the vocal folds are located. The vocal folds are made up of two fibrous sheet of tissue and are responsible for the voice

production. Once the force of the air pressure is great enough, it will create an opening in the vocal folds known as the glottis. In producing a voiced speech segment, such as a vowel (which will be discussed in the next section), the continuous puff of air causes the slit of the glottis to open, setting the vocal folds to oscillate into vibration mode.

For the oro-nasal process, the velum controls the air intake into the nasal or the oral cavity, after it has gone through the larynx and the pharynx.

Finally, the articulation process takes place in the mouth where the finer anatomical structures i.e., upper and lower lips, upper and lower teeth, tongue (tip, blade, front, back) and roof of the mouth (alveolar ridge, palate and velum) move to different positions to produce the various speech sounds.

2.4.2 Excitation of the Speech System

Speech sounds can be classified into 3 distinct classes according to their mode of excitation. They can be classified as voiced, unvoiced or mixed.

Voiced sounds provide a periodic excitation to the system and are produced by forcing air through the glottis with the tension of the vocal cords adjusted so that they vibrate in a relaxation oscillation, thereby producing quasi-periodic pulses of air which excite the vocal tract.

Unvoiced sounds are generated by forming a constriction at some point along the vocal tract, and forcing air through the constriction to produce turbulence.

---

[3] Schematic diagram taken from http://www.haskins.yale.edu

A sound may be simultaneously voiced and unvoiced (mixed). Furthermore, some speech sounds are composed of a short region of silence, followed by a region of voiced speech, unvoiced speech or both. For the purpose of experiments conducted in this dissertation, only voiced sounds were considered.

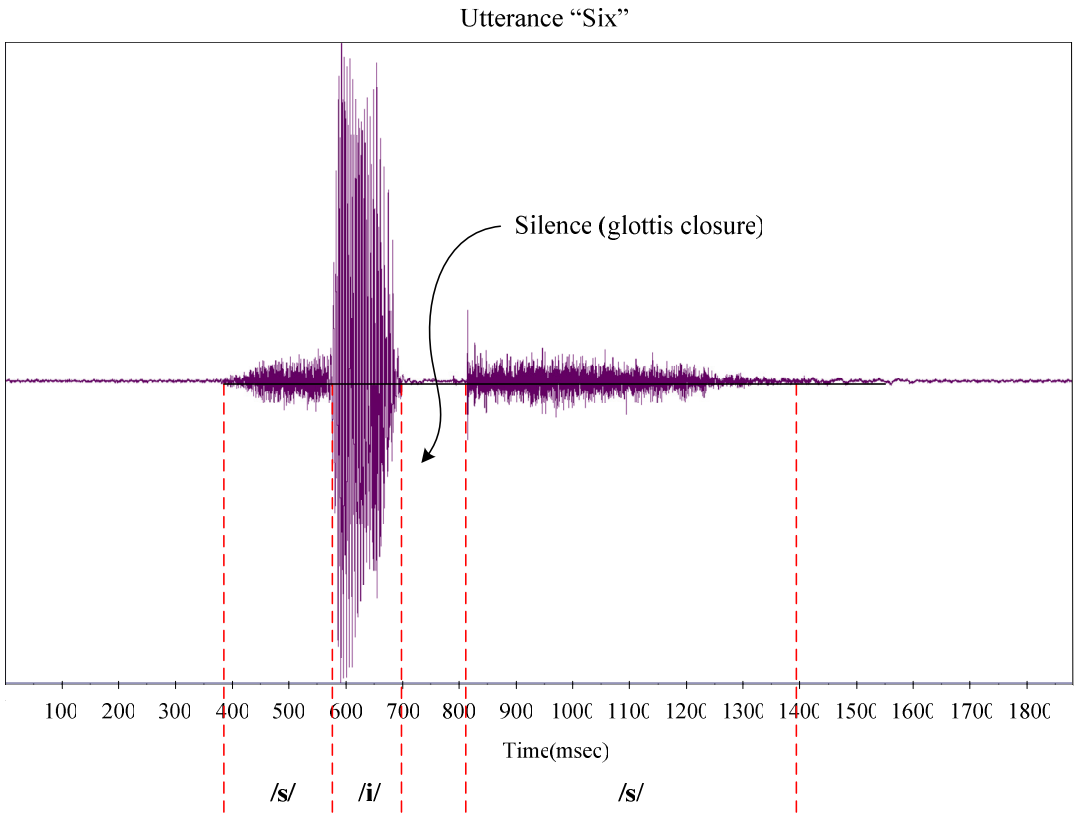

Figure 2.4: A speech signal of an utterance "Six" spoken by a male speaker taken from the SUSAS [45] database.

The difference in the term voiced and unvoiced is best illustrated by an example of the word "six" as shown in Fig. 2.4. The vowel /I/ in the word "six" is classified as voiced. The /s/ sound pronounced in the word "six" has a hissing sound which is generated by forming a constriction and is classified as unvoiced. The silence region

located in Fig 2.4 is where the glottis is at complete closure and no airflow is passing through the vocal folds.

2.4.3 Physiology of Voice Production

Let us now focus on the larynx whereby the physiological characteristic of this special organ forms the voice production. As previously mentioned in *Section 2.4.1*, phonation or voicing occurs in the larynx where the vocal folds produce certain sounds through quasi-periodic vibration. The source of voicing occurs by means of air pushing out from the lungs into the trachea, then up to the glottis (the gap between the vocal folds) providing a periodic excitation from the vibration of the vocal folds. The rhythmic opening and closing of the glottis (vibration of the vocal folds) is due to the sub-glottal air pressure in the trachea. The sequence of events illustrating the phonation process in the larynx is shown in Fig. 2.5, accompanied by its steps described in the following:

(a): The sub-glottal air pressure begins below the true vocal folds.

(b) & (c): As the air pressure builds up, it forces the vocal folds to split open.

(d): Once the vocal folds begin to open, it creates a slit like passage known as the glottis; air then begins to rush out from the trachea through the glottis.

(e): The sub-glottal air pressure continues to force the glottis to open wider and outward creating a rapid rise in airflow.

(f) - (j): The glottis then starts to narrow causing the air pressure to fall, therefore resulting with an increase in airflow velocity. The reason for this is because the same amount of airflow is travelling through a smaller constricted passage and the only

way for the velocity to increase is for the pressure from behind the airflow to be greater than the pressure in front. This in turns, create a suction effect called the Bernoulli force which pulls the vocal folds back together at the lower edge first before fully closing up.

This cycle in steps (a)-(j) then repeats again providing a vibration of the vocal folds. The rate of this vibration is called the fundamental frequency.

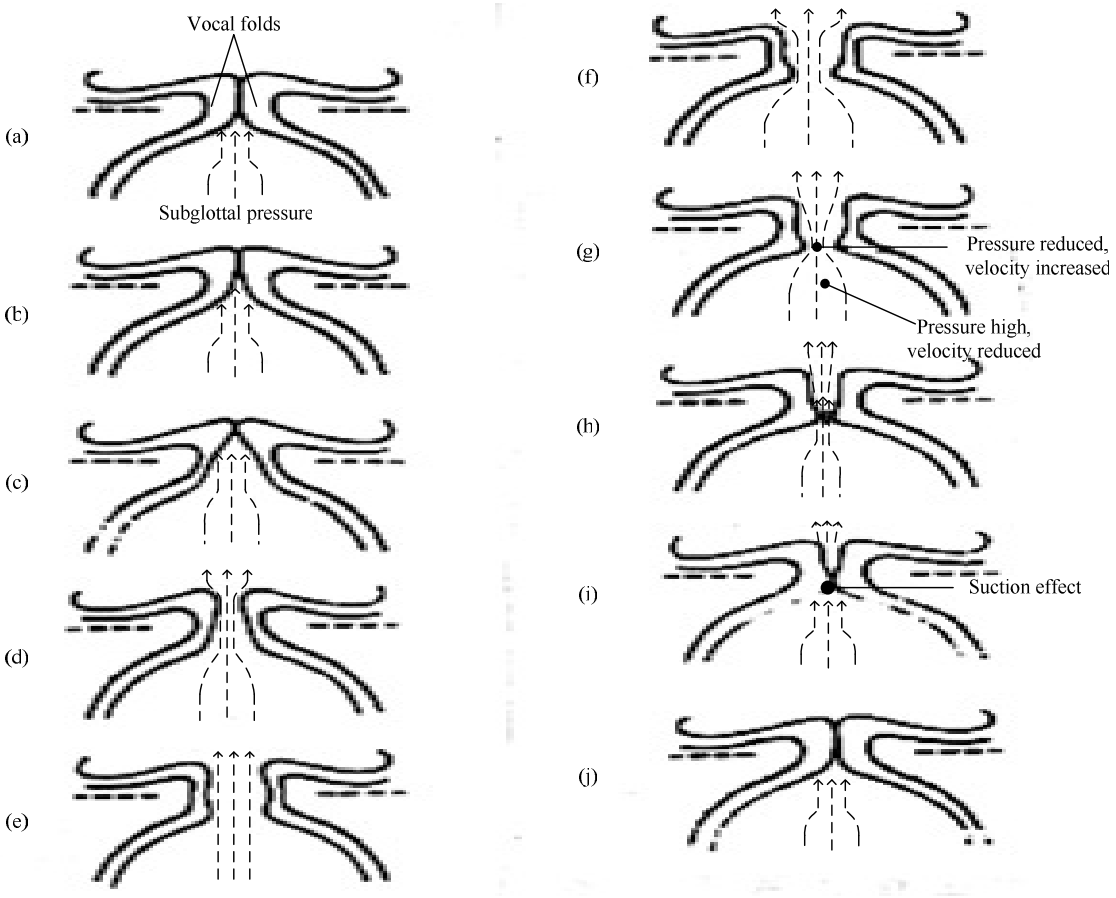

Figure 2.5: Illustration of the sequence of events in the phonation processing that occur in the larynx (cross-sectional view) after Coleman.[4]

[4] http://www.phon.ox.ac.uk/jcoleman/

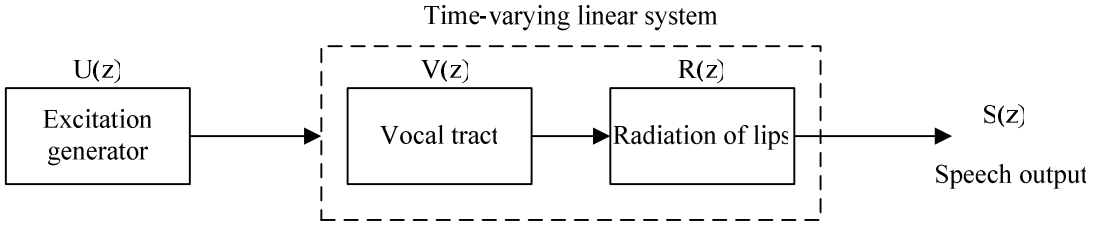

Figure 2.6: General source-system linear model of speech production.

2.5 Modeling the Speech Production System

From a digital signal processing perspective, in which this topic is all about, the speech production system can be imagined as an acoustic filtering operation whereby the anatomy of the speech production is made up of a source (glottis in the larynx) and a filter (pharyngeal cavity (throat), oral cavity (mouth), etc.). The pharyngeal and oral cavities are usually grouped into one unit which is referred to as the vocal tract.

As shown by the general block diagram of the speech production model in Fig. 2.6, it can be modeled into three separate components consisting of the:

I. Excitation Source, U(z)

II. Vocal tract shaping, V(z)

III. Radiation effects, R(z)

The production of speech sounds (voiced) originates from the excitation generator, *u(n)* which creates a periodic train of glottal pulses. The movement of the vocal tract articulators imposes its resonances upon this source excitation so as to produce the different sounds of speech. This process can be represented by the general block diagram of the source-system model of the speech production in Fig. 2.6. From the dotted lines drawn in Fig. 2.6, it can be seen that the vocal tract and radiation effects can

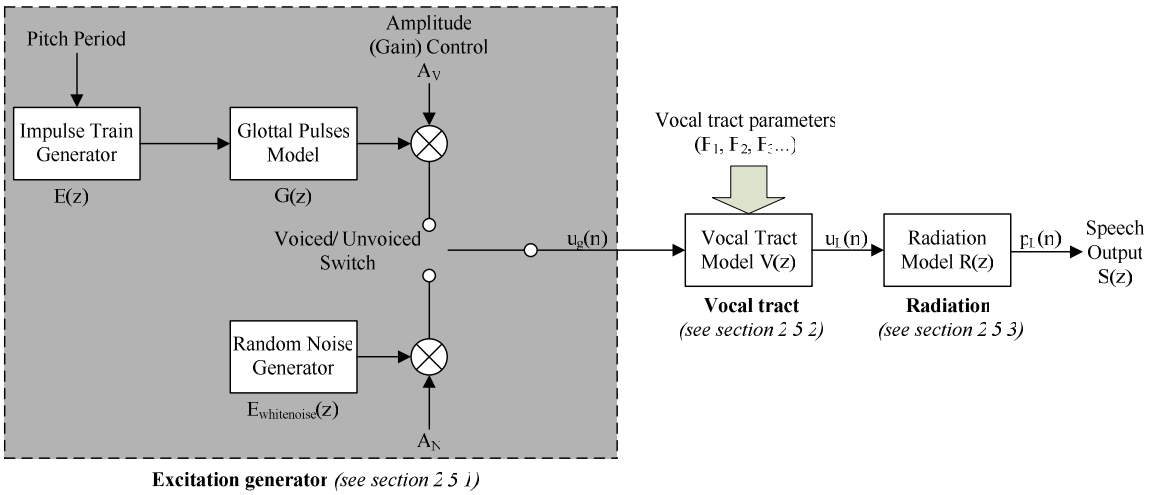

Figure 2.7: General discrete-time linear model for speech production.

be modeled by a time varying linear system. The time-varying nature of speech sounds requires the mode of excitation from the source and the resonance properties of the linear system to change slowly with time. For many speech sounds, it is reasonable to assume that the general properties of excitation and vocal tract articulators remain fixed for periods of 10-30msec. The term linear model is used because each component is combined linearly and is separable. This provides the ability to filter out each component to access its properties individually as given in the z-domain transfer function of the discrete-time linear model of speech:

$$S(z) = U(z)V(z)R(z) \tag{2.1}$$

$$S(z) = E(z)G(z)V(z)R(z) \tag{2.2}$$

A general linear discrete-time model for speech production is shown in Fig. 2.7. The model in Fig. 2.7 is termed a terminal analog model as it is only representative to the true physical system of the speech production process based on its output (terminal)

signals, its internal structure does not mimic the physics of speech production. We will now explain the breakdown of each component and its terminals (input and output) that model the speech production system in the following subsections.

2.5.1 Excitation Generator

The dotted lines in Fig. 2.7 depict the breakdown of the components associated with the excitation generator. The majority of speech sounds can be classified as either voiced or unvoiced. In the production of voiced sounds, the excitation generator, *e(n)* creates a periodic train of equally spaced pulses. This train of pulses then excites the glottal shape filter, *g(n)* which results in the required glottal pulses for voiced sounds. The voiced excitation, *u(n)* can be written as:

$$u(n) = \sum_{i=-\infty}^{\infty} g(n - iP) \tag{2.3}$$

where *g(n)* is the impulse response of the glottal shaping filter and *P* is the period of the train pulses *e(n)* assuming that voicing last forever.

In the voiceless case, the source excitation is characterized as a point of major constriction along the vocal tact (i.e., as in the /s/ in “six” shown in Fig. 2.4) or an explosive form during a stop release. This unvoiced excitation can be modeled with white noise, $E_{whitenoise}(z)$ as shown in Fig. 2.7.

2.5.2 Vocal Tract Model

The acoustic theory of the speech production source-system model is based upon the assumption that the vocal tract can be represented as a concatenation of lossless acoustic

tubes [30], [95]. The reason why these losses tube models are widely used is because it has many properties common with digital filters. However, this approximation of the vocal tract does not take into account losses due to friction, heat conduction, and wall vibration, and therefore it is reasonable to expect the bandwidths of the resonances in the vocal tract to differ from those of a detailed model which includes these losses.

Describing the details in deriving the mathematical representation on the propagation of sound in each of the concatenated tube would be beyond the scope of this chapter. However, we encourage the reader to review books by [Rabiner [95], Pg. 83] and [Deller [30], Pg. 169] to get the needed information. From here on, we will just assume that the lossless tube model representing the vocal tract system is characterized by a set of cross-sectional areas of the tube called the reflection coefficients. Basically, these refection coefficients are used to estimate the area function of the vocal tract.

Looking at the vocal tract model in Fig. 2.7, the input and output terminals includes the effects of the glottis in the larynx, $u_g(n)$ and lips, $u_L(n)$. This relationship can be represented by the transfer function of the vocal tract $V(z)$ of the form:

$$V(z) = \frac{U_{lips}(z)}{U_{glottis}(z)} \tag{2.4}$$

The vocal tract transfer function in Eq. (2.4) can be further expanded to be in its simplified form:

$$V(z) = \frac{G}{1 - \sum_{k=1}^{N} \alpha_k z^{-k}} \tag{2.5}$$

where N is the total number (order) of poles in the transfer function, G is the gain parameter and $\{\alpha_k\}$ are the reflection coefficients of the digital filter that depend upon the area function of the vocal tract.

It can be seen that the vocal tract transfer function in Eq. (2.5) represents an all-pole model. The poles of the transfer function correspond to the resonances (formants) of the vocal tract. An all-pole model is a very good representation of vocal tract effects of non-nasal voiced sounds. For nasals and fricative sounds, both resonances (poles) and anti-resonances (zeros) are required. However, if the order of the poles is high enough, an all-pole model gives a very good representation for almost all speech sounds.

In the z-plane, each complex pole (p_i) root pairs $z = p_i e^{\pm \theta i}$ roughly corresponds to a formant in the vocal tract $V(z)$ spectrum. For the vocal tract function $V(z)$ to be a stable system, each pair of poles in the z-plane should be inside the unit circle. The location of the formant frequencies and formant bandwidths calculated from the poles is given by:

$$FMT_i = \frac{F_S}{2\pi}\theta_i$$
$$\text{where} \quad \theta_i = \tan^{-1}\left[\frac{\text{Im}\,aginary(p_i)}{\text{Re}\,al(p_i)}\right] \tag{2.6}$$

$$FBW_i = -\left(\frac{F_S}{\pi}\right)\ln|p_i| \tag{2.7}$$

where FMT_i and FBW_i represents the i^{th} formant and formant bandwidth respectively, and p_i is its corresponding pole in the upper half of the z-plane that have their imaginary parts positive. F_s denote the sampling frequency in Hz.

Thus, the parameters of the vocal tract model shown in Fig. 2.7 provide the resonance (F_1, F_2, F_3, etc.) and the spectral characteristics of the vocal tract.

2.5.3 Radiation Effect Model

A complete model of the speech production process should include the radiation impedance due to the lips, as in reality, the speech signal pressure wave traveling from the vocal tract tube terminates with the opening between the lips. This is accounted for in Fig. 2.7 where the speech signal output from the source is related to the volume velocity at the lips through the radiation impedance model *R(z)*. Looking at the input and output terminals at the radiation model *R(z)*, the final speech output in its z-domain is given as:

$$S(z) = P_L(z) = R(z)U_L(z) \tag{2.8}$$

The radiation impedance model is usually modeled as a filter consisting of one or two zeros. A approximation to the radiation impedance *R(z)* is given by a differencing filter:

$$R(z) = 1 - z^{-1} \tag{2.9}$$

2.6 Acoustic Feature Characterization of the Speech Production System

Since the components in the discrete time-model shown in Fig. 2.7 are a very suitable choice for modeling the human speech production, it is only natural that any objective measurements in deriving acoustic features drawn from this model should be closely tied to the anatomy of the physiological and perceptual phenomena in the speech production mechanism.

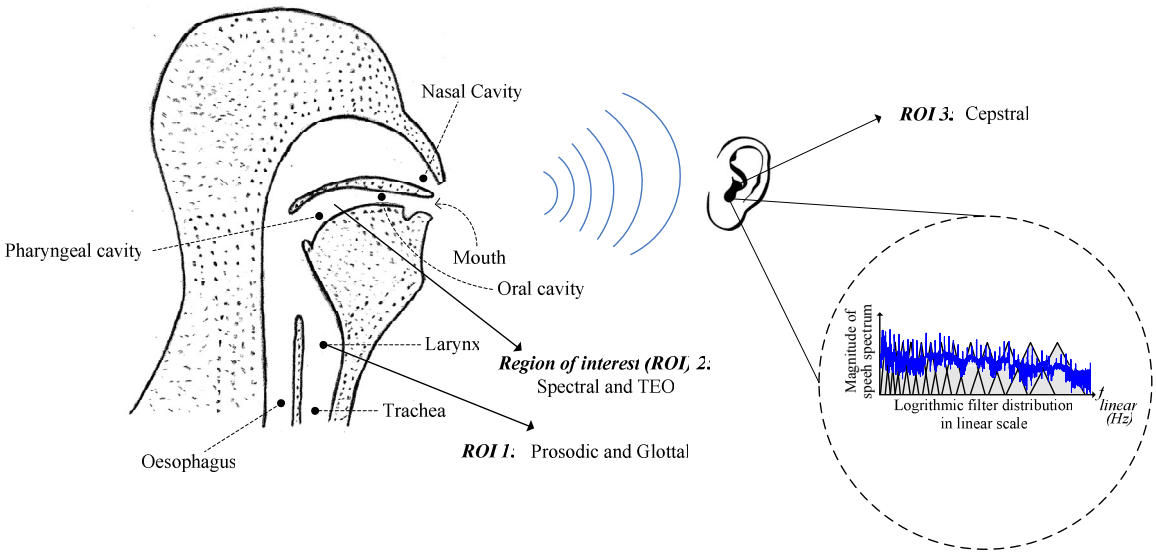

Figure 2.8: Schematic diagram of the human speech production mechanism and the *region of interest* (*ROI*) in characterizing acoustic features of speech.

Fig. 2.8 illustrates five main acoustic feature categories that represents the prosodic (P), glottal (G), spectral (S), TEO-based, and cepstral (C) features. These acoustic feature categories were divided according to the physiological and perceptual components that characterize speech in the human speech production model. In addition, Fig. 2.8 also indicates the *region of interest (ROI)* in the areas of the speech production mechanism where these acoustic feature categories are derived from. As shown in the figure, *ROI 1* (prosodic and glottal feature categories) and *ROI 2* (spectral and TEO-based feature categories) relates to the physiological components whereas *ROI 3* of cepstral feature category relates to the perceptual components. The following gives a brief overview of the acoustic feature categories.

- ***Glottal category:*** The movement of the volume velocity of air flow through the glottis (the gap between the vocal folds) is the excitation source for voiced

speech. The continuous airflow through the glottis which results in quasi-periodic vibration of the vocal folds causes a succession of air pulses known as the glottal flow (pulse). The glottal flow is an important element in defining many speech characteristics. Hence, analyzing the characteristic properties of the voice source is essential in understanding various acoustical cues used in speech production. The glottal pulse is illustrated in Fig. 2.9 where it is formed through the opening and closing of the vocal folds as indicated by the cross-section of the larynx. The glottal pulse waveform, $u_g(t)$ is represented by the closure and opening of the vocal folds. Closure in the glottal waveform refers to the vocal folds being completely closed and opening in the glottal waveform refers to the point when the vocal folds are separated and air flows through the glottis. In Fig. 2.9, one compete cycle (period) of the glottal pulse is denoted as T, the length of the opening phase as T_o and the length of the closing phase as T_c. The reciprocal of the glottal cycle period is known as the fundamental frequency.

- ***Prosodic category:*** The timing, rhythm, intensity and intonation of speech are generally called prosodic features that contribute significantly to the formal linguistic structure of speech communication. The prosodic features may reflect the expressions or feelings of a speaker. For example, prosodic measurements may include the acoustic patterns of fundamental frequency (F_0), the variations between successive F_0, which is called jitter, as well as changes in amplitude, which is called shimmer.

- ***Spectral category:*** From the linear discrete-time source-filter model of speech production shown in Fig 2.7, the source provides the excitation, which is shaped spectrally by the vocal tract (filter). The vocal tract shapes correspond to different sounds with certain spectral characteristics. Therefore, the spectral characteristics of the vocal tract are defined by its shape and length. From an anatomy point of view of the speech production system, the vocal tract is located around the pharyngeal and oral cavities (mouth) as shown by *ROI 2* in Fig. 2.8.

- ***Cepstrum category:*** Like for the spectral category, the cepstrum is related to the filter domain (vocal-tract) represented by the discrete-time model of speech production in Fig. 2.7. However, it is a representation of the perceptual component that characterizes the human voice. The cepstrum is designed for problems that transform speech signals combined by convolution (i.e., voiced speech) to be linearly separable. One of the useful techniques applied to the cepstrum is known as the mel-cepstrum which approximates the human auditory system more closely than the linearly-spaced frequency bands that is used in the normal cepstrum. The mel-cepstrum is based on human perception in how the human auditory system perceives sound as indicated by the *ROI 3* shown in Fig. 2.8.

- ***Non-linear TEO-based category****:* So far, all the acoustic categories (i.e., glottal, prosodic, spectral and cepstrum) that was previously described are based on the assumption that sound propagates linearly along the vocal tract as a plane wave as indicated by the flow of the arrows in Fig. 2.9. This is the reason why the linear

discrete time source-filter model in Fig. 2.7 is a very suitable choice for the representation of these acoustic feature categories in the speech production system. However, according to studies by Teager, [111], [112], [113], this planar propagation flow of sound production may not hold since the true source of excitation actually originates from the vortices located in the false vocal folds (shown in Fig. 2.10) during the closed phase of the vocal folds. The flow of these vortices is nonlinear and is supported by the theory in fluid mechanics [19] as well as by the solution of the Navier-Stokes equation [114]. It is believed that changes in the physiological component of the vocal system induced by stressful conditions such as muscle tension will affect the vortex-flow interactions in the vocal tract [123]. To derive the nonlinear pattern of the vortex-flow interactions, Teager developed an energy operator called the Teager energy operator (TEO) in an attempt to reflect the instantaneous energy of these nonlinear vortex-flow interactions. The analysis of the TEO will be explained more in detail in *Chapter 6*.

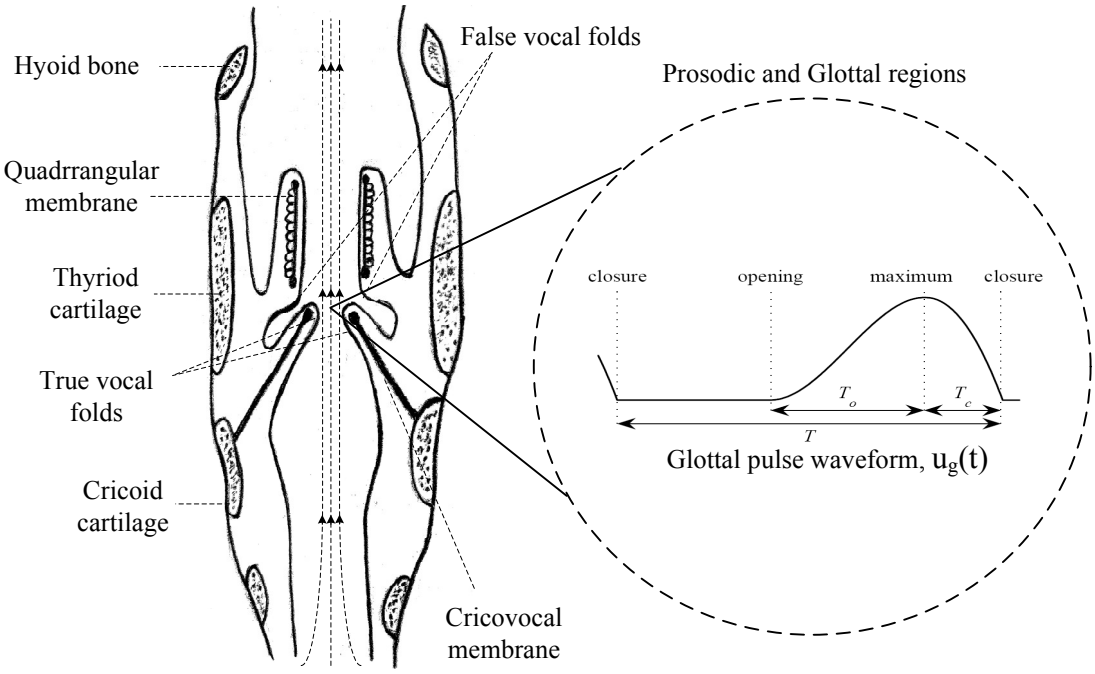

Figure 2.9: Cross-section of the larynx (viewed from the front): Illustration of an ideal glottal pulse waveform (Rosenberg (1971) model) generated from successive opening and closing of the vocal folds.

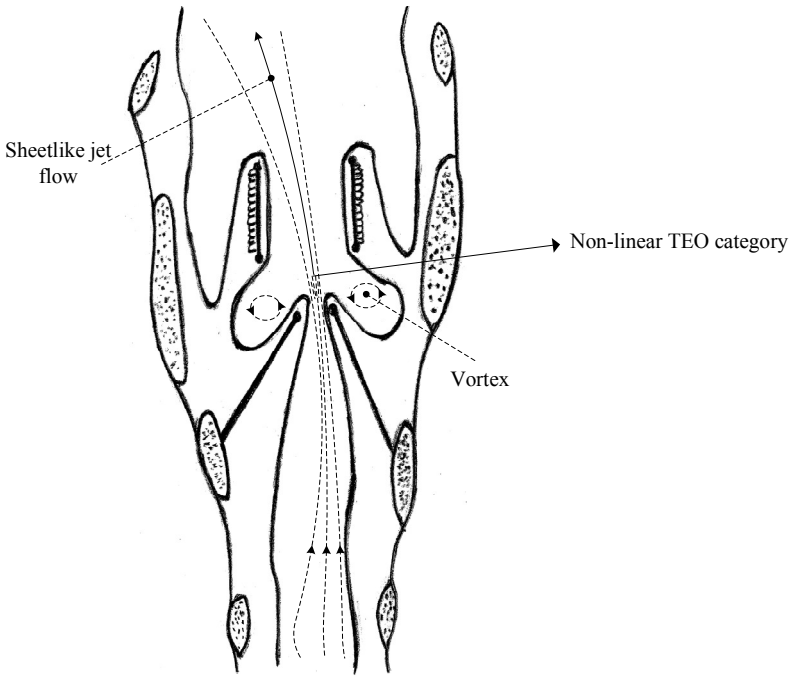

Figure 2.10: Cross-section of the larynx (viewed from the front): A nonlinear interpretation of sound propagation along the vocal tract.

Chapter Three:

PATTERN CLASSIFICATION

"One of the most interesting aspects of the world is that it can be considered to be made of patterns. A pattern is essentially an arrangement. It is characterized by the order of the elements of which it is made, rather than by the intrinsic nature of these elements."

– Norbert Wiener

3.1 Preview

From a signal processing sense, the advancement of technologies has allowed us to process, extract and store vast amount of data. However, extracting all the information from the raw data would be pretty useless if one could not make any sense of it. Therefore, building statistical models from raw data in order to find patterns and trends is crucial in acquiring important information. The process of finding existing patterns from raw data so that the criteria for generalizing decisions can be learned is often known as machine learning or pattern recognition.

The purpose of this chapter is set out to provide an introductory overview to the machine learning techniques of the *1) Gaussian mixture model* and *2) support vector machine* that were implemented in the experiments in this dissertation.

3.2 Gaussian Mixture Model

A widely used model for the distribution of continuous variables is known as the Gaussian (or normal) distribution. For a variable with a single dimension *x*, the Gaussian distribution is of the form:

$$N(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp[-\frac{1}{2\sigma^2}(x-\mu)^2] \tag{3.1}$$

Where μ is the mean and σ^2 is the variance of the normal distribution. The normal distribution is denoted by $N(x \mid \mu, \sigma^2)$ with the argument in the function standing for probability of x given mean μ and variance σ^2.

Translating Eq. 3.1 from a single variable to a D-dimension vector **x**, the multivariate Gaussian distribution is written as:

$$N(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2} \mid \boldsymbol{\Sigma} \mid^{1/2}} \exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})] \tag{3.2}$$

where **x** is a D- component column vector, $\boldsymbol{\mu}$ is the D-component mean vector, $\boldsymbol{\Sigma}$ is the D by D covariance matrix, $\mid \boldsymbol{\Sigma} \mid$ and $\boldsymbol{\Sigma}^{-1}$ are its determinant and inverse respectively. $(\mathbf{x} - \boldsymbol{\mu})^t$ denotes the transpose of $(\mathbf{x} - \boldsymbol{\mu})$.

One of the main practical uses in Gaussian distributions is to model empirical distributions of different random variables. However, this has severe limitations when it comes to modeling on real datasets [12]. To improve these limitations, we therefore consider a linear combination of clusters involving *M* Gaussian densities of the form:

$$p(\mathbf{x}) = \sum_{m=1}^{M} w_m N(\mathbf{x} \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{3.3}$$

where w_m is the mixing coefficients and M is the total number of Gaussian mixtures. From Eq. (3.3) which is termed mixture of Gaussians, it can be seen that each Gaussian density $N(\mathbf{x} \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ has its own mixing coefficient, mean $\boldsymbol{\mu}_m$ and covariance $\boldsymbol{\Sigma}_m$.

Note that these individual Gaussian components are normalized so that:

$$\sum_{m=1}^{M} w_m = 1 \tag{3.4}$$

where $0 \leq w_m \leq 1$. We therefore see that the mixing coefficients satisfy the requirements to be probabilities.

From the sum and product rules in Appendix A, the marginal density is given by:

$$p(\mathrm{x}) = \sum_{m=1}^{M} p(m)\, p(\mathrm{x} \mid m) \tag{3.5}$$

which is equivalent to Eq. (3.3), whereby $p(m) = w_m$ is known as the prior probability in picking the m^{th} mixture component and the densitiy $p(\mathbf{x} \mid m) = N(\mathbf{x} \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ is known as the probability of $\mathbf{x}$ conditioned on m.

Suppose we want to know which Gaussian mixture component did vector $\mathbf{x}$ come from, we can solve the problem by reversing the conditional probability by using Bayes' theorem to give:

$$\begin{aligned} p_{posterior}(\mathbf{x}) &\equiv p(m \mid \mathbf{x}) \\ &= \frac{p(m)\, p(\mathbf{x} \mid m)}{p(\mathbf{x})} \\ &= \frac{w_m N(\mathbf{x} \mid \boldsymbol{\mu}_{m,} \boldsymbol{\Sigma}_m)}{\sum_{l}^{M} w_l N(\mathbf{x} \mid \boldsymbol{\mu}_{l,} \boldsymbol{\Sigma}_l)} \end{aligned} \tag{3.6}$$

In order to find the parameters of $\mathbf{w} \equiv \{w_1,\ldots,w_m\}$, $\boldsymbol{\mu} \equiv \{\boldsymbol{\mu}_1,\ldots, \boldsymbol{\mu}_m\}$, $\boldsymbol{\Sigma} \equiv \{\boldsymbol{\Sigma}_1,\ldots, \boldsymbol{\Sigma}_m\}$ in Eq. (3.3) & Eq. (3.6) we can compute the log likelihood function given by:

$$\ln p(\mathbf{X} \mid \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln[\sum_{m=1}^{M} w_m N(\mathbf{x}_n \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)] \tag{3.7}$$

where $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$. The goal is to find the optimized parameters of each Gaussian mixture component by maximizing the likelihood function using iterative optimization techniques such as the *expectation maximization (EM)* framework. The EM is an iterative method in which each iteration procedure involves a two stage process. The first stage computes the expectation (E)-step of the log-likelihood evaluated using the current estimate of the latent variables. The second stage known as the maximization (M)-step, re-estimates the parameters maximizing the expected log-likelihood found in the E-step. This process is repeated until a convergence criterion is satisfied. The expectation maximization process is illustrated in the plots depicted in Fig 3.1. Plot (a) shows an example of data points in two-dimensional Euclidean space colored in green. Plot (b) shows the first stage of the E-step where the mean $\boldsymbol{\mu}_{init}$, covariance $\boldsymbol{\Sigma}_{init}$ and mixing coefficient w_{int} parameters in each of the two Gaussian components (highlighted in blue and red) are initialized. Next, the posterior probabilities of the data points to each Gaussian component are evaluated. The data points highlighted in blue (○) indicates that the posterior probabilities are closer to the blue Gaussian component while the data points colored in red (●) are closer to the red Gaussian component. The data points that have probabilities belonging to either Gaussian component are depicted by a pink triangle (▲). Plot (c) shows the first M-step in re-estimating the new means, covariance and

mixing coefficient of the Gaussian components from the data points that were newly assigned to either the red or the blue component based on their current posterior probabilities. The log likelihood probability is then maximized by repeating the cycle of the E and M steps until convergence criteria is met as illustrated in plots (d) – (i).

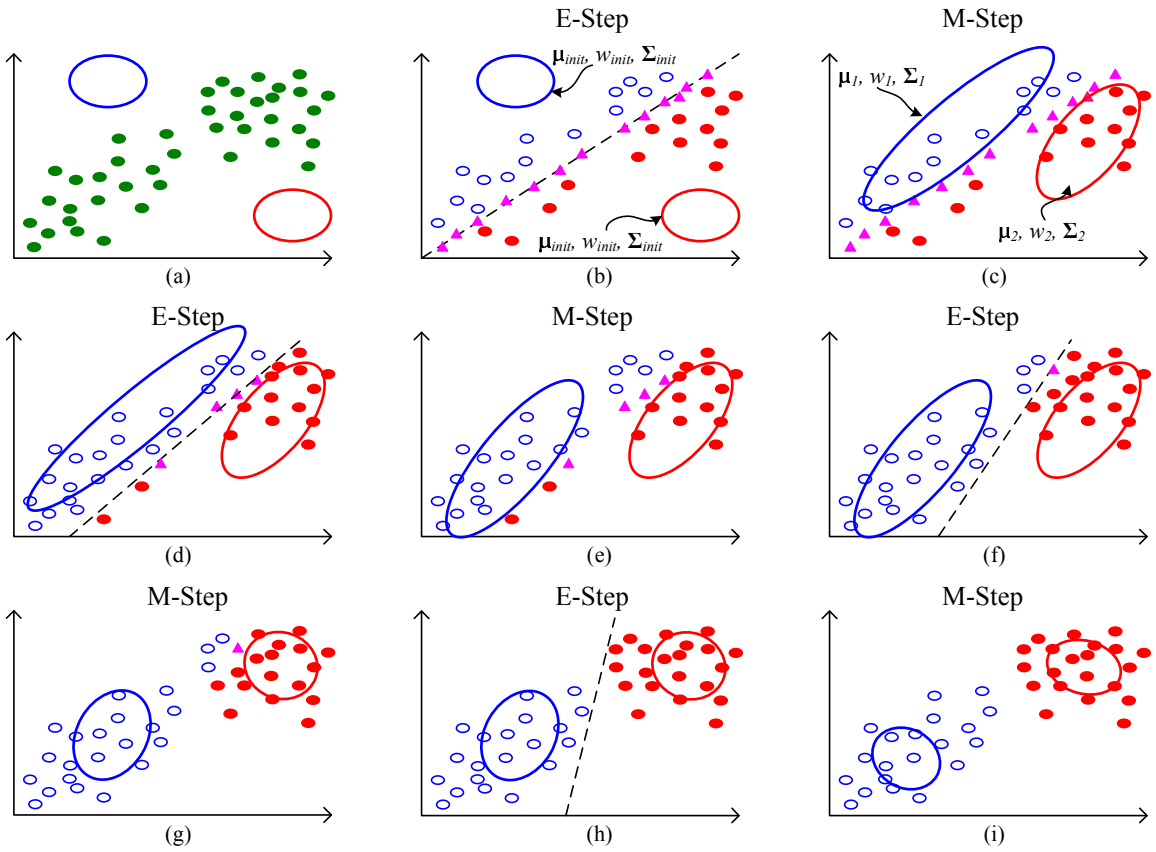

Figure 3.1: Illustration of the expectation maximization (EM) iterative optimization technique for a mixture of two Gaussian components (adapted from Bishop [12]). (a) Green points denote an example of a dataset in two-dimensional Euclidean space. (b) First stage, expectation (E) step: Initialization of the parameters mean $\boldsymbol{\mu}_{init}$, covariance $\boldsymbol{\Sigma}_{init}$, mixing coefficient w_{int} and evaluating the posterior probabilities of the data points to each Gaussian component. (c) Second stage, maximization (M) step: Re-estimating the parameters using the current posterior probabilities and calculating the log likelihood. (d)-(i) show subsequent E and M steps through to the final convergence of the log likelihood.

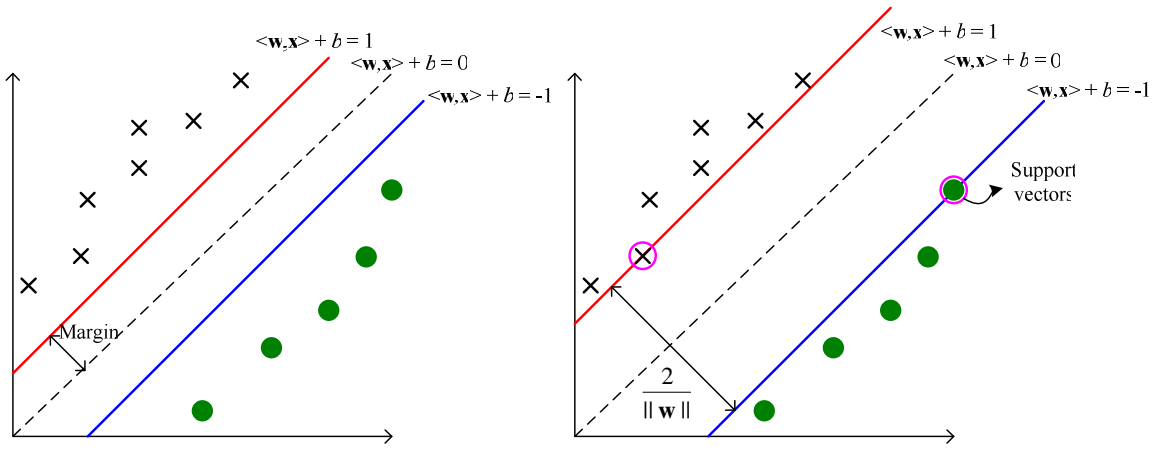

Figure 3.2: A maximal margin hyperplane with its support vectors highlighted in circles.

3.3 Support Vector Machine

Support vector machines (SVM) [13] are linear classifiers that have been applied to many classification problems, generally yielding good performances compared to other algorithms especially in binary classification problems [12]. We begin our discussion on SVM with a two-class classification problem using a simple form of linear classification. Given N number of data points with d-dimensional features (i.e., variables) $\mathbb{R}^d$, where $\mathbf{x} \in \mathbb{R}^d$ is the d-dimensional input vector and $y \in \{-1, 1\}$ is a class label for binary classification. The decision function using linear models is of the form:

$$y = sign(\langle \mathbf{w}.\mathbf{x} \rangle + b) \tag{3.8}$$

where $\mathbf{w}$ is the weight vector, b is the bias, y_i is its associated label, $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are the parameters that control the function of the decision boundary given by Eq. (3.8). This equation assumes that the training dataset is linearly separable in input space.

SVM forms a linear classifier through the concept known as the margin, which is defined as the smallest distance between the decision boundary and any of the samples as illustrated in Fig. 3.2 (left). The objective of the decision boundary in SVM is to maximize the distance of the margin between two parallel hyperplanes which separates two groups (classes). This is illustrated in Fig. 3.2 (right) where the linear classifier defined by the hyperplane $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$ is midway between the separating hyperplanes. The location of the boundary known as support vectors are indicated by the highlighted circles in Fig. 3.2 (right). The margin can be computed explicitly as $2 / \|\mathbf{w}\|$ as shown in Fig 3.2 (right).

So far, we have assumed that the training data is linearly separable in the two-dimensional input space. In the case where the training data cannot be linearly separable, we can map the non-separable data from the input space to a higher dimensional feature space whereby the data is transformed to be linearly separable in which the linear models can be used. Hence, Eq. (3.8) can be re-written in the form:

$$y = sign(\langle \mathbf{w}.\phi(\mathbf{x}) \rangle + b) \tag{3.9}$$

where $\phi : \mathbf{x} \rightarrow \phi(\mathbf{x})$ is a non-linear map from the input space to some feature space. This means that we can build non-linear machines in two steps: first a fixed non-linear mapping transforms the data into a feature space $\phi(\mathbf{x})$, and then a linear machine is used to classify the data in the feature space.

One important property of linear learning machines is that they can be expressed in a dual representation (refer to Cristianini and Shawe-Taylor [26] for more details).

This means that Eq. (3.9) can be expressed as a linear combination of the training points, so that the decision rule can be evaluated using just inner products between the test point and the training point.

$$y = \sum_{i=1}^{l} \alpha_i y_i \langle \phi(\mathbf{x}_i)\phi(\mathbf{x}) \rangle + b \tag{3.10}$$

where the inner product $< \phi(\mathbf{x}_i).\phi(\mathbf{x}) >$ can be replaced with a kernel function K ($\mathbf{x}_i$, $\mathbf{x}$) that obeys Mercer's condition. Mercer's condition states that any positive semi-definite (i.e., K $\geq$0) kernel K ($\mathbf{x}_i$, $\mathbf{x}_j$) can be expressed as a dot product in high-dimensional space. Thus, this is equivalent to mapping the feature vectors into a higher dimensional feature space before using a hyper plane classifier as was illustrated in Fig. 3.2. The resulting support vector machine will give exact separation of the training data in the original input space $\mathbf{x}$, although the corresponding decision boundary is nonlinear. However, in real-world datasets, the class-condition distributions may overlap, in which this case exact separation of the training data can lead to poor generalization.

We therefore need a way to modify the support vector so as to allow some of the training points to be misclassified which requires the solution in searching for the maximum margin classifier of the following optimization problem:

$$\text{minimize} \quad \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{l} \xi_i$$

$$\text{subject to} \ \ y_i(\mathbf{w}^T \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \tag{3.11}$$

where $\mathbf{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$, $\xi \in \mathbb{R}^l$, $\mathbf{x}_i \in \mathbb{R}^d$ is the i^{th} data sample with a d-dimensional feature vector, $y_i \in \{-1, 1\}$ is the class labels, and l is the number of training points. The function ϕ maps the training vectors $\mathbf{x}_i$ into a higher dimensional space. The first constraint dictates that points with equivalent labels are on the same side of the line. The slack variable ξ allows data to be misclassified while being penalized at rate C in the objective function in Eq. (3.11). Therefore this allows SVM to handle non-separable data in real-world situations. The kernel function $K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi(\mathbf{x}_i)^T . \phi(\mathbf{x}_j)$. Below defines several classic kernel functions, each corresponding to a different implicit mapping to feature space:

- linear: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$
- polynomial: $K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma\, \mathbf{x}_i^T \mathbf{x}_j + r)^d$, $\gamma \geq 0$
- radial base function (RBF): $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\, \|\mathbf{x}_i - \mathbf{x}_j\|^2)$, $\gamma \geq 0$

Chapter Four:

LITERATURE REVIEW

"I am convinced that He (God) does not play dice."

– Albert Einstein

4.1 Preview

According to studies by Scherer (1981), the human vocal-auditory modality is considered to be one of the most important channels recognized by psychologists in providing a better understanding to the diverse roles played by different forms of emotion regulation in different types of psychopathology. According to the Diagnostic and Statistical Manual of Mental Disorders version IV (DSM-IV) [6], problems relating to the aspect of emotion or emotion regulation characterized more than 75% of the diagnostic categories of psychopathology [9], [64]. The psychopathology category that is discussed in this dissertation is focused on clinical depression, particularly in speech communication during social interactions as this is where expression and communication of emotional states commonly occurs [101]. The goal of this chapter is designed to: (1) provide a basic understanding of the effects on speech production mechanisms in respect to the physiological, biological and linguistic aspects linking to emotional arousal, (2) discuss briefly on voice as an indicator of emotions and (3) life stresses and (4) provide a review of past literature in objective measurable acoustic descriptors of speech correlating to depression.

4.2 Emotional Arousal on the Physiology of Speech Production

There has been an enormous amount of empirical research directed at describing the physiology of emotion. The aim of these studies which is still ongoing is to search for physiological patterns that might underlie each discrete emotion. This was unlined in historic works derived from James (1884), who reported that emotion can be equated with awareness of a visceral response; and therefore each emotion should be accompanied by a unique pattern of physiological response. Wenger (1950) who tested this proposition shared the same physiological based views and went further on by suggesting that the perception of emotional stimuli depended on the pairing of conditioned and unconditioned stimuli, following which the arousal of the autonomic nervous system (ANS) leads to visceral responses that results in muscular responses and verbal action. Lindsley's (1950, 1951, 1957, 1970) theory of the neurophysiological basis of emotions documented that there are arousal/motivation mechanisms underlying emotion and maintains that it is the limbic system of the human brain that control emotional expression and emotional and motivational behavior. Gellhorn (1964) and Gellhorn and Loufbourrow (1963) suggested that the basis of emotion is the integration of somatic nervous system (SNS) and autonomic nervous system (ANS) activities in what is termed the *ergotropic* and *trophotropic* activities. It was suggested that when emotions are aroused, the *ergotropic* and *trophotropic* balance must be altered by both neurogenic and hormonal processes.

These concepts in the literature of past researchers offer the prospect that physiological measurements might offer a way in accessing a person's emotions directly.

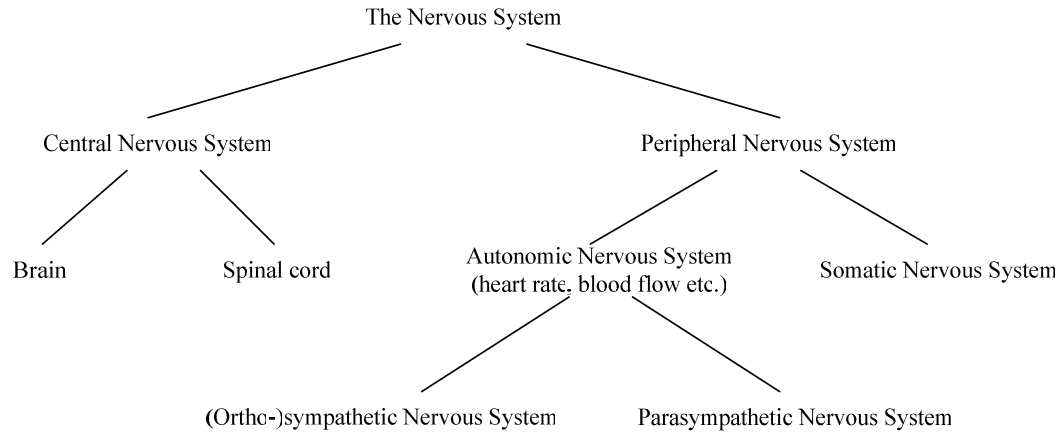

Figure 4.1: Major divisions of the human nervous system.

However, the real discriminative power in physiological measures is limited both in principle and in practice for supporting the hypothesis that emotions are associated with a unique pattern of physiological activity [24].

From the perspective of emotion theory in speech, Scherer (1979) pointed out that there had been no published reports on the effect of emotional arousal on the physiology of speech production [100]. However, Scherer tried to break this barrier, by conducting numerous studies [98], [99], [100], [102], [103], attempting to predict the possible effects of emotional arousal on speech production. The four main processes (as previously described in *Section 2.4.1*) of the speech production mechanism in producing speech sounds include the respiratory, phonation, articulation and oro-nasal. Scherer [100] gathered a considerable amount of evidence that emotional arousal produces changes in these processes. We shall briefly review the physiological and neurological systems that control these mechanisms relating to the speech production system.

Emotions displayed in humans involve the entire nervous system. Fig. 4.1 shows a diagram of the major divisions in the human nervous system. The central nervous

system (CNS) is the part of the nervous system that integrates the information that it receives and coordinates the activity to all parts of the body. It contains the majority of the nervous system and consists of the brain and spinal cord. Together with the peripheral nervous system (PNS), it has a fundamental role in the control of behavior.

The autonomic nervous system (ANS) is the part of the peripheral nervous system controlling visceral functions (heart rate, blood follow, respiration, perspiration, etc.) with most of its actions occurring involuntary. It is divided into two subsystems called the parasympathetic nervous system (PSNS) and (ortho-) sympathetic nervous system (OSNS). Another part of the PNS is the somatic nervous system (SNS) which is associated with voluntary control of body movements.

As was illustrated in the speech chain diagram in Fig. 2.2, the process of speech communication stems from a thought from a speaker's brain. This thought is then converted into a linguistic structure by forming words based on grammatical rules of a particular language. Paralinguistic information such as tonal patterns of pitch, syllable stresses, and timing to form rhythmic patterns are added to convey the speaker's emotional expressions and its overall meaning. The brain in the CNS houses two important structures called the neocortex and the limbic system. The limbic system is part of the brain that houses the primary centers of emotion. It includes the amygdala, which is important in the association of events with emotion. The amygdala comes into play in situations that arouse feelings such as fear, pity anger or outrage. The limbic system is also related to expression and mediation of emotions and feelings, including emotions linked to attachment. The neocortex makes language, including speech and writing possible. It renders logical and formal operational thinking possible and allows us to see

ahead and plan for the future. The neocortex also contains two specialized regions, one dedicated to voluntary movement and one to processing sensory information. On-going communication between the neocortex and the limbic system links thinking and emotions [14] with each influencing the other. This interplay of memory and emotion, thought and action is the foundation of a person's individuality.

Speech production in the linguistic sense is obviously mostly controlled by the neocortex, a process that works mainly via specific motor commands which produce appropriate phonatory and articulatory movements for desired speech sounds. The intended vocal effects are mostly produced by phasic activation of the muscles serving phonation and articulation [103].

The effects of emotional arousal on the vocalization process that are primarily controlled by the limbic system are much more diffuse and complicated. They are produced via tonic activation of the somatic nervous system (in particular the striated musculature), and sympathetic, as well as parasympathetic, activation of the autonomic nervous system. Given the predominance of the striated musculature in producing vocalization, many of these emotional effects are likely to be mediated via the somatic nervous system. However, direct sympathetic or parasympathetic effects, such as respiration changes and the secretion of mucus, will also affect the nature of the vocal output.

4.3 Vocal Indicators of Emotions

The evolutionary bases of emotions came from Cicero [20] and Aristotle [7] who suggested that each emotion is associated with a distinctive voice tone. However, it was

arguably Darwin's (1872) theory of evolution in communication that led to the first comprehensive description of the sounds associated with emotions. This theory was further infused with the description in the previous section that various physiological and neurological changes associated with emotions can directly affect the speech production mechanisms. With the concept of physiological measurements providing a possible way in accessing a person's emotions acoustically through his/her speech, researchers [23], [34] dating as far back as the 1930s, have spent a lot of time and effort establishing kinds of emotional phenomena based on vocal cues from a person's speech. Indeed, with the growth of new methodologies on emotional domains in past to the present years, there have been an explosion in vast amounts of literature on emotions and speech. However, these different methodologies also tend to be associated with different and wider range of emotional domains. This taxonomy of emotion categories used to describe emotional states poses a problem to the research community as the agreement to consider implementing the same categories are not widely accepted, making it difficult to accurately state hypotheses and predictions clearly [67].

The basic primary emotions that are most widely studied are anger, sadness, joy, fear, disgust, surprise [25]. Trained labelers pay particularly close attention to voice cues such as tone, loudness speaking rate, intensity and breaks in patterns of speech in order to make judgments in categorizing emotions. Therefore, it would seem reasonable to hypothesize that acoustic features of prosody would be able to differentiate between different categories of emotions on the basis of vocal displays as they have the closest relation to human perception. Prosody is defined as the pitch, intonation range, tempo, rhythm, pausing that reflects the emotion of the speaker. Prosodic features along with

their functionals have been the most widely studied feature in emotion research. Cowie (2001) provides a very extensive literature review on the types of acoustic analysis applied in the field of emotion recognition in speech. There seems to be a consistency in research findings in the acoustic correlates of pitch and its functionals (i.e., statistical measures over an utterance of range, mean, median and variability), intensity and duration in recognizing primary emotions of anger, happiness, sadness and fear in speech. However, due to the different methodologies defined by different researchers associated with emotional domains as was previously discussed, it is not surprising that there have also been discrepancies. For example, high speech rate was reported from acoustic measurements in emotions of anger [29] and fear [63]. However, the opposite (slower speech rate) was reported for anger in [119] and for fear in [110].

In terms of accuracies in the classification of primary emotions, it is possible to gain recognition rates as high as 100% in databases comprising of acted emotions in speech [11]. Unfortunately, for realistic databases of spontaneous speech, the research evidence available in emotion recognition still suffers from poor overall accuracy rates with the performance of a two-class problem reported to be less than 80% [77] and for a four class problem to be less than 60% [8], [11]. Although the accuracies are low, this is reasonable considering the fact that even reports on perceptual classification of these emotions ranges from 40-60%. These authors suggested that the use of acted rather that natural stimuli may result in emotional portrayals that do not clearly correspond with naturally produced vocal expressing emotion [8].

4.4 Vocal Indicators of Psychological Stress

Stress is the body adjustment which requires physical, mental or emotional adjustment to a change that exceeds the adaptive capacity of an organism, resulting in the interaction of psychological and biological changes that may result in negative health problems [21]. The signs of stressful outcomes in a person and ones failure to cope with it may act as a predictive factor in recognizing psychopathological symptoms such as depression. There is a high possibility that stress related changes can affect vocalization in such a way that the tension state will apply some form of pressure to the speaker which in turn will result in perturbation of the speech production process becoming noticeable in the acoustic signal. Therefore voice and speech cues are important indicators of stress in humans.

One of the key observations in physiological responses on stressful events that effect the speech production has primarily been focused on the sympathetic activation. Sympathetic activation occurs in cases of stressful activities to trigger in what is called "the fight or flight response" theory. This theory which was first described by Cannon [16] states that animals react to threats by giving out a hyper-arousal response. This hyper-arousal response generates a discharge in the sympathetic nervous system which prepares the animal for fighting or fleeing. This "fight or flight" reaction has also been found to occur in humans during times of stress. The sympathetic arousal is characterized by changes in cardiac activity, respiration patterns and muscle tension. The changes in respiration and muscle tension are the ones that have a direct effect on the voice and speech production. Scherer (1981) describes that the respiration is likely to affect the sub-glottal pressure in phonation where general muscle tone will affect the operation of the extra and intralaryngeal mechanisms involved in phonation, as well as the characteristics

of the vocal tract resonance walls and articulatory mechanisms. In addition, disturbances in the coordination of neural impulses which determine phonation and articulation activities may also affect speech production. In studies concerning gender differences in fight or flight reaction, there has been reported evidence that men tend to become more aggressive or more withdrawn than do women under stressful situations due to gender role socialization (i.e., women tend to be more open in discussing their problems within their family or friends) and biological factors [71].

A stimuli or situation in an environment that tends to produce a stressful response are termed as "stressor". Different types of stressors vary widely in literature which includes:

1) Physical or mental activity (i.e., pressure breathing)
2) Physiological stimulation (i.e., narcotics, alcohol, sensory deprivation)
3) Perceptual effect (i.e., noise, poor communication channel)
4) Aversive psychological stimulation (i.e., conflict, offense).

Hansen *et.al* (2000) labels these stressors in different categories depending on the order of the stressor that has the lowest (zero order) to highest (third order) effect on the speech production system. The chronological order of the stressors used in [47] is the same order as the ones shown above (zero order- *physical*, first order - *physiological*, second order - *perceptual* and third order – *psychological*). In relation to psychopathological symptoms, depression and emotions are classified as third order stressor according to studies in [47].

There have been many empirical studies on the vocal correlates of stress particularly in the analysis of the fundamental frequency (F0), because understandably in theory, F0 should reliably increase with the increase in stress or emotion tension [100]. For studies relating to realistic life stress scenarios, particularly those of pilots during in-flight emergencies, several F0 related parameters such as mean, short-term perturbations and long-term variability are among the measures often identified as good correlates with elevated levels of emotional stress [66], [118]. However, recent studies from Zhou (2001) found that nonlinear features derived from the Teager energy operator (TEO), particularly in the proposed critical-band based TEO autocorrelation envelope method, outperformed the F0 parameters in the classification of actual stress in the investigation of motion-fear of subjects on an amusement park roller-coaster ride. The TEO autocorrelation envelope method and F0 yielded a 97.9% and 89.4% correct accuracy rates respectively in the actual stress domain.

The etiology of depression encompasses a variety of factors operating together in giving rise to a depressive episode. One of the factors which have been the focus of attention is in the importance of life stress in the cause of depression [1], [81]. However, this complex relationship between emotional stress and whether it leads to an onset of clinical depression still remains unclear because (1) not all people with severe stress develop or succumb to depression, and (2) not all depressed people report recent life stress. Therefore, although emotional stress might generally be important cues to depression, investigators also proceed into other factors in understanding depression.

4.5 Vocal Indicators of Clinical Depression

As was previously discussed in *Section 4.2*, based on the psychological evidences that emotional arousal produces changes in the speech production system, the use of acoustical properties of speech as indicators of depression and suicidal risk have also been studied since the mid-1980's. Since depression and suicidality manifest themselves through certain emotional changes, these studies are closely related to the studies of emotion recognition in speech. Depressed speech has been consistently characterized by clinicians as dull, monotone, monoloud, lifeless, and "metallic" [86]. From a subjective assessment, Darby & Hollien (1977) conducted a pilot study of severely depressed patients and found that listeners could perceive noticeable differences in prosody characteristic of speech such as pitch, loudness, speaking rate, and articulation from depressed patients before and after treatment. Thus, this had led to numerous amounts of studies in objective measurements using speech parameters that measure these prosodic characteristics.

In Nilsonne (1987), the rate of change of voice fundamental frequency F0 during mental depression was measured. The total of 16 patients were tested during depressive episodes and then after recovery. It was showed that the standard deviation of the rate of F0 change, the mean absolute rate of F0 change, the standard deviation of F0, and the relative occurrence of silent intervals showed a strong correlation with depression.

In Kuny (1993), a sample of 30 depressive patients was investigated during the course of recovery from depression. The recovery progress was assessed based on changes in symptom ratings and through changes in speaking behavior and voice sound characteristics. The results revealed several voice sound characteristics to be closely

related to the time course of recovery from depression. In particular, the parameters “F0-amplitude”, “F0 6dB-bandwidth” and “F0-contour” which assesses a speaker's voice timbre, as well as the parameters “energy” and “dynamics” which assesses a speaker's mean loudness and the variation of loudness over time, displayed consistently high correlations with depressive symptoms.

In Alpert (2001), fluency of speech and prosody were rated and compared for 22 elderly depressed patients and an age-matched normal control group. It was concluded that depression was strongly correlated to speech acoustics.

In 1996, Ellgring and Scherer [33] presented the framework of a major longitudinal study of depressive disorders conducted at the Max-Planck-Institute for Psychiatry, Munich. In the presented experiments 11 female and 5 male depressives were audio-video-recorded while speaking with clinical interviewers. For selected utterances during depressed and recovered mood states, several voice and speech parameters were obtained using digital analysis techniques. The results showed that an increase in speech rate and a decrease in pause duration are powerful indicators of mood improvement in the course of therapy (remission from depressive state). In female, but not in male patients, a decrease in minimum fundamental frequency of the voice predicted mood improvement. These effects were discussed with respect to neurophysiological, cognitive, and emotional factors that have been suggested in the literature as possible causes for the patterns of motor expression observed in depressives. The results also point to the urgent need to systematically study gender differences in depressive speech behavior.

In 1980 professor Stephen Silverman, a clinical psychologist, observed a distinctive quality in the pattern and tone of the voices of patients who were very likely to

attempt suicide in the near future [97]. Due to Silverman's initiative, in 1994 an interdisciplinary research team led by Professor Richard Shiavi at Vanderbilt University was established. In [37], [38], members of the Vanderbilt team investigated speech as a psychomotor symptom of depression and suicidality. Acoustical and statistical analyses were performed on clinically diagnosed populations to determine if the acoustical properties of speech change with depression severity, and if they can be used to classify the mental health condition of individual subjects. In the first stage of research, speech samples of three groups: control (10 female), dysthymic (17 female), and major depressed (21 female) patients were tested. In the second stage of research, acoustical properties of speech of normal (24 male), major depressed (21 male), and high-risk suicidal (22 male) patients were analyzed. Features derived from the formants and the Power Spectral Density (PSD) was found to be the best discriminators of class membership in both male and female patients. The Amplitude Modulation (AM) features emerged as strong class discriminators of the male classes. Features derived from F0 were generally ineffective discriminators in both male and female studies. The latest reports from the Vanderbilt team [90], [91] investigate acoustic properties of speech for near-term suicidal risk assessment. The effects of suicidal state on speech source (vocal cords) and speech filter (vocal tract) were investigated separately. The speech source study investigated the changes in vocal jitter and slope of the glottal flow spectrum as indicators of near-term suicidal state. The study of the speech filter concentrated on analysis of the mel-scale cepstral coefficients (MFCCs). The acoustic features of vocal jitter and the glottal spectral slope sensitivity was tested among 10 male non-depressed (control), 10 male major depressed and 10 male near-term suicidal patients using a two-

sample maximum likelihood (ML) analysis. The mean vocal jitter was found to be a significant discriminator only between suicidal and non-depressed groups with 80% of correct classification results. The slope of the glottal source spectrum was found to be a significant discriminator between major-depressed versus suicidal (75% of correct classifications) and control versus major-depressed (90% correct classifications) classes. Classification based on the combined feature set (jitter and glottal source spectrum) showed an improvement of the correct classification rate, reaching 85% correct results for control versus suicidal classifications, 90% for control versus depressed, and 75% for depressed versus suicidal classifications.

In the most recent studies, Moore (2008) describes an analysis of variation in prosodic feature statistics and glottal features of speech for patients suffering from a depressive disorder and uses these results in an automatic classification of speech into two classes: non-depressed and depressed. The classification based on glottal statistics (F0, Energy Deviation Statistics and Energy Median Statistics) achieved accuracy ranging from 67% to 94%. The classification results based on glottal features showed accuracy ranging from 87% to 100%.

Chapter Five:

DATABASE

"Researchers should make friends with their data"

– Robert Rosenthal

5.1 Preview

The database described in this chapter was obtained through a collaborative effort with the Oregon Research Institute (ORI), USA. The database contained adolescents recruited from community high schools participating in family interactive tasks with their parents that were video and audio recorded. The participants comprised of two groups: (1) adolescents that were diagnosed with major depressive disorder (MDD) and (2) adolescents that were healthy with no clinical disorders. For the purpose of this study, only the audio data from the adolescents' speech samples were extracted in the preparation of the speech corpus.

The material in this chapter will provide: (1) the recruitment procedures in the data collection of participants, (2) the design procedures of the family interactive tasks between parent and adolescent, (3) the database annotation and the preparation of the speech corpus for the depressed and control group.

5.2 Database Collection

5.2.1 Participants

The collection of participants in the database occurred within a span of two years and was formulated by researchers from the Oregon Research Institute (ORI), USA. The participants comprised of 152 adolescents (52 males and 100 females), aged between 14 to 18 years old. The participants represented the West Oregon, USA community sample. For the participant to be eligible in the investigation, adolescents had to meet the research criteria for placement in one of two groups (Depressed, n = 75 or Healthy, n= 77) and live with at least one parent/permanent guardian. Adolescents were excluded if they evidenced comorbid externalizing or substance dependence disorders or were taking either Serotonin Norepinephrine Reuptake Inhibitors or Tricyclic antidepressants; these exclusion criteria were relevant to the collection of cardiovascular data that is not used in the current report.

Through the evaluation of self-report and interview measures, the depressed adolescents met the Diagnostic and Statistical Manual of Mental Disorders (DSM) version IV criteria [6] for a current episode of major depressive disorder (MDD). The median disorder duration was 13.5 weeks (range 2-284). Approximately 43% of the depressed adolescents had experienced a previous episode. The median age at first onset was 14.67 (range 7–18). Rates of current and lifetime comorbidity were 28% and 39%, respectively. The healthy, non-depressed (control) adolescents did not meet the diagnostic criteria for any current psychiatric disorder and had no lifetime history of psychopathology.

For the purposes of the larger study (i.e., research in behavioral psychology) carried out by psychologist from ORI in which data for this report were derived, it was important to ensure similarity on demographic measures between the control and depressed groups of the study. As such, healthy participants were matched to depressed participants on adolescent age, gender, ethnicity, and the socioeconomic characteristics of their schools [108]. In summary, although the two samples were well matched on many demographic variables, the depressed participants came from households with somewhat lower socioeconomic status, and had mothers with higher levels of depressive symptoms. These differences are not surprising, and reflect well-established associations between adolescent depression, low socioeconomic status, and maternal depression [61]. The processes involving the recruitment (i.e., school screening), assessment procedures (i.e., diagnostics in screening depression), questionnaires and interview measures that were carried out by ORI can be found in [108]. The demographic data of the participants are presented in Appendix A- Table A1.

5.2.2 Behavioral Observation Data

Families (adolescents and parents) were required to participate in three different types of family interactions: *event planning interaction (EPI)*, *problem-solving interaction (PSI)* and *family consensus interaction (FCI)*. The three family interaction tasks were designed to access behavioral characteristics of individuals in these interactions. The observations of the family interactions were video and audio recorded in a quiet laboratory room at ORI with the necessary acoustics to provide relatively clean audio samples. Family

members were seated a few feet apart as would be typical for a discussion between familiars. The recording setup of a family interaction is as follows:

(1) Video equipment: VideoBank™ system

(2) Audio equipment: Audio Technica (model: ATW-831-w-a300) lapel wireless microphones

(3) Observer's equipment: OS-3 hand-held microcomputers (observational systems, Inc.)

The video equipment setup in (1) captures two camera feeds, one camera was positioned directly at the adolescent and the other positioned at the parents/parent.

The lapel wireless microphones in the audio equipment setup in (2) were placed on each of the participants (adolescents and parents) shirts at the chest level.

The observer's equipment in (3) was for trained observers to code the family interactions in numeric form of the verbal content and affective behavior of targeted individuals and each family member with whom they interacted.

Although participants were also outfitted with other sensors measuring physiological signals such as electrocardiograph (ECG), impedance cardiogram (ICG), skin conductance, respiratory and blood pressure, they did not impede speech behavior. Each of the three interactions was discussed for 20 minutes, resulting in a total of 60 minutes of observational data (video and audio recordings) for each family. The full 20 minutes were always used and the order of interactions was fixed: EPI, PSI, and FCI. Below summarizes the description of these interactions found in Appendix A- Table A2.

1. ***Event planning interaction (EPI) -*** this discusses the planning of a vacation trip with the family.
2. ***Problem-solving interaction (PSI) -*** upon the start of this interaction, parents and adolescents mutually agreed on two topics of disagreement that were completed from a questionnaire. Each family unit was then asked to discuss the problems and try to come to some resolution that was mutually agreeable to all parties.
3. ***Family consensus interaction (FCI) -*** this family discussion involves the imagination of writing a book chapter for a publisher that reflects the shared perspective on the given theme by both adolescent and their parents.

5.3 Database Annotation

The Living in Family Environments system (LIFE; Hops, Biglan, Tolman, Arthur, & Longoria, 1995) was used to code adolescent affective behavior during family interactions. The LIFE system is a behavioral event-based coding system designed to describe the specific timeline of various emotions (called affect codes) and verbal content (called content codes) displayed by the participants during the course of the interaction. Trained observers, blind to diagnostic status, coded the adolescents' nonverbal affect and verbal content in real time. The LIFE code is composed of 27 content codes and 10 affect codes (contempt, anger, belligerence, neutral, pleasant, happy, caring, anxious, dysphoric and whine). The description of the emotion (affect) codes is presented in Appendix A-Table A3. When recording the participant's display of affect, trained observers pay particular attention to the face, voice, and body posture (the instructions for coding an

affect and verbal content can be found in [49]). Twenty percent of interactions were coded by different pairs of observers. The coding results were positively assessed for an inter-observer agreement [50].

5.4 Speech Corpus – Experimental Group (Depressed and Control)

For the purpose of this dissertation, only speech from the microphones was investigated. The speech was recorded using 2-channels (i.e., adolescent; parents) and only the audio recordings from the channel belonging to the adolescents' microphone were analyzed. The speech of the adolescents was then segmented from the recordings based on the time annotations containing these emotions and verbal content that were LIFE coded by expert observers. The average duration of the speech segments was around 2 to 3 seconds long and its original sampling rate was decimated by a factor of 4. This was done by first using an anti-aliasing filter, followed by down-sampling the audio signal from 44.1 kHz to 11.025 kHz sampling rate.

From the 152 eligible participants that were recruited and assessed by ORI as mentioned in Section 5.2.1, 139 participants (46 males and 93 females) were selected for the speech corpus in our experiments. The rest of the participants recordings were not utilized because of missing video and audio recordings of the interactions or missing LIFE codes of the participants. The depressed group consisted of 68 adolescents (19 males and 49 females) and the control group consisted of 71 adolescents (27 males and 44 females). Table 5.2 and Table 5.3 contain the information (subject ID, gender, diagnosis, number of parents and number of utterances) of the male adolescents in the depressed and control group respectively. The female adolescents information of the

depressed and control group is presented in Table 5.4 and Table 5.5 respectively. The average number of utterances for each adolescent from the speech corpus was approximately 278, 251 and 240 for EPI, PSI and FCI respectively. The ratio of the adolescents' to parents' speech duration was 0.73, 0.71 and 0.67 for EPI, PSI and FCI respectively.

Table 5.2: SPEECH CORPUS OF DEPRESSED PARTICIPANTS – MALE ADOLESCENTS

Subject ID	Gender	Diagnosis	No. of Parents	No. of subject utterances		
				EPI	PSI	FCI
1441	M	Depressed	1	251	248	171
2486	M	Depressed	1	254	346	199
4198	M	Depressed	1	284	231	261
4683	M	Depressed	1	235	246	211
6026	M	Depressed	1	287	262	127
1108	M	Depressed	2	331	250	273
1150	M	Depressed	2	331	306	268
1451	M	Depressed	2	360	149	174
1813	M	Depressed	2	391	185	213
2608	M	Depressed	2	237	244	236
2730	M	Depressed	2	190	247	299
3184	M	Depressed	2	330	286	246
3193	M	Depressed	2	203	233	208
3243	M	Depressed	2	231	200	376
3284	M	Depressed	2	218	277	218
4160	M	Depressed	2	278	326	219
4497	M	Depressed	2	291	298	232
4628	M	Depressed	2	270	255	304
6075	M	Depressed	2	336	237	181
			Total	5380	4826	4416

Table 5.3: SPEECH CORPUS OF CONTROL PARTICIPANTS – MALE ADOLESCENTS

Subject ID	Gender	Diagnosis	No. of Parents	No. of subject utterances		
				EPI	PSI	FCI
2795	M	Control	1	183	139	151
3249	M	Control	1	196	189	159
3457	M	Control	1	288	196	265
3486	M	Control	1	293	200	254
3631	M	Control	1	303	230	157
1082	M	Control	2	224	181	220
1083	M	Control	2	184	146	177
1129	M	Control	2	287	215	230
1172	M	Control	2	270	189	205
1427	M	Control	2	248	181	257
1440	M	Control	2	220	178	236
2001	M	Control	2	335	324	251
2071	M	Control	2	411	257	310
2310	M	Control	2	286	288	230
2485	M	Control	2	281	293	178
3369	M	Control	2	329	254	251
3394	M	Control	2	292	284	223
3841	M	Control	2	351	362	283
3978	M	Control	2	185	148	189
4067	M	Control	2	257	287	199
4145	M	Control	2	309	266	193
4263	M	Control	2	210	232	175
4401	M	Control	2	223	274	251
5041	M	Control	2	294	290	268
6070	M	Control	2	343	279	205
6367	M	Control	2	297	348	285
6568	M	Control	2	289	279	311
			Total	7388	6509	6113

Table 5.4: SPEECH CORPUS OF DEPRESSED PARTICIPANTS – FEMALE ADOLESCENTS

Subject ID	Gender	Diagnosis	No. of Parents	No. of subject utterances		
				EPI	PSI	FCI
1949	F	Depressed	1	277	127	132
2143	F	Depressed	1	266	264	282
2439	F	Depressed	1	231	253	184
2462	F	Depressed	1	250	214	242
2523	F	Depressed	1	307	166	259
2739	F	Depressed	1	231	253	274
2829	F	Depressed	1	284	232	253
2900	F	Depressed	1	214	248	191
2939	F	Depressed	1	189	174	146
3098	F	Depressed	1	321	285	255
3146	F	Depressed	1	225	235	202
3177	F	Depressed	1	184	199	199
3359	F	Depressed	1	239	138	150
3564	F	Depressed	1	265	301	280
3712	F	Depressed	1	273	229	212
3960	F	Depressed	1	204	178	165
3992	F	Depressed	1	197	161	204
4057	F	Depressed	1	217	213	233
4462	F	Depressed	1	250	263	227
4490	F	Depressed	1	211	165	199
4583	F	Depressed	1	206	207	184
4734	F	Depressed	1	278	222	261
4947	F	Depressed	1	207	193	126
5085	F	Depressed	1	287	227	234
1160	F	Depressed	2	315	187	249
1845	F	Depressed	2	222	94	128
1969	F	Depressed	2	339	258	271
2274	F	Depressed	2	352	206	187
2378	F	Depressed	2	337	309	159
2389	F	Depressed	2	289	323	300
2449	F	Depressed	2	211	245	206
2482	F	Depressed	2	381	287	241
2519	F	Depressed	2	334	261	291
2659	F	Depressed	2	401	318	311
2924	F	Depressed	2	265	319	282
3244	F	Depressed	2	270	205	242
3313	F	Depressed	2	393	290	274
3355	F	Depressed	2	197	176	197
3826	F	Depressed	2	250	298	186
3835	F	Depressed	2	256	205	27
3984	F	Depressed	2	315	211	224
4193	F	Depressed	2	267	315	344
4194	F	Depressed	2	415	329	290
4470	F	Depressed	2	264	337	303
4901	F	Depressed	2	323	332	337
4920	F	Depressed	2	309	181	185
4949	F	Depressed	2	381	335	319
5083	F	Depressed	2	300	313	320
6375	F	Depressed	2	316	249	269
			Total	13515	11730	11236

Table 5.5: SPEECH CORPUS OF CONTROL PARTICIPANTS – FEMALE ADOLESCENTS

Subject ID	Gender	Diagnosis	No. of Parents	No. of subject utterances		
				EPI	PSI	FCI
1004	F	Control	1	232	288	179
1996	F	Control	1	211	231	144
2713	F	Control	1	328	227	286
3461	F	Control	1	170	186	194
3719	F	Control	1	320	258	225
4002	F	Control	1	250	259	191
4262	F	Control	1	248	270	230
4485	F	Control	1	232	271	252
4702	F	Control	1	235	280	267
4708	F	Control	1	244	266	293
4712	F	Control	1	333	289	261
5205	F	Control	1	297	339	236
1065	F	Control	2	179	195	357
1455	F	Control	2	367	212	287
1810	F	Control	2	320	222	269
1820	F	Control	2	356	290	309
1822	F	Control	2	364	320	221
1922	F	Control	2	295	258	213
1935	F	Control	2	268	152	208
1943	F	Control	2	276	285	296
2036	F	Control	2	369	266	249
2148	F	Control	2	294	178	251
2207	F	Control	2	367	225	313
2279	F	Control	2	393	241	314
2331	F	Control	2	285	253	266
2638	F	Control	2	285	283	228
2673	F	Control	2	203	262	266
2718	F	Control	2	370	264	258
2850	F	Control	2	340	301	291
2853	F	Control	2	80	320	298
3118	F	Control	2	404	344	353
3251	F	Control	2	255	195	255
3273	F	Control	2	234	215	226
3510	F	Control	2	195	381	248
3945	F	Control	2	282	330	363
4020	F	Control	2	353	370	279
4062	F	Control	2	327	351	272
4148	F	Control	2	197	180	282
4378	F	Control	2	249	313	301
4396	F	Control	2	275	351	290
4452	F	Control	2	267	254	240
4503	F	Control	2	342	283	259
4889	F	Control	2	246	356	277
6273	F	Control	2	348	214	236
			Total	12485	11828	11533

Chapter Six:

SPEECH ANALYSIS METHODOLOGY

"An idea that is not dangerous is unworthy of being called an idea at all."

– Oscar Wilde

6.1 Preview

The analysis methodology described in this chapter in modeling speech contents of depressed and control adolescents was based on the proposed framework illustrated by the general block in Fig. 6.1. Firstly, from the speech corpus that was discussed in *Chapter 5*, detection of voiced frames was implemented in the pre-processing stage for both the training and testing phases and is explained in *Section 6.2*. Secondly, from these voiced frames as will be discussed in *Section 6.3*, acoustic features that represented objective measurements in the human speech production were extracted. Thirdly, *Section 6.4* deals with the statistical analyses that were carried out to discard any acoustic features that were statistically non-significant in distinguishing the speech of depressed adolescents from that of control adolescents. Finally, machine learning techniques were introduced in *Section 6.5* whereby the selected extracted acoustic features that were statistically significant in the training phase were modeled into their respective classes (depressed and control class) and the accuracies in classification were determined from the testing set of statistically significant features. The dotted line in Fig. 6.1 indicates using 50% training/testing set.

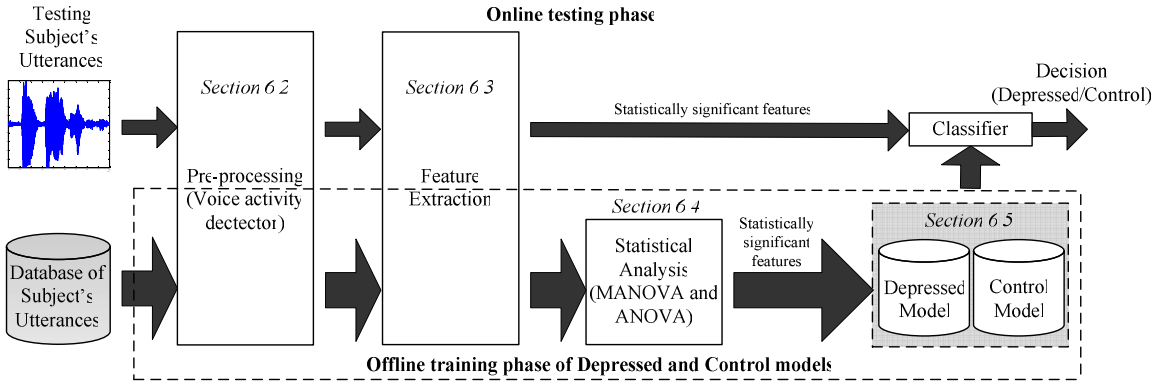

Figure 6.1: Block diagram in modeling speech of depressed and control adolescents.

6.2 Pre-processing – Voice Activity Detector

We used a frame based linear prediction (LP) technique extracted from the toolbox in [18] to detect the voiced regions of the speech signal.. First, the speech signal was normalized based on the maximum amplitude and then was segmented into 25 msec with 50% overlapping frames using a rectangular window. If the final frame segmented was less than 25 msec, the frame was appended with random noise of 30db below the peak amplitude to fill it out to 25 msec. Next, the segmented frames of the speech signal were filtered with a zero phase filter to remove any low-frequency drift. The 13^{th} order linear prediction coefficients (LPCs) were then calculated per frame. Energy of the prediction error and the first reflection coefficient r_1 were calculated and a threshold was empirically set to detect the voiced frames.

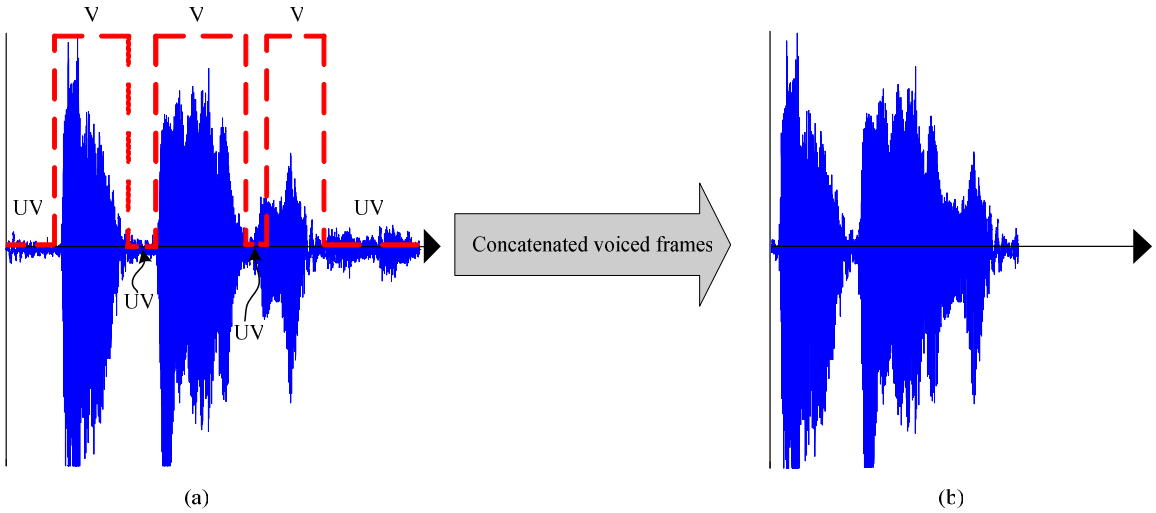

Figure 6.2: Illustration of the voice activity detector. (a) Voiced (V) segments and unvoiced (UV) segments detected. (b) Concatenation of only the voiced frames.

Eq. (6.1) explains the calculation of the first reflection coefficient r_1, where N is the number of samples in the analysis frame and $s(n)$ is the speech sample. A threshold of 0.2 determined empirically by [51] was set for the reflection coefficients to classify a frame as voiced (reflection coefficient > 0.2) or unvoiced (reflection coefficient < 0.2).

$$r_1 = \frac{\frac{1}{N}\sum_{n=1}^{N-1} s(n)s(n+1)}{\frac{1}{N}\sum_{n=1}^{N} s(n)s(n)} \tag{6.1}$$

In this algorithm, silent frames were also considered as unvoiced. All unvoiced frames were next removed in the pre-processing stage with the detected voiced frames concatenated for feature extraction as illustrated in Fig. 6.2.

6.3 Feature Extraction

Similar to the procedure in [84], we also proposed the grouping of acoustic features into categories and subcategories that are closely related to the human speech production model. As was explained in *Section 2.6* and illustrated in Fig. 2.8, our investigations lies in the regions of interest (ROI) of acoustic features that are associated with the human speech production mechanisms. The acoustic features in our study were grouped into five main feature categories that represented the TEO-based, cepstral (C), prosodic (P), spectral (S), and glottal (G) features. Acoustic features grouped into these categories are closely related to the physiological and perceptual components that characterized speech in the human speech production model. The physiological components are related to the feature categories of TEO-based, prosodic (P), spectral (S), and glottal (G). The TEO-based feature category is derived from the nonlinear speech production model and measures the nonlinear airflow in the vocal tract, whereas, the feature categories of prosodic (P), spectral (S), and glottal (G) are derived from the linear speech production model of sound propagation along the vocal tract. The feature category of cepstral (C) which is also derived from the linear speech production model, relates to the perceptual aspect. We will now discuss the acoustic features implemented in each category.

6.3.1 Prosodic Category

6.3.1.1 Fundamental frequency

The autocorrelation method for fundamental frequency (F0) estimation was used to determine changes in speaking behavior in response to factors relating to stress, intonation and emotional changes. Prior to the experiment, a small-scale test was

conducted in which the autocorrelation method was compared with the cepstrum method and the average magnitude difference function (AMDF) method using the Roger Jang's audio toolbox [54] . It was observed that all the three methods were giving consistently similar results. Since the speech recording was setup in a controlled environment in a laboratory using high quality sound recording equipment and lapel microphones attached to the chest of each individual speaker; the noise level was relatively low making it more likely for the autocorrelation method to provide stable results. In our experiments, the fundamental frequency F0 was calculated with the autocorrelation method. For the pitch tracking algorithm, the autocorrelation function (ACF) calculated the pitch periods by searching for the maximum point of the ACF within 40 Hz to 1000 Hz range in each frame of the speech signal. The pitch estimation algorithm employed is a time domain approach which estimates the similarity between a frame and its delayed version via the autocorrelation function (ACF):

$$acf(lag) = \sum_{n=0}^{N-1-lag} s(n)s(n+lag) \tag{6.2}$$

The value of lag in terms of sample points that maximizes *acf (lag)* over a specified range was selected as the pitch period. This operation in the autocorrelation for pitch estimation is illustrated in Fig. 6.3. Fig. 6.3 (a) shows an example of an input speech frame being measured for similarities with its shifted delayed version of the frame via the ACF as depicted in Fig. 6.3 (b). Fig. 6.3 (c) shows the plot of the ACF. The blue line shows the original ACF plot and the green line is the truncated plot with their frequency limits

imposed within the 40 Hz to 1000 Hz range. Searching for the maximum point in the ACF plot within these frequency limits is chosen as the pitch period. It is expected for a periodic signal to correlate well with itself at very short delays and at delays corresponding to multiples of the pitch periods.

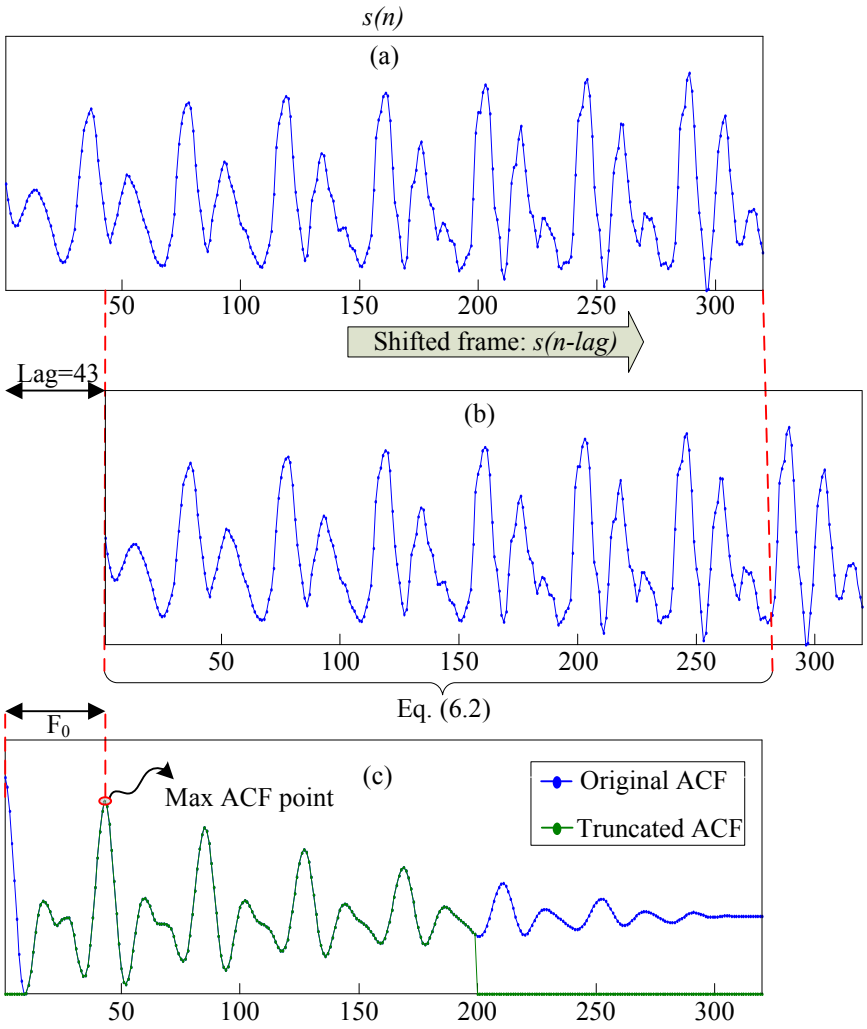

Figure 6.3: Pitch estimation using the autocorrelation function: (a) speech frame; (b) shifted speech frame; (c) Plot of the autocorrelation function.

6.3.1.2 Log energy

The logarithmic of short-term energy (LogE) within a frame, E_s (m) is also an important feature which measures the loudness of the speech signal and is given by:

$$E_s(m) = \log \sum_{n=m-N+1}^{m} s^2(n) \tag{6.2}$$

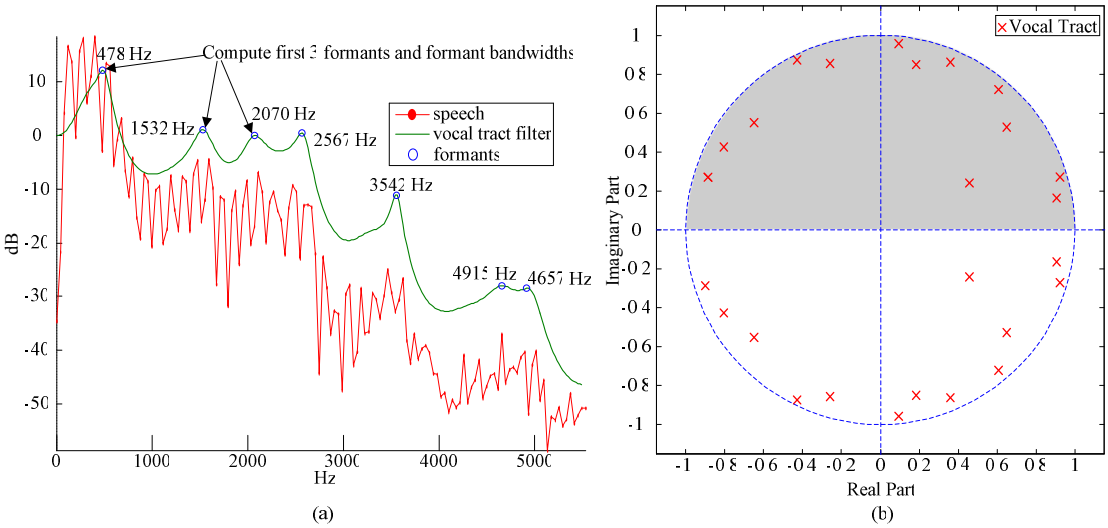

Figure 6.4: Formants estimation. (a) Vocal tract spectra. (b) Pole-zero plot of vocal tract spectrum.

6.3.1.3 Formants & formant bandwidths

The shape and length of the vocal tract is characterized by a set of resonance frequencies also known as formants that play a vital role in providing important information relating to the physical characteristics of a speaker. Formants carry the identity of sound. The location of the formants and its bandwidths of voiced speech are important in many applications [36]. The tendency in modeling the peaks of the spectral envelope using linear prediction makes it a suitable choice in finding the formants of the vocal tract. The

vocal tract is modeled as an all-pole filter. A 13th order linear prediction (LP) filter was employed to calculate the formant frequencies and bandwidths from the roots of the polynomial LP predictor that have their imaginary parts positive (i.e. lying in the upper half of the unit). Referring back to *Chapter 2*, Eq. (2.6) and Eq. (2.7) gives the transformation of complex pole (p_i) root pairs $z = p_i e^{\pm \theta i}$ and sampling frequency F_s into the location of its formant frequency (FMT) and 3-dB bandwidth (FBW).

Fig 6.4a shows an illustration of an input speech frame together with its spectra shape of the vocal tract filter. The resonances (formants) of the vocal tract are highlighted by circles (O) located at each peak of the spectrum. The corresponding pole and zero plot of the vocal tract spectrum are presented in Fig 6.4b. The poles are highlighted by X in the pole and zero plot in modeling the vocal tract using a 13th order LP filter. As shown in Fig. 6.4b, complex conjugate poles always come in pairs, both having the same real part, but with imaginary parts of equal magnitude and opposite signs. The upper half of the circle corresponds to the frequency range from 0 Hz to half the sampling frequency Fs. Therefore, as previously said, only the poles with their imaginary parts positive that are lying in the upper half of the unit circle which is shaded in grey in Fig. 6.4b, were used in the calculation of the formants and bandwidths. For our experiments, only values of the first three formants (FMT_1-FMT_3) and formant bandwidths (FBW_1-FBW_3) below its Nyquist frequency were taken.

6.3.1.4 Jitter

Frequency perturbation also called jitter refers to the short-term (cycle to cycle) fluctuations in fundamental frequency (F0). It is obtained by measuring the F0 of the

glottal cycle, subtracting it from the previous F0 values, and dividing it by the average value of F0.

$$Jiiter = \frac{\frac{1}{N-1}\sum_{i=1}^{N=1} | F_{0i} - F_{0i+1} |}{\frac{1}{N}\sum_{i=1}^{N} F_{0i}} \tag{6.3}$$

where *i* is the frame number and *N* is the total number of frames.

6.3.1.5 Shimmer

Shimmer is calculated in a similar fashion as the jitter calculation previously described; however, the period to period variability of the signal peak to peak amplitude (A_i) is calculated instead.

$$Shimmer = \frac{\frac{1}{N}\sum_{i=1}^{N-1} | A_i - A_{i+1} |}{\frac{1}{N}\sum_{i=1}^{N} A_i} \tag{6.4}$$

where *i* is the frame number and *N* is the total number of frames.

In clinical treatment, jitter and shimmer have been widely used to describe the pathological characteristics of voice [85].

6.3.2 Cepstral Category

6.3.2.1 Mel frequency cepstral coefficients (MFCC)

The mel frequency cepstral coefficients (MFCC) have been widely used in speech processing and were considered because it has been effectively used in speech content characterization [105]. A mel is a unit of measure of perceived pitch. It uses filter-banks

based on the human auditory system that transforms linear frequencies to a logarithmic scale (mel-scale) that is more appropriate in approximating the perceived pitch and harmonic content of speech signals, since the human ear perceives sound in a nonlinear manner. This is useful since our whole understanding of speech is through our ears. The relationship between linear frequency scale (F_{linear}) and the mel-scale (F_{mel}) is illustrated in Fig. 6.5 and explained in Eq. (6.5). The positions of the linear frequencies are mapped to the mel-scale given by:

$$F_{lmel} = \frac{C \log_{10}}{\log_{10} 2}[1 + \frac{F_{linear}}{C}] \tag{6.5}$$

where C is the corner frequency which controls the slope of the curve. Fig. 6.5a shows the mapping of the linear frequency to the mel-scale with different values of C (i.e., 100, 300, 500 and 700).

From the spectral regions of the voiced frames, nonlinear filter positions in the linear frequency scale (f_{linear}) were calculated by first placing linearly spaced triangular filters in the mel frequency scale (F_{mel}) as illustrated beside the y-axis in Fig. 6.5a. Next, these triangular spaced filters in the mel-scale are mapped back to the linear frequency scale using Eq. (6.5). This is shown in Fig. 6.5b where the mapping of the filters is approximately linear below 1 kHz and logarithmic above. Using this filter bank we compute the MFCCs. The output Y(i) of the i^{th} filter in the linear frequency scale is defined in Eq. (6.6), where S(.) is the signal spectrum, $H_i(.)$ is the i^{th} filter, and m_i and n_i are boundaries of the i^{th} filter.

$$Y(i) = \sum_{j=m_i}^{n_i} \log_{10}[S(j)]H_i(j) \tag{6.6}$$

Eq. (6.7) describes the computation of n[th] MFCC, k_i is the center frequency of the i^{th} filter, and N and N_{cb} are number of frequency sample points and number of filters, respectively.

$$C(n) = \frac{2}{N}\sum_{i=1}^{N_{cb}} Y(i)\cos(k_i \frac{2\pi}{N} n) \tag{6.7}$$

For our experiments, the corner frequency of C =700 was implemented because it has been documented that it provides closer approximation to the mel-scale for frequencies below 1000 Hz [40].

6.3.3 Spectral Category

6.3.3.1 Spectral centroid

Spectral centroid (SC) indicates the centre of a signal's spectrum power distribution. It is the calculated weighted mean of frequencies present in the signal, with their magnitudes as weights.

$$SC = \frac{\sum_{n=0}^{N-1} f(n)S(n)}{\sum_{n=0}^{N-1} S(n)} \tag{6.8}$$

where *S(n)* represents the magnitude of frequency bin number *n* of the speech power spectrum, *f(n)* represents the center frequency bin and *N* is the total number of frequency bins.

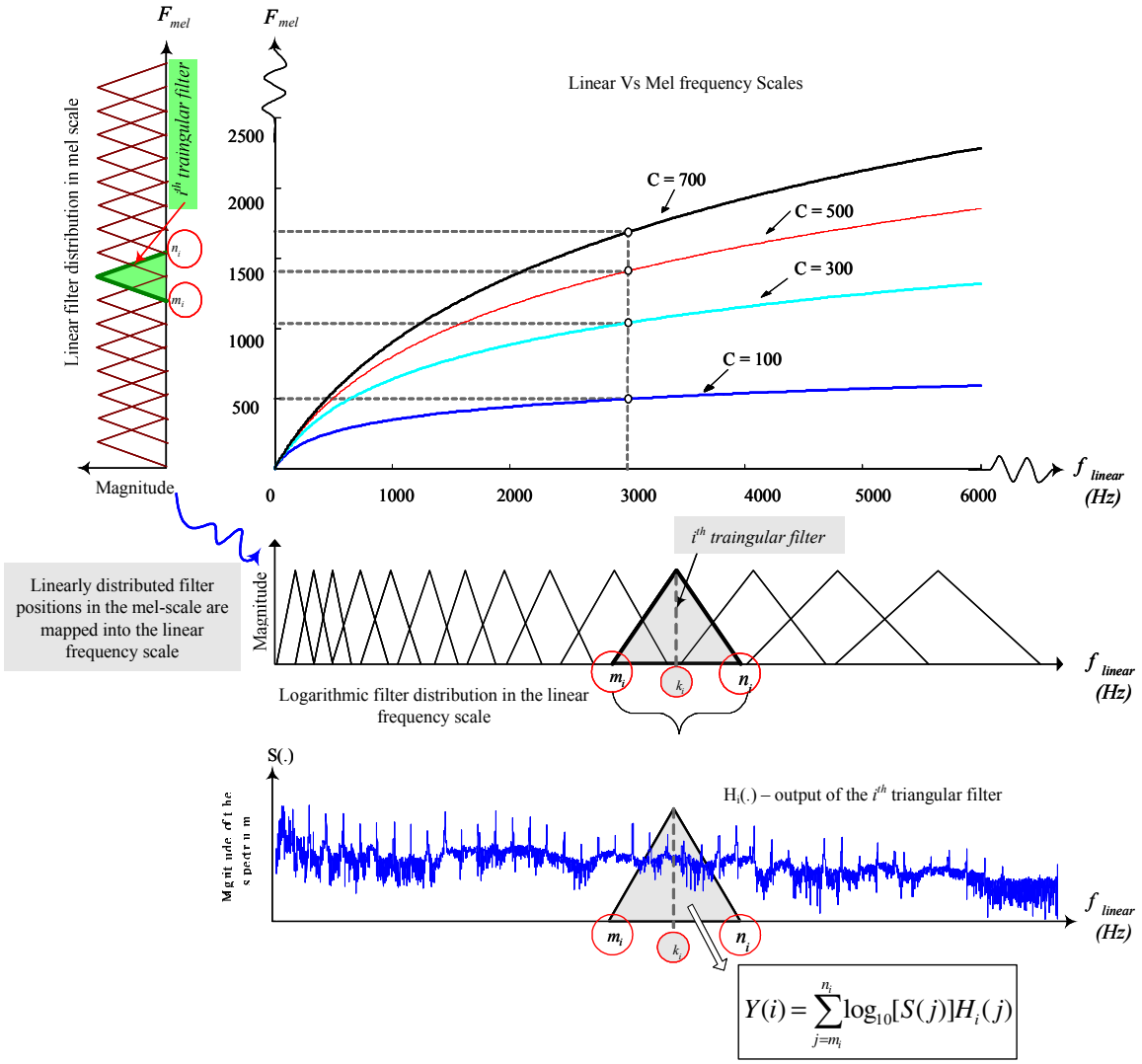

Figure 6.5: Plots of filter distributions in mel-scale versus linear frequency scale (illustration reproduced after Maddage (2006)).

6.3.3.2 Spectral flux

Spectral flux (SF) is a measure of how quickly the power spectrum of a signal is changing, calculated by comparing the power spectrum for one frame against the previous frame. It is the measure of the amount of frame to frame variance in the spectral shape. A steady sound will exhibit little flux, while a modulating sound will exhibit more flux. The calculation for SF measures the ordinary Euclidean norm of the delta spectrum power between two adjacent frames and is given by:

$$SF = \left\| \; |S(k)| - |S(k+1)| \; \right\| \tag{6.9}$$

where k is the sample index corresponding to a frequency and S(k) is the power of the speech signal at the corresponding frequency band.

6.3.3.3 Spectral entropy

Spectral entropy (SE) is the means of measuring the amount of information based on Shannon's information theory and it has been applied to emotion recognition in speech [69]. The probability density of the spectrum in each frame is estimated by normalizing the frequency components and is given as:

$$P[S(k)] = \frac{S(k)}{\sum_{k=1}^{M/2} S(k)} \tag{6.10}$$

where M is the FFT sample length.

The spectral entropy is then defined as:

$$H = -\sum_{k=1}^{M/2} P[S(k)] \log_2 P[S(k)] \tag{6.11}$$

6.3.3.4 Spectral roll-off

Spectral roll-off (SR) is the point where the frequency that is below some percentage (set as 80% for our experiments) of the power spectrum resides. The equation for SR is:

$$\sum_{n=0}^{K-1} S(n) = 0.80 \sum_{n=0}^{N-1} S(n) \tag{6.12}$$

where n is the frequency bin index, N is the total number of frequency bins, $S(n)$ is the amplitude of the corresponding frequency bin, and K is the spectral roll-off number.

6.3.3.5 Power spectral density

The one sided power spectral density was computed based on the Welch spectral estimator method using a 4096-point fast Fourier transform (FFT) with a 5ms non-overlapping hamming window size. The total power spectral for frequency 0-2000 HZ, its sub-bands PSD_1 (0-500Hz), PSD_2 (500-1000Hz), PSD_3 (1000-1500Hz), PSD_4 (1500-2000Hz) and the ratio of power from each spectral sub-band to the total power were calculated. The PSD have been effectively used to discriminate between speech of control and depressed adults [38].

6.3.4 Glottal Category

The glottal pulse and shape have documented to play an important role in the analysis of speech in clinical depression [84], [91]. In a recent study [84], the influence of glottal features combined with prosodic and vocal tract features have been documented to improve overall discrimination of speech from the depressed and control group. The process of acquiring the glottal flow characteristics is by inversing the estimated vocal tract and lip radiation filters from the source of the speech signal. However, separating the glottal flow from the vocal tract by locating the timing instances where the vocal folds are assumed closed is a difficult task to achieve. This is due, in part, to the different speaking styles that vary from speaker to speaker. For example, it is more difficult to capture the glottal closure in speaking styles that yield a higher pitch when modeling the vocal tract because of the rapid motion of the vocal folds. The other problems relate to analysis methods (such as LP analysis) which make assumptions that do not hold when in separating glottal/vocal tract interaction. For our study in glottal flow extraction, we used the TTK Aparat glottal inverse filtering toolbox [2]. The glottal inverse filtering method implemented here was based on a slightly modified iterative adaptive inverse filtering algorithm (IAIF) [3]. Instead of using a linear predictive (LP) filter, a discrete all-pole modeling (DAP) was implemented to model the vocal tract as it is less sensitive to the biasing of formants caused by nearby harmonic peaks. Like for the LPC as previously described for the formant calculation, the number of formants (or resonances) to model the vocal tract in the DAP was set to 13 (F_s/1000+2). This was done to be sure that there was at least one formant for every kilohertz band in the vocal tract transfer function. Once the glottal flow was estimated, quantitative analysis of the glottal flow pulses was

performed in the time-domain and in the frequency-domain. It should be noted that glottal waveform extraction is still a matter of study and accurate representations are still difficult to determine and verify. In this study, the glottal-timing (GLT) and the glottal frequency (GLF) were used to represent the glottal flow parameters in the time and in the frequency domain respectively.

6.3.4.1 Glottal timing

The glottal flow can be divided into a few phases that are illustrated from the glottal flow pulse in Fig. 6.6b. This is shown by the mark boundaries from the dotted lines indicating the glottal timing interval for the opening phase (OP), closing phase (CP) and closed phased (C) that describes the glottal pulse shape. It has been suggested that the glottal OP can be subdivided into two timing instances referred to as the primary opening (▲) and secondary opening (●) [94]. The duration of the primary and secondary opening of OP is denoted by T_{o1} and T_{o2} respectively. The duration of CP is denoted by T_c and the period of the glottal cycle is denoted by T. Once these instances are acquired, several timing and frequency parameters can be easily calculated. In GLT, the timing parameters used is the open quotients (OQ_1 & OQ_2), approximation of the open quotient (OQ_a), quasi-open quotient (QOQ), speed quotients (SQ_1 & SQ_2), closing quotient (CIQ), amplitude quotient (AQ) and normalized amplitude quotient (NAQ).

6.3.4.2 Glottal frequency

For the GLF, the frequency parameters used are the difference of the first and second harmonics (labeled H_1 and H_2 in Fig. 6.6d) in decibels of the glottal flow power spectrum

(DH12), harmonic richness factor (HRF) and the parabolic spectral parameter (PSP). Table 6.1 shows a brief summary of their parameters and an in-depth description of them can be found in [2].

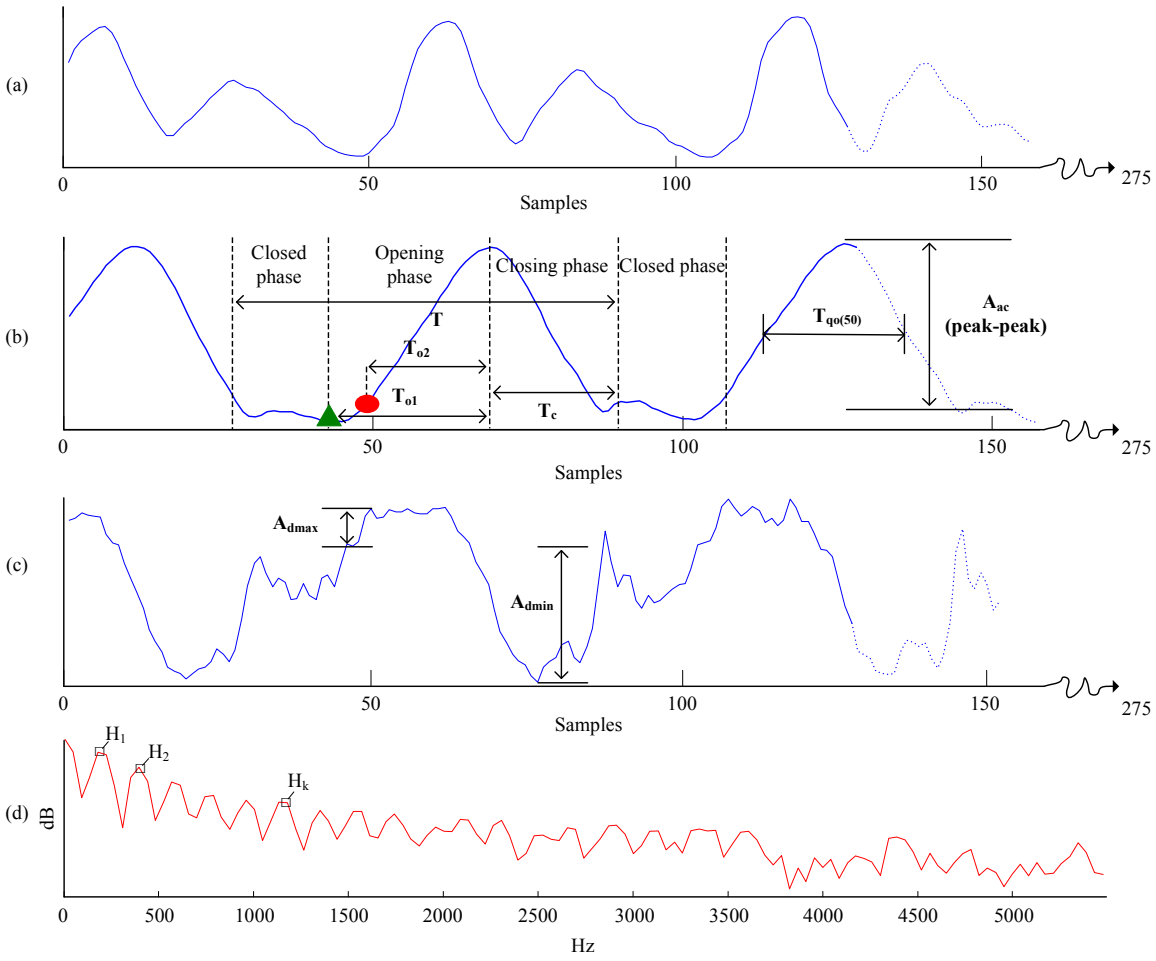

Figure 6.6: Glottal inverse filtering. (a) 25msec speech frame. (b) Glottal flow estimate. (c) Glottal flow derivative. (d) Glottal flow spectrum.

Table 6.1: GLOTTAL FEATURES CALCULATIONS – TIMING PARAMETERS (GLT) & FREQUENCY PARAMETERS (GLF)

	Feature parameter	Symbol	Description	Calculation method
GLT	1	OQ_1	The ratio of the primary opening phase to the length of glottal cycle duration.	$\frac{T_{o1}+T_c}{T}$
	2	OQ_2	The ratio of the seconday opening phase to the length of glottal cycle duration.	$\frac{T_{o2}+T_c}{T}$
	3	OQ_a	Approximates the opening to OQ for an ideal LF pulse.	$A_{ac}(\frac{\pi}{2A_{d\max}}+\frac{1}{A_{d\min}})f_o$
	4	QOQ	The time of the open phase duration that is 50% above the peak to peak amplitude of the glottal flow.	$\frac{T_{qo(50)}}{T}$
	5	SQ_1	The ratio of timing duration of the primary opening phase to the closing phase.	$\frac{T_{o1}}{T_c}$
	6	SQ_2	The ratio of timing duration of the secondary opening phase to the closing phase.	$\frac{T_{o2}}{T_c}$
	7	AQ	The ratio of timing duration of the closing phase to the length of the glottal cycle.	$\frac{T_c}{T}$
	8	CIQ	The ratio of the peak-to-peak amplitude of the glottal flow to the minimum peak of the pulse derivative.	$\frac{A_{ac}}{A_{d\min}}$
	9	NAQ	Normalized AQ by dividing it by the length of the glottal cycle duration.	$\frac{AQ}{T}$
GLF	10	PSP	Fits a second-order polynomial to the glottal flow spectrum on a logarithmic scale computed over a single glottal cycle.	Refer to [4]
	11	DH12	Difference of the first and second harmonics in decibels.	$H1-H2$
	12	HRF	The ratio of the sum of harmonics magnitude above the first harmonic (H1) to the magnitude of the first harmonic.	$\frac{\sum_{k>2}H_k}{H_1}$

6.3.5 Teager Energy Operator (TEO)-Based Category

TEO based features have shown good performances in stress recognition [123]. In the emotional state of anger or stressed speech, the fast air flow causes vortices located near the false vocal folds which provides additional excitation signals other than pitch [111], [116]. To model this time-varying vortex flow, Teager [113] proposed a non-linear energy-operator called the Teager energy operator (TEO) which computes an energy

profile (also known as the TEO profile). The Teager energy operator in a discrete form [55] is defined in Eq. (6.13), where Ψ[.] is the Teager energy operator and x(n) is the n^{th} speech sample point.

$$\psi[x(n)] = x^2(n) - x(n+1)x(n-1) \tag{6.13}$$

6.3.5.1 TEO critical-band autocorrelation envelope

Several TEO based features have been proposed in the literature and we computed the TEO critical band based autocorrelation envelope (TEO-CB-Auto-Env) feature that is based on the method discussed in [123]. As proposed by Zhou (2001), it is more convenient to break the bandwidth of the speech spectrum into smaller bands (also known as critical bands) before calculating the TEO profile in Eq. (6.13) for each independent band. This slight modification of the algorithm is carried out to capture the variations of pitch harmonics across the different critical bands. Gabor band-pass filters are implemented to separate voiced utterances into 15 critical bands. Each discrete-time Gabor band-pass filter impulse response is given by:

$$\begin{gathered} h(n) = \exp(-b^2 n^2)\cos(\Omega_c n) \\ \text{with } -N \le n \le N,\ b = \alpha T_s, \text{ and } \Omega_c = 2\pi f_c T_s \end{gathered} \tag{6.14}$$

where T_s is the sampling period, f_c is the centre frequency of the band-pass filter. The integer N is chosen to truncate the Gaussian envelope of $h(n)$ essentially to zero.

The value of α is chosen to control the bandwidth [39] using the rule that:

$$\alpha = rmsBW \times \sqrt{2\pi} \tag{6.15}$$

where the *rmsBW* is the bandwidth of the band-pass filter. Given the centre frequencies and bandwidths of each Gabor band-pass filter, the Gabor band-pass filter is performed by convolving the truncated *h(n)* with the speech signal.

In our implementation, 512-point Gabor band-pass filters for the 15 critical bands were used. The TEO profile was calculated for each of the bands. The Gabor-filtered TEO stream was then segmented into frames. Finally, the autocorrelation of the TEO output was computed and the area under the normalized autocorrelation envelope (depicted by the shaded area in the last plot of Fig 6.7b) was calculated to give the TEO-CB-Auto-Env features. Fig. 6.7a shows a basic block diagram in the computation of the TEO-CB-Auto-Env feature coefficients. We followed approximately the same frequency range for our 15 critical bands as in [123] (CB1: 100~300, CB2: 300~500, CB3: 500~700, CB4: 700~900, CB5: 900~1100, CB6: 1100~1300, CB7: 1300~1500, CB8: 1500~1900, CB9: 1900~2300, CB10: 2300~2700, CB11: 2700~3100, CB12: 3100~3500, CB13: 3500~4100, CB14: 4100~4700, and CB15: 4700~5500) Hz. Fig. 6.7b shows an example of the plots corresponding to the computation described in Fig 6.7a of the Gabor filter, TEO profile waveform and autocorrelation envelope for critical band 9 (1900Hz~2300Hz).

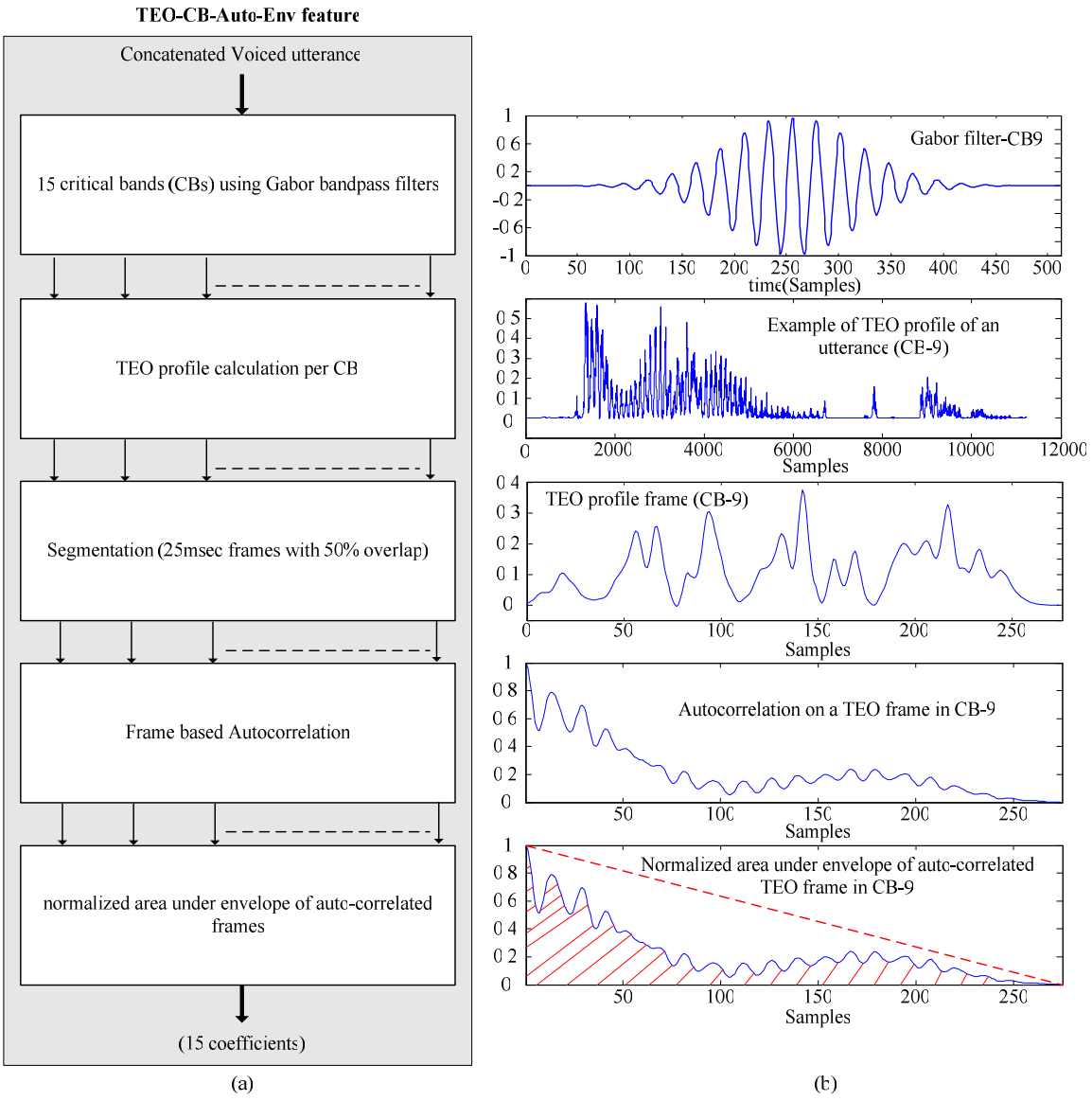

Figure 6.7: TEO-CB-Auto-Env feature: (a) feature extraction implementation. (b) An example of the Gabor filter, TEO profile and the autocorrelation envelope for an utterance within the 9th critical band (CB).

6.3.6 Delta (Δ) and Delta-Delta (Δ-Δ) Coefficients

The inclusion of the first and second order derivatives (delta and delta-delta) which can capture the temporal information among neighboring frames are calculated as follow:

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta (c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{6.16}$$

where d_t is a delta coefficient at time t and it is computed in terms of the corresponding static coefficients from $c_{t-\Theta}$ to $c_{t+\Theta}$. The window size is set $\Theta=9$ to obtain both delta and delta-delta coefficients. The same formula (6.16) is applied to the delta coefficients to obtain the delta-delta coefficients. The delta and delta-delta were incorporated in all the acoustic features that were mentioned in *Section 6.3.1-6.3.5*.

6.4 Statistical Setup

The acoustic feature categories previously described in *Section 6.3* were statistically examined for significance in characterizing speech of depressed and control adolescents. This was done as a preliminary step to ensure that feature coefficients that gave a statistically non-significant result were removed in the modeling of depressed and control speech. Assumptions of parametric testing using Kolmogorov-Smirnov (KS) test were also examined for each feature coefficient to check if they were normally distributed within each of the depressed and control classes. Error bars were also graphed to visually check for differences in their mean values between each dependent variable. In order to identify relationships that might exist between the feature coefficients, a multivariate

analysis of variance (MANOVA) followed by a one-way analysis of variance (ANOVA) procedure was conducted on pair-wise comparison of the depressed and control classes.

6.5 Modeling and Classification Setup

6.5.1 Gaussian Mixture Model

GMM has been effectively used in speech information modeling tasks such as speaker recognition and spoken language identification [12]. The HTK package (Young *et. al 2002*) was used for the GMM-based depressed and control content modeling. The HTK-toolbox was originally designed for the training of the Hidden Markov Models (HMM). However, it can also be used for the training of the GMM, as a one state self-loop continuous density HMM is equivalent to a GMM. As was described in *Section 3.2*, an integral part of the GMM is to optimize the parameters of each Gaussian mixture component by maximizing the likelihood function using the expectation maximization (EM) framework which is an iterative optimization technique. The EM algorithm implemented in the HTK toolbox to train the models is known as the Baum-Welch (also called forward-backward) algorithm and differs from the other standard EM algorithms.

The EM algorithm was used for estimating parameters of mean, covariance and mixture weight of each Gaussian component in the GMM. Each model was trained with three iterations every time the number of Gaussian mixture components was increased. For computational efficiency, diagonal covariance matrices were used in the Gaussian component instead of full covariance.

6.5.2 Optimized Parallel Support Vector Machine (OPSVM)

Although, the support vector machine (SVMs) systems have shown to yield generally good results when applied to many classification problems [12], the computational complexity of these algorithms is very high, which prohibits applications involving large scale problems. Given a two-class problem and a training set of N vector-class pairs: $\{[\mathbf{x}_1, y(\mathbf{x}_1)], \ldots, [\mathbf{x}_N, y(\mathbf{x}_N)]\}$, where $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d$ are the D-dimensional feature vectors, and $y(\mathbf{x}_i) \in \{-1, +1\}$ are the actual class labels for vectors $\mathbf{x}_i$, the classification labels for vectors $\mathbf{x} \in \mathbb{R}^d$ are produced by the decision function $s(\boldsymbol{x})$ defined as:

$$s(\mathbf{x}) = sign\left(\sum_{i=1}^{N} y(\mathbf{x}_i)\alpha_i K(\mathbf{x}, \mathbf{x}_i) + b\right) \tag{6.17}$$

where $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_N\}$ and b are the SVM model parameters, and $K(\mathbf{x}, \mathbf{x}_i)$ is a positive definite kernel chosen to be the Gaussian type of the *Radial Basis Function* (RBF) as explained in *Section 3.3*.

Due to the large size of the data set, the training process based on a single SVM was found to be extremely slow and therefore not useful in practical applications. To train an SVM, solving a quadratic optimization problem is needed. Computing $K(\mathbf{x}, \mathbf{x}_i)$ for every training pair would require $O(N^2)$ computation, and solving may take up to $O(N^3)$ [22], where the number of parameter is N. Therefore, an optimized parallel implementation (also known as mixture of experts) of SVM had to be used to reduce the computational effort associated with the SVM training when applied to a very large data set. Following an approach described in [22], a single SVM was replaced by a

parallel configuration of M support vector machines (SVMs), with each SVM being trained on a different subset of the entire training data to determine an optimal set of coefficients $\alpha_k = \{\alpha_1^k, ..., \alpha_N^k\}$, k=(1,...,M) for each SVM. The training data sub-sets were mutually exclusive, and randomly selected. The LIBSVM toolbox [17] was used to train each SVM. The LIBSVM implementation was carried out by interfacing the binary executable files given in the LIBSVM toolbox with MATLAB.

The OPSVM training was conducted as a two-stage process. In the first stage of the training of the OPSVM as illustrated in Fig. 6.8, optimal model parameters for each SVM were determined and trained using the following procedures:

1. Divided the training set into random subset of size N/M where N is the total number of training examples and M is the number of SVMs to be trained for each subset.
2. Scaled each subset to be [0, 1].
3. Empirically modeled each training data subset using a SVM with radial based function (RBF) in the kernel. Searching for the most appropriate (γ, C) pair of the RBF was performed through a grid search using 5-fold cross-validation on the training dataset.
4. Determined the SVM model parameters of α and b by minimizing the quadratic optimization problem found in Eq. (3.11).

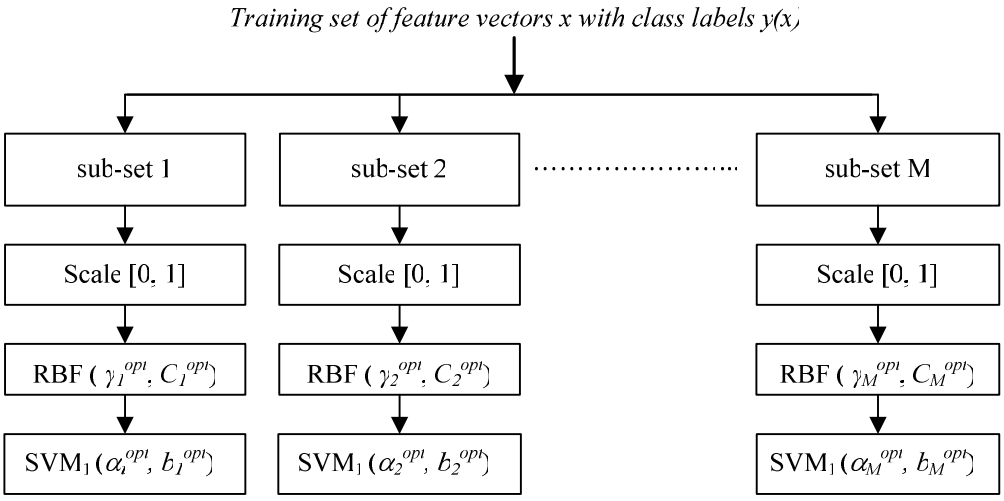

Figure 6.8: Stage I of the OPSVM training process - finding the optimal model parameters and training each SVM model.

Assuming that $s_k(x)$ is the output from the k^{th} SVM, the output from the combined configuration of all SVMs was calculated as a hyperbolic tangent of a weighted sum of outputs from individual SVMs:

$$r(\mathbf{x}) = \tanh\left(\sum_{k=1}^{M} w_k s_k(\mathbf{x})\right) \tag{6.18}$$

where w_k is a constant weight for k^{th} SVM. Instead of the neural network approach proposed in [22], a global optimization algorithm was designed to determine the weights vector $w=\{w_1,...,w_M\}$. The algorithm minimized the following objective function:

$$g_{obj}(\mathbf{w}) = \sum_{i=1}^{N} \left[r(\mathbf{x}_i) - y(\mathbf{x}_i)\right] \tag{6.19}$$

In the second stage of the OPSVM as illustrated in Fig. 6.9, the optimal weight vector **w**, were derived using a simple global optimization procedure [80] referred to as the *Range Controlled Evolutionary Simulated Annealing* (RCESA).

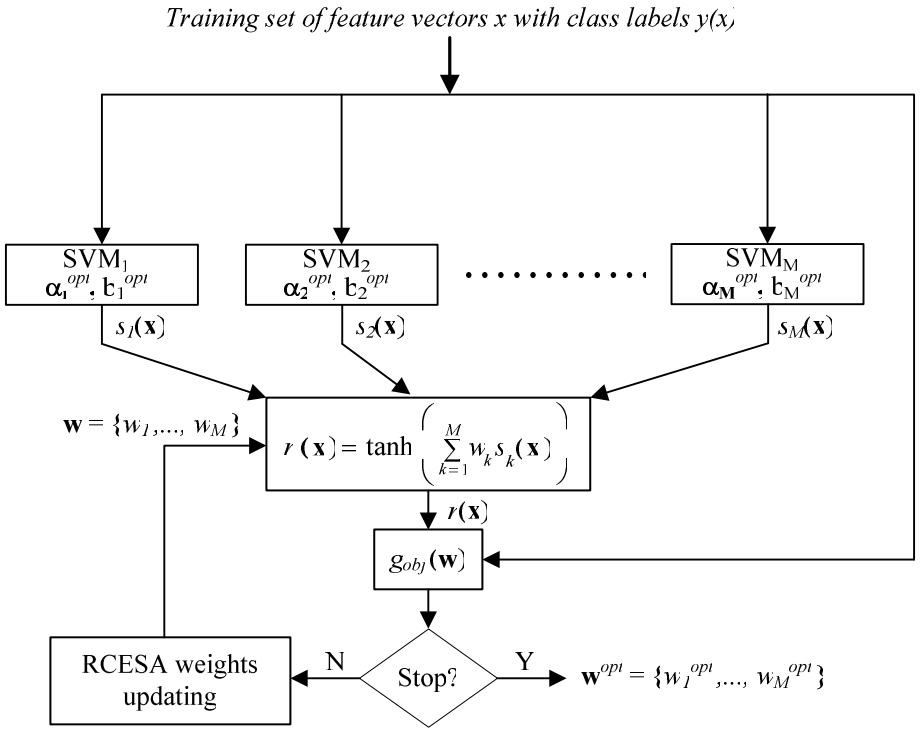

Figure 6.9: Stage II of the OPSVM training process - finding the optimal weight vector **w**.

The RCESA optimization procedure used to derive the optimal set of weights $\{w_{1,...,} w_M\}$ combined two meta-heuristic search strategies: *Evolutionary Search* and *Simulated Annealing* and introduced a range control in the solution space. It is referred to as the *Range Controlled Evolutionary Simulated Annealing* (RCESA). The RCESA algorithm was designed to increase the search efficiency as well as to improve the search sensitivity to the minima. In addition to the traditional Boltzmann acceptance criterion [93] based on the change in the objective function value (energy change), the RCESA restricts the search step size in the vector space to a limited range decreasing exponentially with iterations. The combined control of the energy and range aims to improve the algorithms' efficiency in detecting local "*downhill movements*", which is

important when the desired solution is located at the bottom of a "*long narrow valley*" [93]. Also, the algorithm starts a new iteration from the "*best so far*" solution, thus improving convergence rates.

At the beginning, the algorithm scans a wide range of the feature space and accepts solutions leading to decrease or to a small increase of the objective function value. As the algorithm progresses, the probability of accepting solutions leading to higher values of the objective function decreases rapidly to zero. At the same time the search range becomes increasingly confided to a small space around the final solution. As illustrated in the flowchart of Fig. 6.10, the evolutionary aspect was incorporated into the algorithm by generating not a single solution but a whole population of solutions, and choosing the best one. Each iteration *iter* starts from the "*best so far*" solution with a new value of temperature T_{iter}, new vector range R_{iter}. The temperature and the vector range decreases exponentially with iterations, whereas the population size is kept constant throughout iterations.

After finding the optimal set of weights, the classification process was conducted for the test vectors using Eq. (6.17) and Eq. (6.18) as illustrated in Fig. 6.11.

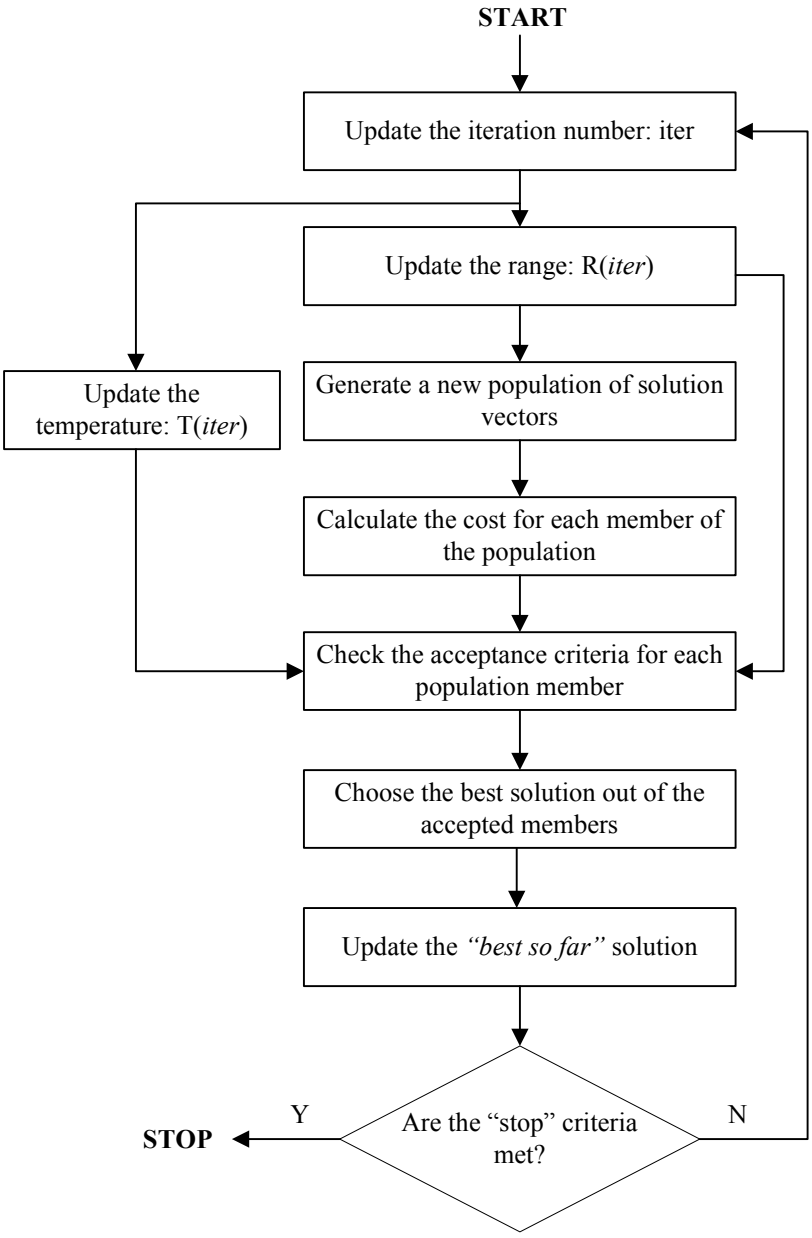

Figure 6.10: The flowchart of the RCESA optimization algorithm.

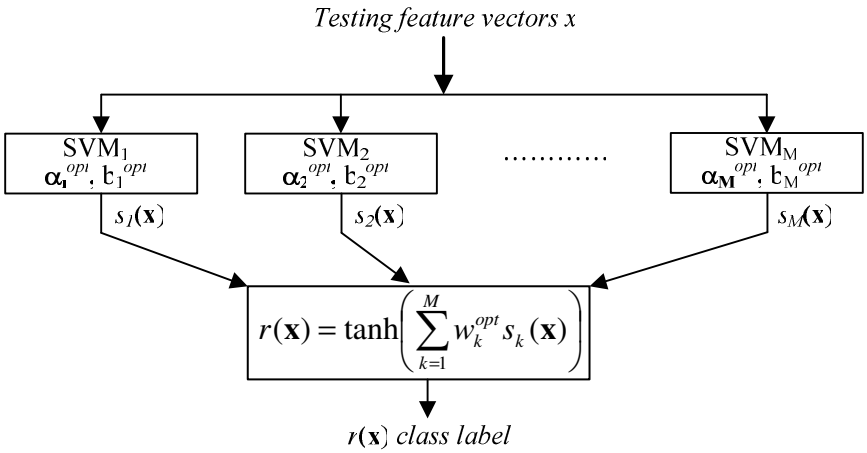

Figure 6.11: The OPSVM classification process.

Chapter Seven:

EXPERIMENTAL RESULTS ON ACOUSTIC FEATURE CATEGORIES IN SPEECH OF DEPRESSED AND CONTROL ADOLESCENTS

"New ideas pass through three periods: 1) It can't be done. 2) It probably can be done, but it's not worth doing. 3) I knew it was a good idea all along!"

– Arthur C Clarke

7.1 Preview

This chapter is directly focused on the experimental results based on the depression detection framework explained in *Section 6.1*. Investigation on the 5 acoustic feature categories of the Teager-based, cepstral, prosodic, spectral and glottal features were analyzed to distinguish which acoustic feature categories or combinations of feature categories correlated with depression in speech of adolescents. Two methods of machine learning techniques of the Gaussian mixture model and support vector machine were carried out in the modeling and classification of speech in the depressed and control group. We begin with a brief introduction of the steps taken in the experimental setup, followed by the results of these experiments in the later sections.

7.2 Experimental Setup

Experiments with the framework outlined in *Section 6.1* in detecting clinical depression in speech of adolescents' were carried out using the database described in *Chapter 5*. To

avoid over-fitting the data, the data was divided so that approximately 50% of the adolescents, including 33 depressed (23 females and 10 males) and 34 control subjects (21 females and 13 males) were used for testing, and the remaining data was used for training the depressed and control models. A series of experiments that were conducted are briefly explained below and the results of these experiments are discussed in the following sections.

- $EXP_{preliminary}$: Optimizing the number of filters and number of coefficients in the acoustic feature of mel frequency cepstral coefficient (MFCC) in the cepstral category.
- *EXP1*: Removing feature coefficients that gave a statistically non-significant result in the modeling of depressed and control speech from MANOVA and ANOVA.
- *EXP2*: Using two feature categories i.e., TEO-based & cepstral (C), the effectiveness of gender-independent and gender-dependent modeling techniques for depressed and control adolescent classification was examined.
- *EXP3*: Next, testing on different lengths (durations) of utterances from the testing set was examined.
- *EXP4*: Using the best gender modeling strategy and optimal test utterance length found in EXP2 & EXP3, the effectiveness of other feature categories of prosodic (P), spectral (S), glottal (G) and their combinations for depressed and control adolescent classification was investigated.

- *EXP5*: From our database described in *Chapter 5*, the study of feature categories proposed in recent published work by others was carried out.
- *EXP6*: Next, the top feature category out of the TEO-based & C categories that yielded the highest classification accuracy was selected based on their performances in EXP2 and combined with the other feature categories (P, S and G) and their combinations as described in EXP4. In addition, the test of statistical significance on the classification accuracy improvements were determined by using McNeamar's test on paired feature categories.
- *EXP7*: Due to the high performance in modeling speech contents, GMM was employed for modeling speech of depressed and control adolescents in EXP2-EXP6. In the final experiment, the best feature category combination obtained in EXP6 was compared with the SVM classifier because of the advantageous properties of its generalization capabilities for solving two class problems.
- *EXP8*: Finally, a feature selection filter approach was implemented to select relevant and informative feature coefficients of the top acoustic feature category using the optimal classifier.

7.3 Evaluation Methods

The main objective was to correctly classify the test adolescents (alternatively called subjects) as either depressed or control. The subject based correct classification accuracy (SBCCA) is written as follows:

$$\left.\begin{array}{l}\text{Subject based}\\ \text{correct classification}\\ \text{accuracy (SBCCA)}\end{array}\right\} = \frac{\left(\begin{array}{l}\text{Number of correctly}\\ \text{classified subjects}\end{array}\right) * 100\%}{\text{Total number of subjects}} \tag{7.1}$$

To determine the number of correctly classified subjects in Eq. (7.1), the utterance-based correct classification accuracy (UBCCA) was first calculated using the following formula:

$$\left.\begin{array}{l}\text{Utterance based}\\ \text{correct classification}\\ \text{accuracy (UBCCA)}\end{array}\right\} = \frac{\left(\begin{array}{l}\text{Number of correctly}\\ \text{classified utterances}\end{array}\right) * 100\%}{\text{Total number of utterances}} \tag{7.2}$$

If UBCCA for a given subject was greater than 50% for the depressed class, then the subject was classified as depressed. Since, the predicted classes of the test subjects were known; the total number of correctly classified subjects could therefore be calculated and used in Eq. (7.1) to determine the SBCCA values. In addition, the correct classification of depressed and control subjects were measured in terms of sensitivity, specificity and the overall accuracy defined as follows:

True positive (TP) = Number of depressed subjects classified as depressed
False negative (FN) = Number of depressed subjects classified as control
True negative (TN) = Number of control subjects classified as control
False positive (FP) = Number of control subjects classified as depressed

$$\text{Sensitivity} = \frac{TP}{TP + FN} \times 100\% \qquad \text{Specificity} = \frac{TN}{TN + FP} \times 100\% \tag{7.3}$$

$$\text{Overall accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \times 100\% \tag{7.4}$$

In general, since the aim of this research work was to provide an early stage screening device to assists psychologists in their diagnosis of clinical depression before going through a full scale clinical evaluation, the objective was to identify more depressed subjects (higher specificity) rather than to screen out negative cases (higher specificity). Thus, the highest overall classification accuracy was achieved by obtaining an optimal sensitivity to specificity ratio (ideally>1) and at the same time keeping the ratio between sensitivity and specificity at a reasonable margin (i.e., ratio should not be >2) to avoid class skews. However, in some cases, it was not possible to achieve reasonably high accuracy without making the sensitivity to specificity ratio <1.

All classification results were cross-validated based on four turns using different training and testing sets.

7.4 MFCC – Optimized Number of Filters and Coefficients ($EXP_{Preliminary}$)

First, optimization of the parameters in the MFCC was carried out to maximize the detection accuracy of depressed and control subjects. The overall correct classification accuracy was plotted as a function of: *1) number of filters* (from 10 to 60 with a step size of 10; and *2) number of coefficients* (from 6 to 30 with a step size of 6). Based on these plots, we selected the number of filters and the number of coefficients that were giving the highest overall classification accuracy. For the MFCC parameter optimization, half of the utterances per subject (10 min of speech) from each interaction (EPI, PSI and FCI) were used and divided equally for training 1024 Gaussian mixtures by iteratively increasing the number of Gaussian mixture components by a factor of two (i.e., 2, 4, 8, 16, 32, to 1024) for depressed and control classes. As shown in Fig. 7.1, it was found that

30 filters in the filter bank and 12 coefficients maximized the overall correct classification accuracy of 55.8%. The optimized parameters were then used on the full database for the MFCC feature extraction in further experiments.

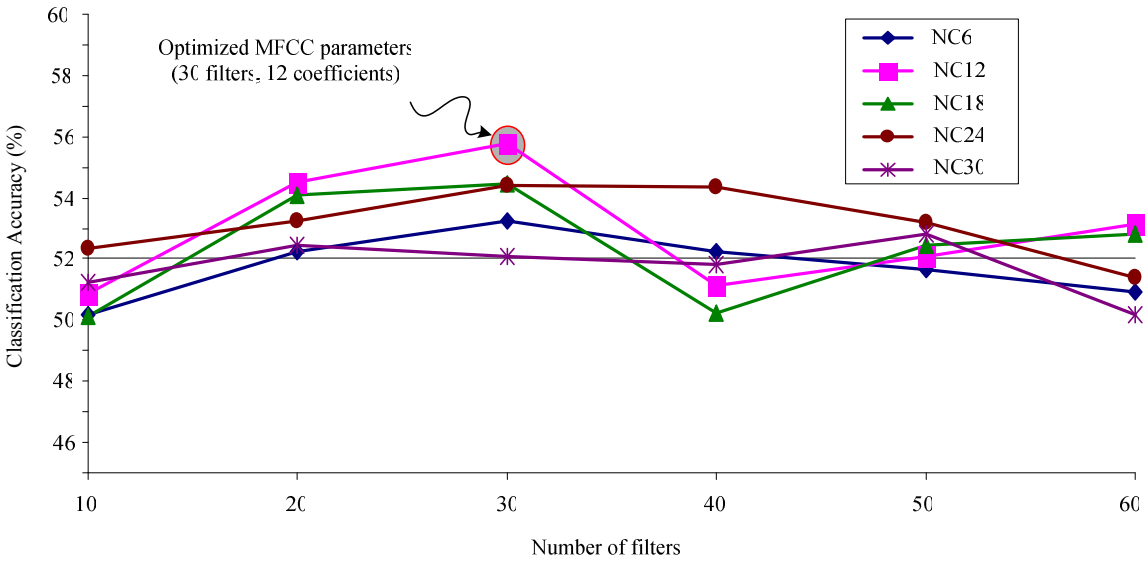

Figure 7.1: Maximizing the mel frequency cepstral coefficients (MFCC) classification accuracy by tuning the number of filters and number of coefficients (NC) using 1024 Gaussian mixtures.

Table 7.1: MANOVA AND ANOVA ANALYSIS ON THE SUBCATEGORY FEATURES FOR BOTH MALE AND FEMALE ADOLESCENTS

Category	Sub-category features[a]	No. of feature coeff.	Significance (male)			Significance (female) – In all interactions of EPI, PSI and FCI, all the feature sub-categories are used; i.e. 186 coefficients from all the features for each interaction
			EPI	PSI	FCI	
TEO	TEO-CB-Auto-Env	45	+	+	+	
Cepstral (C)	MFCC	36	+	+	+	
Prosodics (P)	F0	3	-	+	+	
	LogE	3	+	+	+	
	FMTS & FBWS	18	+	+	+	
	Jitter	3	-	+	-	
	Shimmer	3	+	+	+	
Spectral (S)	Centroid	3	+	-	-	
	Flux	3	+	+	+	
	Entropy	3	+	+	+	
	Roll-off	3	+	+	+	
	PSD	27	+	+	+	
Glottal (G)	GLT	27	+	+	+	
	GLF	9	+	+	+	
	Total	186	180	183	180	

[a] All features include their delta (Δ) and delta-delta (Δ-Δ)

7.5 Statistical Results – MANOVA and ANOVA (*EXP1*)

Table 7.1 presents the extracted acoustic features grouped into categories and subcategories for both male and female subjects from each interaction. A total of 14 acoustic features comprising of 186 feature coefficients, which included their delta (Δ) and delta-delta (Δ-Δ) coefficients for all the frames in the utterances, were statistically examined for significance in characterizing speech of depressed and control adolescents.

Results of the Kolmogorov-Smirnov (KS) test [35] for normal distribution of the dataset indicated that the entire feature coefficients were normally distributed (p>0.05). Instead of combining all the feature coefficients in MANOVA, which is considered a sub-optimal approach unless there is a good theoretical basis for doing so [35], individual subcategories representing the acoustic features along with their delta and delta-delta

coefficients were examined with MANOVA. The reason behind this approach was that incorporation of the delta and delta-delta coefficients in previous work [76], [77] has shown to result in an increase in classification results, and therefore correlations should exists between the feature coefficients in each sub-categorical feature in MANOVA.

In MANOVA, multivariate group tests were performed on each sub-categorical feature using *Wilks's lambda*. Features in the subcategory that met a significance level of $p<0.05$ were retained. Otherwise, the feature was then followed up with a one-way analysis of variance (ANOVA) on each feature coefficient. Each feature coefficient that also met a significance level of $p<0.05$ for ANOVA were kept in that subcategory. Otherwise, if all the feature coefficients in the sub-categorical feature were still non-significant, the sub-categorical feature was discarded.

Selected acoustic features in the subcategories and the number of coefficients are listed in Table 7.1. The plus sign indicates that the sub-categorical feature produced a statistically significant result and the minus sign indicates that the result was statistically non-significant. We found that all the features (in the sub-category) were statistically different between depressed and control speech of female adolescents in all the three interactions. However, for speech in the male adolescents, a few features were not significant, as indicated by a minus sign in Table 7.1.

As an illustration, Fig. 7.2 shows the error bar graphs of F0 that displays the means and 95% confidence intervals between both groups of depressed (D) and control (C) adolescents. As shown in Fig. 7.2, for the event planning interaction (EPI) task in male subjects, the error bars in both groups overlap considerably, indicating that these samples were unlikely from different groups. In other words, there are maybe no

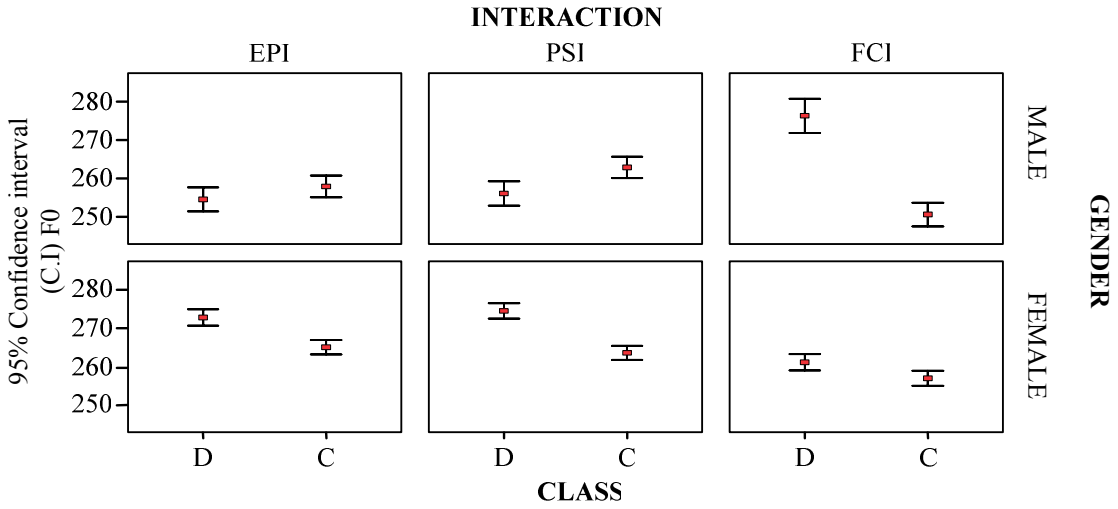

Figure 7.2: The error bars of F0 for depressed and control adolescents.

significant differences between the means of these two samples. Indeed, as graphed in Fig. 7.2 and summarized in Table 7.1, the acoustic feature of F0 for male subjects in EPI did not indicate a significant difference and therefore was discarded.

7.6 Effectiveness of Gender Independent vs. Gender Dependent Modeling (*EXP2*)

As noted in the database description in *Chapter 5*, although participants were matched on a range of demographic variables, we only considered gender in our analyses because the development of gender differences in depressive symptoms has been documented to occur during early adolescence [89]. Accordingly, the examination of how gender differences might affect classification accuracies in clinical depression was analyzed. For the purpose of this experiment, two feature categories (Table 7.1) of TEO-based and cepstral (C) were selected as a starting point for the analysis. These two types of features have been previously reported as effective discriminators of stress and emotion in speech

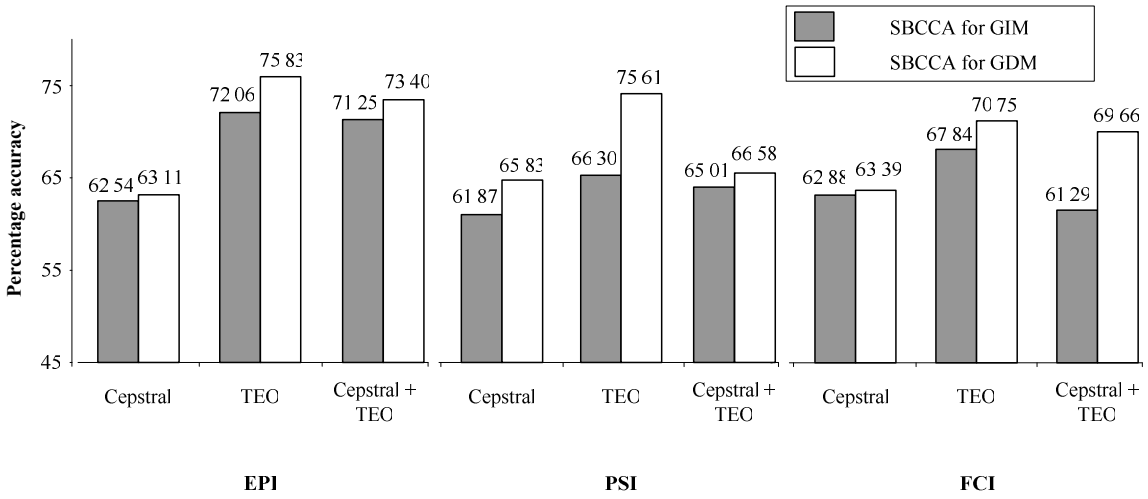

Figure 7.3: SBCCA for the TEO-based and cepstral features using GIM and GDM.

[123] and also in speaker and language detection [105]. Since depression is often characterized as an affect (emotion) regulation [108] disorder, it was expected that the cepstral and TEO-based based features could also provide good results in the depression detection in speech. The gender-dependent depressed and control class models were generated separately from the feature vectors of male and female subjects using the GMM based training procedure. The gender-independent depressed and control class models on the other hand, were trained by combining together feature vectors from both male and female subjects. Similar testing length as in [91] of 0.5 min for each utterance was achieved by concatenating only the voiced sections from the utterances belonging to each adolescent. Depressed and control classification using gender-dependent models were carried out assuming that the gender of the test adolescent was known.

Fig. 7.3 shows the overall classification performances based on subject based correct classification accuracies (SBCCA) for both gender independent modeling (GIM) and gender dependent modeling (GDM). In all the interactions, as can be observed in Fig.

Table 7.2: TEO-BASED CATEGORY PERFORMANCE USING SBCCA WITH 0.5 MIN TEST UTTERANCES ON GIM AND GDM - SENSITIVITY AND SPECIFICITY RESULTS

Training & Testing feature: TEO							
Modeling Strategy		EPI		PSI		FCI	
		Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
GIM (Male & Female)		74.59	69.52	51.17	81.43	53.42	82.26
GDM	Male	76.39	82.97	81.39	71.57	62.22	84.89
	Female	76.95	75.00	84.71	64.77	75.65	60.23

7.3, GDMs outperformed the GIMs for the calculation of SBCCA. It was also observed in Fig. 7.3 that the TEO-based category consistently outperformed the cepstral category and cepstral+TEO-based category combination in both the GDMs and GIMs. Compared to TEO-based category with GIMs, TEO-based category with GDMs improved SBCCA by 3.8%, 9.3% and 2.9% for EPI, PSI and FCI respectively.

Table 7.2 shows the sensitivity and specificity in the SBCCA of the TEO-based feature category for GIM and GDM. From the sensitivity results in the table, it can be observed that GDMs performed better than GIMs in detecting depressed subjects when both males and females are modeled separately in all interaction contexts. GDMs for the males resulted in an accuracy improvement in sensitivity of 1.8%, 30.2% and 8.8% for EPI, PSI and FCI respectively compared to GIMs. For the females, GDMs also resulted in an accuracy improvement in sensitivity of 2.4%, 33.5% and 22.2% for EPI, PSI and FCI respectively compared to GIMs.

7.7 Optimal Test Utterance Length for Analysis (*EXP3*)

In previous research, test utterances of 20 seconds [38] and 0.5 minute [91] in length have been used for depressed and control subject classification. To determine the optimal

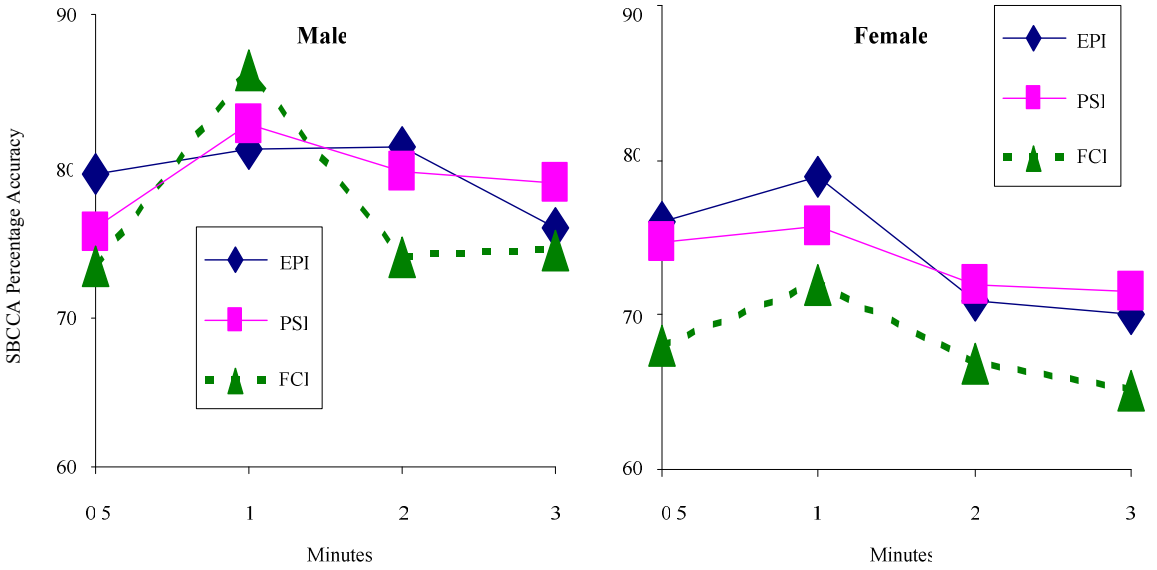

Figure 7.4: Classification accuracies using different concatenated test utterances length for TEO-based feature category using the GMM classifier.

duration of speech samples to be analyzed, we carried out EXP3 to examine SBCCA with different duration (i.e. 0.5mins, 1min, 2min and 3min) of concatenated voiced utterances using the GMM classifier. In these experiments, we used the feature category of TEO-based with the GDMs because this setting outperformed the others from previous experiments in EXP2. Experimental results are shown in Fig. 7.4. With reference to the accuracies using 1min utterances, we noticed around 7.1%, 5.1% and 7.0% average SBCCA drops (of all interactions) for the male subjects and 2.6%, 5.6% and 6.6% SBCCA drops for the female subjects using 0.5min, 2min and 3min utterance durations respectively. It can be observed in Fig. 7.4, that the SBCCA measure was consistently achieving the highest value for utterance length of 1 minute; this length of the test utterances was therefore chosen as a length to be used in our subsequent experiments.

Table 7.3: CLASSIFICATION PERFORMANCE OF PROSODIC, SPECTRAL, AND GLOTTAL FEATURE CATEGORIES USING SBCCA WITH 1 MIN TEST UTTERANCES - MALE RESULTS.

Training/Testing Features		EPI			PSI			FCI		
		Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
P		51.39	68.27	59.83	45.83	55.91	50.87	64.72	53.02	58.87
S		62.50	60.03	61.26	61.68	41.62	51.64	44.44	57.69	51.07
G		57.50	61.68	59.59	69.44	79.67	74.56	80.83	51.24	66.03
P + S		61.94	61.95	61.95	62.22	54.40	58.31	63.89	51.92	57.91
P	+ G	62.50	58.79	60.65	54.72	77.20	65.96	61.11	68.68	64.90
S	+ G	65.28	60.03	62.65	65.56	42.86	54.21	64.17	51.37	57.77
P + S	+ G	78.06	54.95	66.50	76.67	57.69	67.18	64.44	73.76	69.10

Table 7.4: CLASSIFICATION PERFORMANCE OF PROSODIC, SPECTRAL, AND GLOTTAL FEATURE CATEGORIES USING SBCCA WITH 1 MIN TEST UTTERANCES - FEMALE RESULTS.

Training/Testing Features		EPI			PSI			FCI		
		Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
P		62.75	70.45	66.60	60.79	73.86	67.33	60.92	63.64	62.28
S		42.38	87.50	64.96	61.04	55.12	58.08	67.17	71.59	69.38
G		73.38	70.45	71.91	72.13	53.41	62.77	75.33	71.59	73.46
P + S		50.58	88.64	69.61	76.21	60.23	68.22	63.79	68.18	65.99
P	+ G	67.92	70.45	69.19	77.71	53.41	65.56	67.79	69.32	68.55
S	+ G	76.29	67.05	71.67	75.42	57.95	66.69	87.75	52.27	70.01
P + S	+ G	73.21	71.59	72.40	83.46	67.05	75.25	66.96	73.86	70.41

7.8 Effectiveness of Prosodic, Spectral, Glottal Feature Categories, and Their Combinations *(EXP4)*

Further analyses were conducted on the other different feature categories representing prosodic (P), spectral (S) and glottal (G) using the GDMs. Table 7.3 and Table 7.4 presents the overall accuracy results for the male and female adolescents respectively that is based on the subject based correct classification accuracy (SBCCA) for feature categories P, S, G and their different combinations in all the interactions (EPI, PSI and FCI). From the tables, a few key findings can be observed.

Firstly, in Table 7.3 for the males, the influence of G on individual categories P and S (i.e., P+G and S+G) improved classification accuracy compared to P and S alone. Compared to P alone, P+G increased SBCCA by 0.8%, 15.1% and 6% in the EPI, PSI

and FCI respectively. Also for the males, compared to S alone, S+G increased accuracy by 1.4%, 2.6% and 6.7% for the EPI, PSI and FCI respectively.

Secondly, in Table 7.4 for the females, combining G with other feature categories also improved classification rates in the EPI and FCI. In the EPI and FCI for females, P+G showed a 2.6% and 6.3% improvement over P alone. In the EPI, PSI and FCI, S+G showed a 6.7%, 8.6% and 0.6% improvement over S alone. However in PSI, a slight decrease of 1.8% was shown for P+G when compared to P alone.

Thirdly, as shown in Tables 7.3 and 7.4 for both males and females respectively, combining G with P+S also improved the classification accuracy for all the interactions. In the case for males, P+S+G improved accuracy over P+S by 4.6%, 8.9% and 11.2% for EPI, PSI and FCI respectively. For the females, P+S+G also improved accuracy over P+S by 2.8%, 7% and 4.4% for EPI, PSI and FCI respectively.

Fourthly, for the EPI and FCI, P, S, G and their different combinations were better discriminators in the female than the male samples. A similar trend emerged for the PSI, with the exception that G and P+G for females showed a decreased in SBCCA rates over the males of 11.8% and 0.4% respectively.

Based on the above observations, it is more likely that feature category G and its combination with P and S can improve SBCCA for both male and female subjects.

Table 7.5: CLASSIFICATION PERFORMANCE OF FEATURE CATEGORIES PROPOSED IN [84] USING SBCCA WITH 1 MIN TEST UTTERANCES - MALE RESULTS

Training/Testing Features	EPI			PSI			FCI		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
P_c	54.44	60.44	57.44	59.72	56.32	58.02	50.83	72.39	61.61
$P_c + V_c$	51.39	68.27	59.83	45.83	55.91	50.87	64.72	53.02	58.87
P_c + G	54.17	79.12	66.64	71.67	57.69	64.68	75.28	55.22	65.25
$P_c + V_c$ + G	62.50	58.79	60.65	54.72	77.20	65.96	61.11	68.68	64.90

Table 7.6: CLASSIFICATION PERFORMANCE OF FEATURE CATEGORIES PROPOSED IN [84] USING SBCCA WITH 1 MIN TEST UTTERANCES - FEMALE RESULTS

Training/Testing Features	EPI			PSI			FCI		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
P_c	67.13	70.45	68.79	60.47	62.94	61.70	57.83	71.59	64.71
$P_c + V_c$	62.75	70.45	66.60	60.79	73.86	67.33	60.92	63.64	62.28
P_c + G	91.71	61.36	76.54	56.79	72.73	64.76	80.25	52.27	66.26
$P_c + V_c$ + G	67.92	70.45	69.19	77.71	53.41	65.56	67.79	69.32	68.55

7.9 Study of Feature Categories Proposed in Recent Published Work by Others (*EXP5*)

The classification performances of the proposed feature categories of prosodic (P_o) vocal tract (V_o) and glottal (G_o) defined in [84] was examined with our database. This examination was aimed at determining if we could establish any similar trends or performances in defining our own feature categories and sub-categories with those proposed in [84]. A major point to take note in the analysis was that the feature analysis strategy in [84] used different implementations in analysis methods in extracting the acoustic features. Additionally, statistical quantifiers were used in [84] (i.e., mean, variance, median, percentile, etc.) over frames of extracted features of an utterance or sentences with the number of features ranging from 156-872 (depending on the timing/grouping of the data). Whereas, features proposed in our study did not include any feature statistics. Instead, only the short-term features (i.e., features that are extracted on a speech frame basis) were computed.

Results are presented in Table 7.5 and Table 7.6. Note that although the same groupings of feature categories as [84] were implemented, the sub-categorical features were slightly different. Also, in [84] the sub-categorical features of formants and formant bandwidths were partitioned to have another feature category that represented measurements of the vocal tract shape and length (denoted V_o in Table 7.5 and Table 7.6). However, for our grouping (see Table 7.1), the prosodic feature category contained the sub-categorical feature of formants and formant bandwidths. Therefore, it is not surprising that the results for P_o+V_o and P_o+V_o+G_o for males in Table 7.5 were the same as P and P+G respectively in Table 7.3. This was also the same for the females in Table 7.6, whereby the results for P_o+V_o and P_o+V_o+G_o were the same as P and P+G in Table 7.4.

Similar to results based on the combinations of glottal features, an improvement in classification rates was seen for Po+Go compared to Po alone for both male and female adolescents (Table 7.5 and Table 7.6). Comparing Table 7.3 and Table 7.5 in the classification accuracy for the male adolescents and comparing Table 7.4 and Table 7.6 in the accuracy for the female adolescents, it was observed that using our entire feature category combinations of P+S+G yielded better classification accuracies when compared to the grouping of Po+Vo+Go.

For the males, P+S+G (Table 7.3) gave a 5.9%, 1.2% and 4.2% classification accuracy increase in the EPI, PSI and FCI, respectively as compared to P_o+V_o+G (Table 7.5). For the females, P+S+G (Table 7.4) gave a 3.2%, 9.7% and 1.9% classification accuracy increase in EPI, PSI and FCI respectively as compared to P_o+V_o+G_o (Table 7.6).

Table 7.7: INFLUENCE OF TEO-BASED FEATURE CATEGORY ON CLASSIFICATION PERFORMANCE USING SBCCA WITH 1 MIN TEST UTTERANCES - MALE RESULTS

Training/Testing Features		EPI			PSI			FCI		
		Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
TEO		81.67	81.04	81.36	86.94	78.98	82.96	80.83	92.45	86.64
P	+ TEO	81.94	80.91	81.43	89.17	75.27	82.22	75.28	92.72	84.00
S	+ TEO	76.11	81.04	78.58	75.56	83.38	79.47	75.28	85.44	80.36
G	+ TEO	86.67	77.34	82.00	81.67	81.04	81.36	75.83	82.97	79.40
P + S	+ TEO	76.11	79.12	77.62	78.33	60.99	69.66	78.06	76.10	77.08
P + G	+ TEO	83.89	75.41	79.65	84.44	82.83	83.64	78.33	82.97	80.65
S + G	+ TEO	83.89	75.27	79.58	84.72	73.35	79.04	78.06	77.47	77.76
P + S + G	+ TEO	81.39	79.12	80.25	84.44	78.85	81.65	78.06	81.32	79.69

Table 7.8: INFLUENCE OF TEO-BASED FEATURE CATEGORY ON CLASSIFICATION PERFORMANCE USING SBCCA WITH 1 MIN TEST UTTERANCES - FEMALE RESULTS

Training/Testing Features		EPI			PSI			FCI		
		Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
TEO		80.64	77.27	78.96	81.38	70.02	75.70	72.08	71.94	72.01
P	+ TEO	70.08	68.18	69.13	58.96	73.86	66.41	67.88	65.91	66.89
S	+ TEO	58.83	72.73	65.78	56.58	68.18	62.38	64.96	63.64	64.30
G	+ TEO	78.42	79.55	78.98	69.08	65.91	67.50	69.21	65.91	67.56
P + S	+ TEO	70.00	71.59	70.80	55.58	68.18	61.88	69.04	61.36	65.20
P + G	+ TEO	75.21	61.36	68.29	80.63	64.77	72.70	71.29	67.05	69.17
S + G	+ TEO	69.96	72.73	71.34	78.46	60.23	69.34	68.21	67.05	67.63
P + S + G	+ TEO	66.92	73.86	70.39	71.33	78.41	74.87	67.00	71.50	69.25

7.10 Performance Analysis by Combining TEO-based Category with Prosodic, Spectral, and Glottal Categories (*EXP6*)

In EXP2, the TEO-based features showed the best overall performance, therefore in the next stage of our experiments, the effect of combining the TEO-based features with the P, S and G features was investigated. Table 7.7 and Table 7.8 presents the classification results in terms of sensitivity, specificity and overall accuracy for males and females respectively when TEO-based category was added to the different combinations of P, S and G. Table 7.9 summarizes the influences in percentage increase or decrease in accuracies when the feature category of TEO-based was added to P, S and G, compared with having P, S and G and their different combinations alone as shown in Table 7.3 and Table 7.4.

Table 7.9: INFLUENCE OF TEO-BASED CATEGORY IN PERCENTAGE ACCURACY IMPROVEMENT WHEN ADDED TO PROSODIC, SPECTRAL, AND GLOTTAL CATEGORIES

Training/Testing Features		Overall Accuracy increase (+) / decrease (-)					
		EPI		PSI		FCI	
		MALE	FEMALE	MALE	FEMALE	MALE	FEMALE
P	+ TEO	**+21.60%**	+2.43%	**+31.35%**	-0.92%	**+25.13%**	+4.61%
S	+ TEO	**+17.32%**	+0.82%	**+27.83%**	+4.30%	**+29.29%**	-5.08%
G	+ TEO	**+22.41%**	+7.07%	**+6.80%**	+4.73%	**+13.37%**	-5.90%
P + S	+ TEO	**+15.67%**	+1.19%	**+11.35%**	-6.34%	**+19.17%**	-0.79%
P + G	+ TEO	**+19.00%**	-0.90%	**+17.68%**	+7.14%	**+15.75%**	+0.62%
S + G	+ TEO	**+16.93%**	-0.30%	**+24.83%**	+2.65%	**+19.99%**	-2.38%
P + S + G	+ TEO	**+13.75%**	-2.01%	**+14.47%**	-0.38%	**+10.59%**	-0.76%

Table 7.10: MCNEMAR'S TEST OF STATISTICAL SIGNIFICANCE IN PERCENTAGE ACCURACY IMPROVEMENT WHEN TEO-BASED CATEGORY WAS ADDED TO PROSODIC, SPECTRAL, AND GLOTTAL CATEGORIES

McNemar's test on paired feature categories		EPI		PSI		FCI	
		Male P-value	Female P-value	Male P-value	Female P-value	Male P-value	Female P-value
P	P+TEO	**0.012**	0.561	**0.000**	0.603	**0.000**	0.313
S	S+TEO	**0.045**	0.584	**0.001**	0.440	**0.001**	0.281
G	G+TEO	**0.003**	0.143	**0.035**	0.422	**0.033**	0.200
P+S	P+S+TEO	**0.030**	0.451	**0.030**	0.137	**0.007**	0.432
P+G	P+G+TEO	**0.009**	0.222	**0.014**	0.118	**0.035**	0.400
S+G	S+G+TEO	**0.015**	0.137	**0.000**	0.544	**0.025**	0.538
P+S+G	P+S+G+TEO	**0.035**	0.454	**0.011**	0.450	**0.082**	0.161

*Accuracies highlighted in bold indicate the McNemar s test results for the statistically significant accuracy increments ($p<0.05$).

For the males, when TEO-based feature category was combined, a significant accuracy increment was observed in all the interactions. Interestingly, in most cases, the TEO-based feature category on its own showed higher classification accuracy for both male (Table 7.7) and female adolescents (Table 7.8) throughout all the interactions.

Since the addition of the TEO-based features to the other feature categories (prosodic, glottal, spectral and their different combinations) was clearly increasing the classification accuracy, statistical analysis was applied to determine if these improvements were statistically significant (95% confidence intervals).

For our case, we decided that the McNeamar's test on paired feature categories between four fold cross-validation classification accuracies would be more appropriate, as other test for significance (i.e., t-test) would have low statistical power due to the few number of cross-validation turns in the classification accuracy. As shown in Table 7.10, the increment in accuracies were statistically significant (p <0.05) for all the males (highlighted in bold).

7.11 Comparison with SVM Classifier (*EXP7*)

The results that have been discussed so far were based on the GMM. To examine whether the classification accuracies were biased with respect to GMM, the best results obtained with GMM i.e., TEO-based category with GMM were compared with the results of TEO-based category with SVM.

SVM was implemented in the form of optimized parallel support vector machine (OPSVM) discussed in *Section 6.5.2*. The number of subsets in OPSVM to be trained was varied from 10 to 30 with a step size of 10. The number of subsets (or SVMs) in OPSVM that gave the highest classification was chosen in the model selection. The optimal number of subsets (or SVMs) in OPSVM which maximized the SBCCA with the TEO-based category was 20. Table 7.11 shows the results of equal weights and optimal weights calculated from the global optimization algorithm in the OPSVM. In the male sample, compared to using equal weights, the optimal weights in OPSVM increased the SBCCA by approximately 5.4%, 10.3% and 3.9% in the EPI, PSI and FCI respectively. In the female sample, SBCCA increments in using optimal weights were approximately 10.2%, 6.6% and 8.9% in the EPI, PSI and FCI respectively. Although OPSVM yielded

very similar results as compared with the GMM modeling technique, the computational time required in training the models in our dataset was less efficient compared to the GMM.

Table 7.11: OPSVM CLASSIFICATION RESULTS FOR TEO-BASED FEATURE CATEGORY USING SBCCA WITH 1 MIN TEST UTTERANCES

Event Planning Interaction (EPI)						
SVM weights	Male			Female		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
Equal	100	46.15	73.08	58.33	81.82	70.08
Optimal	83.50	73.36	78.43	78.75	81.82	80.29
Problem Solving Interaction (PSI)						
SVM weights	Male			Female		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
Equal	93.88	30.77	62.33	54.17	81.82	68
Optimal	75.10	70.24	72.67	58.33	90.91	74.62
Family Consensus Interaction (FCI)						
SVM weights	Male			Female		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
Equal	77.78	84.62	81.2	70.83	68.18	69.51
Optimal	77.78	92.31	85.05	75	81.82	78.41

7.12 Feature Selection on Top Feature Category and Optimal Classifier *(EXP8)*

Having decided from previous experiments (EXP1-EXP7) on the gender modeling strategy (i.e., *GDM*), optimal test utterance length (i.e., *1 min*), top feature category (i.e., *TEO-based*) and optimal classifier (i.e., *GMM*) in obtaining the highest classification accuracy performances in clinical depression detection from the speech of adolescents, further evaluation was conducted to optimize the top performing acoustic feature category of TEO-based. This was carried out by removing feature coefficients that did not contain important information through a feature selection filter approach so that it could improve the generalization capabilities of the classifier.

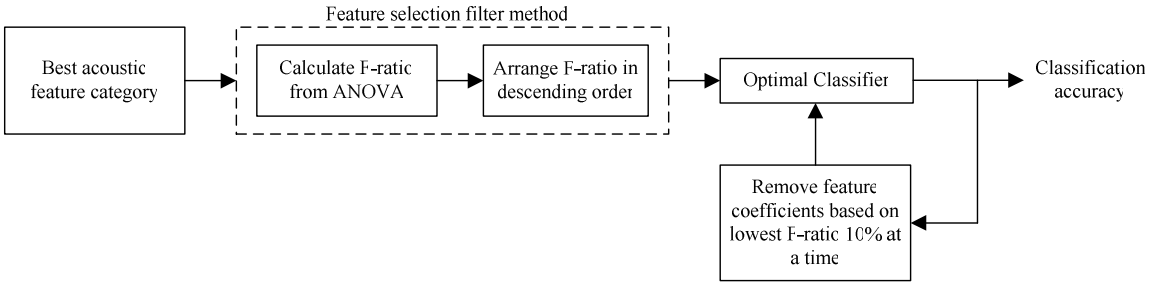

Figure 7.5: Framework of feature selection filter approach using top acoustic feature category and optimal classifier.

The filter approach method implemented here was a feature ranking method based on F-ratio scores from ANOVA. Looking back at *Section 7.5*, the statistical analysis of ANOVA produces an F-ratio score which measures the discrimination between classes by comparing the amount of systematic variance in the data to the amount of unsystematic variance. The larger the F-ratio the more likely the means of the feature coefficients between the classes are more discriminative. We chose the filter approach as it was computationally less expensive due to the size of our dataset.

Therefore, from our experiments, using the filter approach on the feature category of TEO-based with the GMM classifier, the following procedures as illustrated in Fig. 7.5 were implemented:

(1) The F-ratio was calculated for every feature coefficient and arranged in descending order.

(2) The classification accuracy of the filter that is based on the F-ratio scores was empirically measured by removing the feature coefficients with the lowest F-ratio scores 10% at a time.

For the males, Table 7.12, Table 7.13 and Table 7.14 presents the ranking of the TEO-based feature coefficients in descending order that was based on their ANOVA F-ratio scores for the tasks of EPI, PSI and FCI respectively.

Table 7.15, Table 7.16 and Table 7.7 presents the F-ratio scores for the females in the tasks of EPI, PSI and FCI respectively. The arrows in the tables indicate the top percentage of feature coefficients in the TEO-based category that were based on their F-ratio scores.

Fig. 7.6 and Fig 7.7 plots the classification accuracies of the filter approach on the TEO-based category that had features coefficients based on the lowest F-ratio scores removed 10% at a time for the males and females respectively. The highlighted circles in the figures indicate the highest overall classification accuracy of the filter obtained for each interaction. For the males, as shown in Fig. 7.6, the highest overall classification performances was achieved when the top 10% of feature coefficients for the EPI task, top 10% of feature coefficients for the PSI task and top 30% of feature coefficients for the FCI task based on the ranking of F-ratio scores were kept. For the females, as shown in Fig. 7.7, the highest overall classification performances came from the top 10% of feature coefficients for the EPI task, top 20% of feature coefficients for the PSI task and top 30% of feature coefficients for the FCI task. The breakdown of the highest classification performances of the filter method showing the sensitivity, specificity and overall accuracy in all the interactions for both the male and female adolescents is given in Table 7.18. A point to take note here is that searching for the best set of parameters by ranking their F-ratio scores from highest to lowest should ideally give the highest classification accuracy from the classifier when the TOP ranking F-ratio scores are kept (i.e., TOP 10%

of features kept). However, features that perform well or poorly or their own can have better or worse performances as part of ensemble features [44].

Comparing the accuracies with the original feature coefficients in the TEO-based category (Tables 7.7 & 7.8) and feature coefficients that were selected after the filter method (Table 7.18), the filter method used in the feature selection improved the accuracies throughout all the interactions in both genders. After applying the feature selection, the accuracies were improved by 2.4%, 2.2% and 1% in the tasks of EPI, PSI and FCI respectively for the males. For the females, the percentage improvement was 1%, 5.3% and 1.8% in the tasks of EPI, PSI and FCI respectively.

Table7.12: ANOVA F-RATIO SCORES OF THE TEO-BASED CATEGORY ARRANGED IN DESCENDING ORDER (EPI TASK - MALES)

	Sorted feature	F-ratio
	TEO CB15	67.37
	TEO CB13	56.33
	TEO CB14	46.47
	TEO CB1	40.65
Top 10% of features kept	TEO CB2	31.76
	TEO CB3	14.91
	TEO CB13 Δ-Δ	11.80
	TEO CB5	11.75
Top 20% of features kept	TEO CB15 Δ-Δ	11.68
	TEO CB12	11.19
	TEO CB14 Δ-Δ	10.68
	TEO CB15 Δ	9.80
	TEO CB13 Δ	8.97
Top 30% of features kept	TEO CB11 Δ-Δ	7.77
	TEO CB12 Δ-Δ	7.72
	TEO CB4	7.44
	TEO CB11	7.43
Top 40% of features kept	TEO CB14 Δ	7.21
	TEO CB9	4.89
	TEO CB10 Δ-Δ	4.76
	TEO CB10	4.35
	TEO CB6	4.00
Top 50% of features kept	TEO CB4 Δ-Δ	3.08
	TEO CB9 Δ	2.99
	TEO CB11 Δ	2.99
	TEO CB9 Δ-Δ	2.92
Top 60% of features kept	TEO CB7	2.79
	TEO CB8	2.76
	TEO CB12 Δ	2.67
	TEO CB8 Δ-Δ	2.43
	TEO CB10 Δ	2.33
Top 70% of features kept	TEO CB1 Δ-Δ	1.86
	TEO CB1 Δ	1.28
	TEO CB8 Δ	1.21
	TEO_CB3 Δ	1.06
Top 80% of features kept	TEO CB2 Δ	1.03
	TEO CB7 Δ	1.02
	TEO CB5 Δ-Δ	0.79
	TEO CB3 Δ-Δ	0.57
	TEO CB5 Δ	0.50
Top 90% of features kept	TEO CB4 Δ	0.39
	TEO CB2 Δ-Δ	0.19
	TEO CB7 Δ-Δ	0.17
	TEO CB6 Δ-Δ	0.02
	TEO CB6 Δ	0.02

Table 7.13: ANOVA F-RATIO SCORES OF THE TEO-BASED CATEGORY ARRANGED IN DESCENDING ORDER (PSI TASK - MALES)

	Sorted feature	F-ratio
	TEO CB15	78.63
	TEO CB13	63.67
	TEO CB14	48.71
	TEO CB15 Δ-Δ	20.03
Top 10% of features kept	TEO CB1	18.25
	TEO CB13 Δ-Δ	16.63
	TEO CB14 Δ-Δ	16.12
	TEO CB15 Δ	8.97
Top 20% of features kept	TEO CB12	8.80
	TEO CB13 Δ	8.45
	TEO CB11	6.25
	TEO CB14 Δ	5.95
	TEO CB12 Δ-Δ	5.48
Top 30% of features kept	TEO CB11 Δ-Δ	4.85
	TEO CB4	3.69
	TEO CB7	2.82
	TEO CB12 Δ	2.64
Top 40% of features kept	TEO CB5	1.81
	TEO CB2	1.67
	TEO CB11 Δ	1.26
	TEO CB5 Δ-Δ	1.14
	TEO CB8 Δ-Δ	1.07
Top 50% of features kept	TEO CB10 Δ-Δ	0.97
	TEO CB6 Δ-Δ	0.70
	TEO CB4 Δ-Δ	0.62
	TEO CB1 Δ	0.59
Top 60% of features kept	TEO CB8	0.58
	TEO CB10	0.56
	TEO CB6	0.55
	TEO CB3	0.53
	TEO CB4 Δ	0.53
Top 70% of features kept	TEO CB9 Δ-Δ	0.48
	TEO CB2 Δ	0.34
	TEO CB10 Δ	0.28
	TEO CB2 Δ-Δ	0.24
Top 80% of features kept	TEO CB7 Δ	0.17
	TEO CB9 Δ	0.17
	TEO CB5 Δ	0.13
	TEO CB7 Δ-Δ	0.12
	TEO CB3 Δ-Δ	0.06
Top 90% of features kept	TEO CB1 Δ-Δ	0.04
	TEO CB8 Δ	0.04
	TEO CB3 Δ	0.03
	TEO CB9	0.03
	TEO CB6 Δ	0.01

Table 7.14: ANOVA F-RATIO SCORES OF THE TEO-BASED CATEGORY ARRANGED IN DESCENDING ORDER (FCI TASK - MALES)

	Sorted feature	F-ratio
	TEO CB13	310.28
	TEO CB14	274.25
	TEO CB15	274.20
	TEO CB11	268.24
Top 10% of features kept	TEO CB12	265.38
	TEO CB10	227.41
	TEO CB9	224.05
	TEO CB8	155.08
Top 20% of features kept	TEO CB6	118.99
	TEO CB5	118.75
	TEO CB2	109.70
	TEO CB7	74.28
	TEO CB4	54.24
Top 30% of features kept	TEO CB3	50.28
	TEO CB1	44.63
	TEO CB13 Δ	34.31
	TEO CB11 Δ	30.20
Top 40% of features kept	TEO CB12 Δ	29.91
	TEO CB9 Δ	29.18
	TEO CB10 Δ	28.01
	TEO CB14 Δ	27.14
	TEO CB15 Δ	26.70
Top 50% of features kept	TEO CB8 Δ	13.84
	TEO CB13 Δ-Δ	13.59
	TEO CB14 Δ-Δ	13.11
	TEO CB11 Δ-Δ	12.51
Top 60% of features kept	TEO CB12 Δ-Δ	12.22
	TEO CB15 Δ-Δ	11.45
	TEO CB9 Δ-Δ	11.31
	TEO CB5 Δ	9.78
	TEO CB6 Δ	9.52
Top 70% of features kept	TEO CB2 Δ-Δ	8.28
	TEO CB3 Δ	8.02
	TEO CB4 Δ	7.93
	TEO CB10 Δ-Δ	7.44
Top 80% of features kept	TEO CB2 Δ	7.31
	TEO CB7 Δ	6.94
	TEO CB8 Δ-Δ	6.75
	TEO CB1 Δ-Δ	4.36
	TEO CB1 Δ	3.71
Top 90% of features kept	TEO CB6 Δ-Δ	2.89
	TEO CB3 Δ-Δ	2.35
	TEO CB5 Δ-Δ	2.16
	TEO CB4 Δ-Δ	1.83
	TEO CB7 Δ-Δ	1.63

Table 7.15: ANOVA F-RATIO SCORES OF THE TEO-BASED CATEGORY ARRANGED IN DESCENDING ORDER (EPI TASK - FEMALES)

	Sorted feature	F-ratio
	TEO CB5	50.83
	TEO CB9	32.26
	TEO CB6	27.24
	TEO CB10	20.36
Top 10% of features kept	TEO CB8	16.90
	TEO CB11	13.56
	TEO CB3	8.36
	TEO CB15_Δ-Δ	7.94
Top 20% of features kept	TEO CB12	7.65
	TEO CB4	6.62
	TEO CB7 Δ-Δ	6.44
	TEO CB3 Δ	6.11
	TEO CB14 Δ-Δ	5.62
Top 30% of features kept	TEO CB13 Δ-Δ	5.55
	TEO CB2 Δ-Δ	5.43
	TEO CB3 Δ-Δ	5.40
	TEO CB2 Δ	5.29
Top 40% of features kept	TEO CB1 Δ	4.73
	TEO CB14 Δ	4.47
	TEO CB12 Δ-Δ	4.40
	TEO CB1	4.19
	TEO CB15 Δ	4.18
Top 50% of features kept	TEO CB11 Δ	4.14
	TEO CB7 Δ	4.04
	TEO CB4 Δ-Δ	3.78
	TEO CB13 Δ	3.68
Top 60% of features kept	TEO CB2	3.59
	TEO CB7	3.37
	TEO CB10 Δ	3.10
	TEO CB12 Δ	3.02
	TEO CB9 Δ	2.16
Top 70% of features kept	TEO CB8 Δ	1.87
	TEO CB11 Δ-Δ	1.65
	TEO CB4 Δ	1.57
	TEO CB5 Δ	1.40
Top 80% of features kept	TEO CB6 Δ	1.39
	TEO CB6 Δ-Δ	1.34
	TEO CB8 Δ-Δ	1.22
	TEO CB13	1.09
	TEO CB9 Δ-Δ	1.04
Top 90% of features kept	TEO CB14	0.74
	TEO CB10 Δ-Δ	0.60
	TEO CB1 Δ-Δ	0.33
	TEO CB15	0.11
	TEO CB5 Δ-Δ	0.01

Table 7.16: ANOVA F-RATIO SCORES OF THE TEO-BASED CATEGORY ARRANGED IN DESCENDING ORDER (PSI TASK - FEMALES)

	Sorted feature	F-ratio
	TEO CB10	137.61
	TEO CB11	136.80
	TEO CB12	135.08
	TEO CB13	128.56
Top 10% of features kept	TEO CB14	113.73
	TEO CB9	99.09
	TEO CB15	91.93
	TEO CB5	80.73
Top 20% of features kept	TEO CB8	63.15
	TEO CB6	56.59
	TEO CB1 Δ	36.57
	TEO CB4	35.91
	TEO CB2 Δ	35.43
Top 30% of features kept	TEO CB7 Δ	33.77
	TEO CB2 Δ-Δ	30.65
	TEO CB3 Δ	28.91
	TEO CB1	28.89
Top 40% of features kept	TEO CB4 Δ	24.41
	TEO CB3 Δ-Δ	24.03
	TEO CB6 Δ	22.53
	TEO CB1 Δ-Δ	21.43
	TEO CB7 Δ-Δ	18.80
Top 50% of features kept	TEO CB4 Δ-Δ	18.09
	TEO CB8 Δ	17.48
	TEO CB5 Δ	15.89
	TEO CB9 Δ-Δ	14.00
Top 60% of features kept	TEO CB5 Δ-Δ	13.52
	TEO CB8 Δ-Δ	11.67
	TEO CB6 Δ-Δ	11.31
	TEO CB7	8.58
	TEO CB9 Δ	8.46
Top 70% of features kept	TEO CB10 Δ	8.06
	TEO CB10 Δ-Δ	6.20
	TEO CB11 Δ	5.82
	TEO CB12 Δ	5.43
Top 80% of features kept	TEO CB3	4.72
	TEO CB14 Δ	4.46
	TEO CB11 Δ-Δ	4.33
	TEO CB15Δ	4.05
	TEO CB2	3.54
Top 90% of features kept	TEO CB13Δ	3.28
	TEO CB14 Δ-Δ	2.78
	TEO CB15 Δ-Δ	2.70
	TEO CB13 Δ-Δ	2.56
	TEO CB12 Δ-Δ	2.16

Table 7.17: ANOVA F-RATIO SCORES OF THE TEO-BASED CATEGORY ARRANGED IN DESCENDING ORDER (FCI TASK - FEMALES)

	Sorted feature	F-ratio
	TEO CB3 Δ	52.50
	TEO CB2 Δ	51.30
	TEO CB4 Δ	48.75
	TEO CB1 Δ	46.45
Top 10% of features kept	TEO CB7 Δ	43.69
	TEO CB11	40.15
	TEO CB6 Δ	37.28
	TEO CB1 Δ-Δ	35.91
Top 20% of features kept	TEO CB12	34.20
	TEO CB10	31.70
	TEO CB7 Δ-Δ	30.06
	TEO CB3 Δ-Δ	28.92
	TEO CB5	28.77
Top 30% of features kept	TEO CB5 Δ	27.67
	TEO CB8 Δ	25.48
	TEO CB8 Δ-Δ	25.04
	TEO CB2 Δ-Δ	24.78
Top 40% of features kept	TEO CB9	21.67
	TEO CB9 Δ	20.08
	TEO CB13	17.97
	TEO CB4 Δ-Δ	17.12
	TEO CB10 Δ	16.13
Top 50% of features kept	TEO CB6	15.96
	TEO CB14	14.30
	TEO CB5 Δ-Δ	14.04
	TEO CB12 Δ	13.10
Top 60% of features kept	TEO CB15 Δ	12.19
	TEO CB14 Δ	11.72
	TEO CB11 Δ	11.55
	TEO CB9 Δ-Δ	11.43
	TEO CB6 Δ-Δ	9.20
Top 70% of features kept	TEO CB13 Δ	8.74
	TEO CB15 Δ-Δ	8.71
	TEO CB10 Δ-Δ	8.09
	TEO CB2	7.62
Top 80% of features kept	TEO CB15	6.26
	TEO CB14 Δ-Δ	6.13
	TEO CB13 Δ-Δ	6.08
	TEO CB11 Δ-Δ	4.54
	TEO CB12 Δ-Δ	4.16
Top 90% of features kept	TEO CB3	3.86
	TEO CB8	3.60
	TEO CB7	0.40
	TEO CB4	0.24
	TEO CB1	0.13

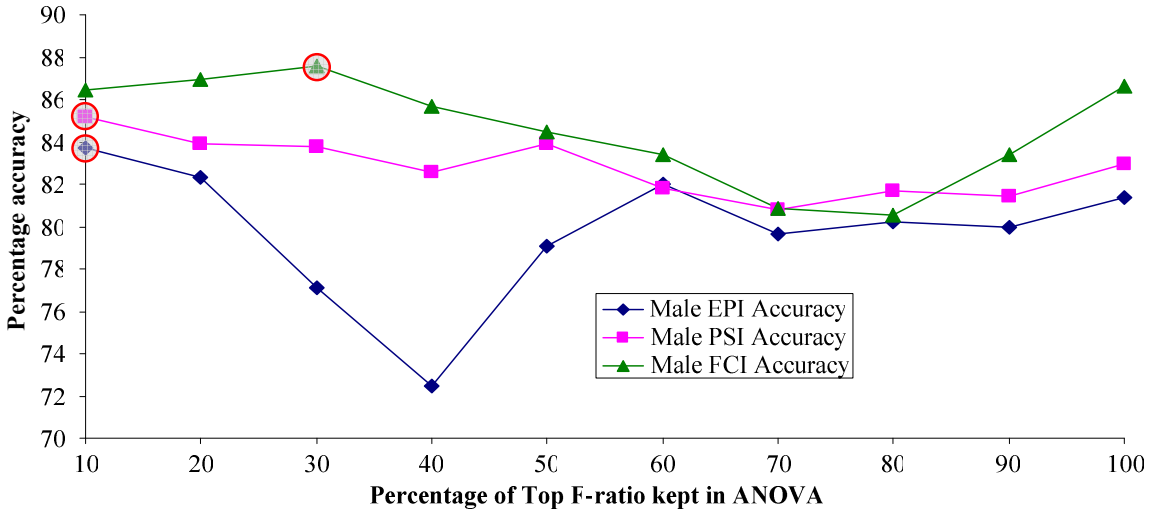

Figure 7.6: Classification accuracies of filter with 10%-90% of feature coefficients that were kept based on ranking of F-ratio from ANOVA (Male adolescents).

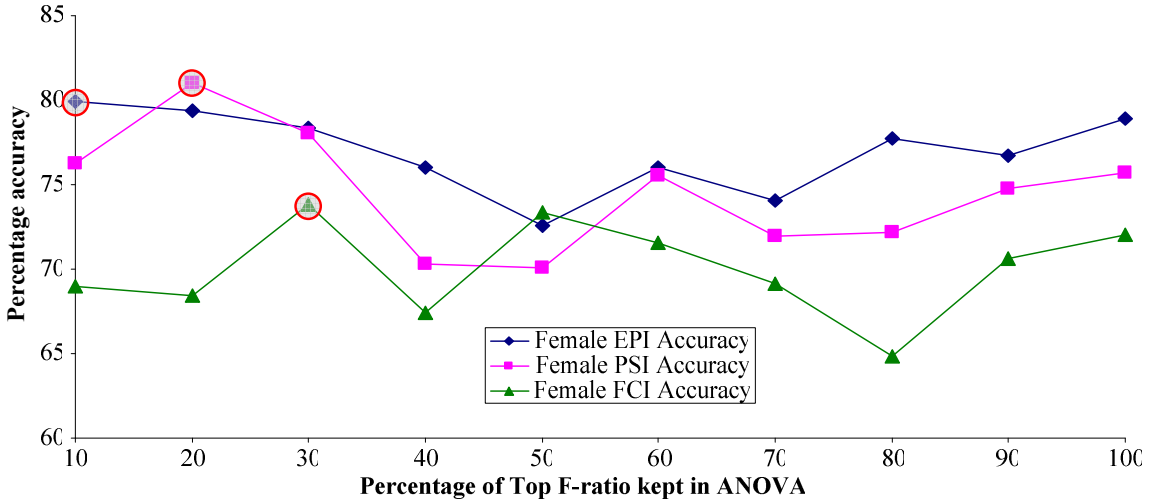

Figure 7.7: Classification accuracies of filter with 10%-90% of feature coefficients that were kept based on ranking of F-ratio from ANOVA (Female adolescents).

Table 7.18: CLASSIFICATION ACCURACIES AFTER FEATURE SELECTION FILTER APPROACH FOR THE TEO-BASED CATEGORY

Event Planning Interaction (EPI)					
Male			Female		
Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
84.45	82.96	83.71	81.46	78.41	79.93
Problem Solving Interaction (PSI)					
Male			Female		
Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
89.44	80.91	85.18	85.15	76.83	80.99
Family Consensus Interaction (FCI)					
Male			Female		
Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
80.83	94.37	87.60	74.58	73.08	73.83

7.13 Summary

The purpose behind this chapter was to investigate on the characterization on acoustic speech feature categories i.e., Teager energy operator (TEO)-based, Cepstral (C), Prosodic (P), Spectral (S) and Glottal (G) that are closely related to the physiological and perceptual components in the human speech production model for the detection of clinical depression in adolescents.

As a preliminary step, statistical analyses were carried out in revealing to what extend different acoustic feature categories were correlated with clinical depression. Since the acoustic features were collected during different family interaction tasks and our subjects represented different genders, these factors were also included in the statistical analysis. The multivariate analysis of variance (MANOVA) and the one-way analysis of variance (ANOVA) tested if the mean values of features for different tasks and genders were statistically significant between depressed and control groups. It was

found that only five out of the total of 14 acoustic features/interactions/gender combinations, listed in Table 7.1 had statistically insignificant ($p>0.05$) difference between the means. These combinations included: F0/EPI/male, jitter/EPI&FCI/male and centroids/PSI&FCI/male. As a result, the uncorrelated features/interactions/gender combinations were excluded from subsequent investigations, which aimed at determining features providing the strongest subject-based discrimination between speech of depressed and control adolescents.

In evaluating the classification performances, the classification decision was based on the total number of subjects correctly classified as depressed or control. The subject was classified as depressed if more than 50% of the utterances belonging to the subject were classified as depressed otherwise the subject was labeled as a control.

The following summarizes the key elements that were found in our experiments on the acoustic correlates of depression in speech of adolescents:

- Gender dependent modeling (males and females modeled separately) was more effective in discriminating features of the depressed and control classes compared to gender independent modeling (males and females modeled together).
- Test utterances with concatenated voiced segments of 1 min in duration gave the optimal classification results.
- In line with past research [84], the glottal feature category when added to other feature categories, increased the overall discrimination between speech of depressed and control classes.

- The TEO-based feature category outperformed all other feature categories or feature category combinations in all the interactions for both the genders.
- A feature selection filter approach method that is based on the ranking of ANOVA F-ratio scores further improved the classification accuracies on the TEO-based feature category in all the interactions for both the genders.

Chapter Eight:

DISCUSSION AND CONCLUSIONS

"Research is to see what everybody else has seen and to think what nobody else has thought"

– Albert Szent-Gyorgyi

8.1 Research Summary

Clinical depression is one of the most prevalent of all psychopathology disorders and if left untreated may lead to many negative consequences (i.e., suicidality, poor physical health, economic cost). Understanding the cause of depression has long been a complex and challenging task particularly in the field of psychology due to the many potential psychological variables. The advancement in technology have allowed psychologists to collaborate with different interdisciplinary fields (i.e., neuro-imaging techniques in cognitive neuroscience, image analysis in facial expression recognition and speech information analysis) in using computer-aided data processing techniques to provide a better understanding in the psychological factors relating to the development of clinical depression. The topic of this research focuses on developing objective acoustic measures in speech analysis for the diagnosis of depression in adolescents whereby it has been well documented that the voice of depressed individuals are slow, uniform, monotonous and expressionless with the person having the fear of expressing himself or herself [86].

Although there have been vast amount of literature pertinent to deriving acoustic correlates of depression from speech [5], [27], [33], [65], [88], the main motivation behind this research came from the few studies conducted by members of the research team from Vanderbilt University (France *et al.* (2000); Ozdas *et al*. (2004)) and Georgia Institute of Technology University (Moore *et al.* (2008)) that actually implemented these acoustic correlates as potential indicators in building a diagnostic tool that uses speech analysis in detecting depression [38], [84], [91] and suicidality [38], [91].

Therefore, inline with these studies, the key expected outcome of this research is to provide a speech-based depression detection system that could serve as a screening tool to assist mental health professionals in identifying clinically depressed persons. As this system is intended as the first stage of a diagnostic process and not a definitive identifier (i.e., detection via this type of system would normally be followed by full clinical evaluation of these screened as potentially depressed), the objective was to identify more depressed subjects (higher sensitivity) rather than to screen out negative cases (higher specificity).

Due to the significant differences between adult and adolescent speech [48], and given the fact that there has been a dramatic increase in the incidence of depressive symptoms and disorders in adolescents [70], this differs from other studies in that our research is specifically designed to present an initial attempt in investigating the acoustic parameters of speech as indicators of depression in adolescents (and not in adults).

The speech corpus (see *Chapter 5*) prepared for this study was obtained through a collaborative effort with the Oregon Research Institute (ORI) [108] that contained 139

adolescents (aged between 14-18 years old) with their respective parents participating in three different types of family interactive tasks:

1. Event planning interactions (EPI)
2. Problem solving interaction (PSI)
3. Family consensus interaction (FCI)

The duration of each interactive task was 20 minutes. The tasks were designed to create different types of interactional contexts between the adolescent and parents. Out of these 139 adolescents, 68 (19 males and 49 females) of them were diagnosed by psychologist from ORI to have major depressive disorders (MDD) and the rest of the 71 (27 males and 44 females) were diagnosed to be healthy, non-depressed (control) adolescents.

This study reports on an investigation of 14 acoustic features used as potential indicators of clinical depression that were grouped into five main feature categories extracted from adolescents' speech recorded during the three family interactional tasks. The proposed acoustic feature categories were formed based on the physiological and perceptual similarities of the speech production model (see *Section 2.6* for the schematic diagram of the speech production mechanism indicating the regions of interest in characterizing the five acoustic feature categories). The five main acoustic feature categories are listed below and the sub-categorical acoustic features grouped into these main categories can be found in *Section 6.3*.

- Teager energy operator (TEO)-based
- Cepstral (C)

- Prosodic (P)
- Spectral (S)
- Glottal (G)

Previous psychological studies reported significant differences in depressive symptoms between adolescent males and females [89]. Therefore, in experiments described in *Section 7.6*, the influence of gender differences in depression detection was first investigated using feature categories of TEO and C. The TEO-based and C feature categories were selected as our starting point since they have been effectively employed in speaker, language, and emotion content modeling [105], [123]. Experimental results indicated higher overall accuracies from all three interactions in feature categories of C (average of 1.7%), TEO (average of 5.3%) and their combinations of TEO+C (average of 4%) with gender dependent modeling (males and females modeled separately) than with gender independent modeling (males and females modeled together). This is consistent with those previous psychological studies that have suggested significant variations in depressive symptoms based on the gender [89]. It was also observed that the TEO-based feature category outperformed the feature category of C and their combinations of TEO+C in both modeling strategies.

In order to build an accurate screening system for clinically depressed subjects, it was important to determine how the decision timing of the test utterance length would affect the overall performances of the system. In *Section 7.7*, our experiments based on the TEO-based feature category indicated that utterances with 1 minute of speech content improved the overall classification accuracies in all of the three interactions.

In *Section 7.8,* experiments with the other feature categories of P, S and G yielded overall accuracy improvements for the male subjects when feature category G was combined with P (i.e., P+G) , S (i.e., S+G) or P+S (i.e., P+S+G) in all the interactions. For the female subjects, this trend was similar except for the feature categories of P+G combination in the PSI task, whereby there was a 1.8% average accuracy drop compared to feature category of P alone.

As stated by Moore (2003), it would be difficult to present a direct comparison in the works gathered by different studies related to the acoustic properties of speech as indicators of depression due to the different implementations in analysis methods used in extracting the acoustic features. For example, various studies use different forms of feature statistics (i.e., mean, variance, median, percentile, etc.) over frames of extracted features of an utterance or sentences. Furthermore, in our study no feature statistics were computed, instead only the short-term features (i.e., features that are extracted on a speech frame basis) were applied. However, we believe that although direct comparison cannot be made, the categorization of acoustic features into the grouping of feature categories should still provide a similar pattern in classification rates as long as these extracted acoustic features (be it short-term features or different feature statistics) still represented the same feature categories in characterizing the physiological and perceptual similarities of the speech production model.

Therefore, in searching for a relatively independent reference point that would allow us to verify our findings, feature categories similar to recent published research [84] were examined on our database in *Section 7.9*. Consistent with past research [84], implementation of both our proposed feature categories and feature categories from [84]

demonstrated that the critical role of the glottal feature category, which when added to the other stand-alone feature categories, increased the overall discrimination between speech of depressed and control classes. However, in our proposed categories for females, the increase in accuracy for the feature categories of P+G was not shown during the task of PSI. These findings could be due to the fact that the PSI task is the interaction that is most likely to elicit conflictual behavior, which in turn could contribute to an increase in pitch during angry and loud speech in these stressful scenarios. The rapid motion of the glottis caused by the increased in pitch does not always yield complete closure. Therefore, increased pitch could yield difficulties in obtaining reliable information about the changes in the glottal waveform. These difficulties are especially pronounced for females as they tend to exhibit higher pitch [83].

It was also observed that our final combinations of P+S+G (Table 7.3 and Table 7.4) gave higher classification results compared to the combinations of the proposed feature categories of P_o+V_o+G_o (Table 7.5 and Table 7.6) in [84] for both genders throughout all the interactions. One possible reason for the increase was that in the spectral (S) feature category, the acoustic sub-categorical feature of power spectral density (PSD) was included and it has been noted in past research that PSD provides a superior discrimination between the speech of control and depressed adults [38].

From experiments in *Section 7.10*, it was observed that by adding the TEO-based feature category to different combinations of the P, S and G features, the classification accuracy increased in all cases for the males and in some cases for the females. Most interestingly, the TEO-based features, when used on their own, clearly outperformed all other features and feature combinations. This pattern held for all three interactions across

both genders. It was seen for the males that the TEO-based feature category yielded correct classification scores of 81.36%, 82.96%, and 86.64% respectively. For the females, the TEO-based feature category yielded correct classification scores of 78.87%, 75.70%, and 72.01% respectively.

Experiments carried out to this point (*Section 7.4 – Section 7.10*) were based on the modeling and classification using the Gaussian mixture model (GMM). The GMM was employed in the modeling and classification of these five acoustic feature categories as it has been effectively used in speech information modeling tasks [12]. However, so as not to bias the classification results to just one classifier, the best performing feature category of TEO with GMM was compared with the modeling and classification of TEO with the support vector machine (SVM). Due to our large data set, an optimized parallel implementation of the SVM (OPSVM) as described in *Section 6.5.2* was implemented to reduce the computational effort. From the experiments in *Section 7.11,* the overall accuracies for the TEO-based category with OPSVM gave very similar results compared to the TEO-based category with GMM. However, the computational time required in training the models of our dataset in the OPSVM was still less efficient compared to the GMM. Therefore, we opted for the GMM for the rest of our experiments.

So far, it can be seen that a large number of acoustic features were evaluated in this study and it would seem practical to purse a data reduction approach (i.e., principle component analysis) with all the acoustic features combined. A data reduction strategy such as PCA, which would estimate latent components that capture the patterns of co-variation between the acoustic features examined, would have the advantage of reducing the number of variables to be studied. On the other hand, the PCA and related techniques

would also have the disadvantage that one is no longer examining observed variables, but rather factor scores that are unique to the data set collected in a particular study. As such, it is difficult to generalize these factors score to future research or applications. On balance, given the exploratory nature of this research, and the desire to identify features that can be directly utilized in future studies and devices, we believe that it is more appropriate to examine observed variables rather than to pursue a data reduction strategy, even though this does have the down side of multiplicity. On the contrary, we decided in a feature selection approach on the feature category of TEO which was the most effective feature category at discriminating speech of depressed and control adolescents.

The TEO-based feature category comprised of 45 feature coefficients. As described in *Section 7.12*, after a simple filter approach that selected the top feature coefficients based on the ranking of F-ratio scores obtained from one-way analysis of variance (ANOVA), the overall accuracies was slightly improved when compared to using the original number of feature coefficients in the TEO-based feature category for both the genders in all interactions. After applying feature selection on the feature category of TEO, the overall accuracies for the male adolescents were 83.71% (*Sensitivity = 84.45%*; *Specificity = 82.96%*), 85.18% (*Sensitivity = 89.44%*; *Specificity = 80.91%*) and 87.60% (*Sensitivity = 80.83%*; *Specificity = 94.37%*) in the tasks of EPI, PSI and FCI respectively.

For the female adolescents, the percentage overall accuracy was 79.93% (*Sensitivity = 81.46%*; *Specificity = 78.41%*), 80.99% (*Sensitivity = 85.15%*; *Specificity = 76.83%*) and 73.83% (*Sensitivity = 74.58%*; *Specificity = 73.08%*) in the tasks of EPI, PSI and FCI respectively.

Looking across the different interactions, it was also observed that although the overall accuracy and the specificity measures did not provide consistent results, there is a clear pattern within the sensitivity measure which shows that the PSI provides consistently higher results for both male and female subjects. This again can be attributed to the fact that the PSI evokes situations most likely to elicit conflicting behavior and therefore produces more pronounced changes in speech acoustics in identifying depression.

In summary, the classification accuracies in detecting clinical depression in adolescents' speech strongly depended on the gender and on the type of acoustic features. Inline with previous research [84], our study reinforces the importance of glottal features in the discrimination between speech of depressed and control adolescents. Finally, the non-linear approach of the TEO-based feature category shows the highest correlation with depression in the speech of both the male and female adolescents.

8.2 Interpretation of Major Findings – Why Do TEO-Based and Glottal Features Significantly Improve the Detection Accuracies of Clinically Depressed Subjects?

It was observed that the glottal features boosted the accuracy of discrimination between speech of depressed and control adolescents. TEO-based features also appear to be powerful discriminants of depression in speech. Both observations are maybe closely related to the physical impact of depression on the speech production processes through the vocal folds and vocal tract (tube extending from vocal folds to the lips). In order to explore this further, it is helpful to briefly discuss the main processes in speech production.

The classical source-filter theory (see *Section 2.5*) of voice production assumes that the air flow through the vocal folds (source) and the vocal tract (filter) is unidirectional. During phonation, the vocal folds vibrate. One vibration cycle includes the opening and closing phases in which the vocal folds are moving apart or together, respectively. The number of cycles per second determines the frequency of the vibration, which is subjectively perceived as pitch or objectively measured as the fundamental frequency (F0). The sound is then modulated by the vocal tract configuration and the resonant frequencies of the vocal tract are known as formants. However, recent advances in theoretical acoustics and computational modeling, and experiments in mechanical models, point out additional nonlinear and turbulent air flows which occur during phonation. In his original report, Teager [113] presented results of the intra-oral air velocity measurements made with an array of hot wire anemometers. The results indicated that the air flow is not purely laminar and a flow separation occurs causing an active sound generation in the mouth. Assuming that speech is an amplitude and frequency (AM-FM) modulated signal, the TEO parameter represents a measure of instantaneous energy calculated not only as a function of signal amplitude, but frequency as well [79], [92], [123]. This indicates that the TEO values contain information about spectral distribution of the signal energy and show sensitivity to the presence of additional harmonics and cross-harmonics in the speech signal [123].

The experimental studies of the vocal flow formation [10], [57], [58], [59], [106], [109], [113] on the other hand, provide strong evidence that the glottal air flow has a nonlinear character with a laminar flow component as well as additional turbulent components called vortices. In [57], [58], [59] two types of vortices were identified; each

occurring in a specific part of the vibration cycle, and at a certain location relative to the glottis. During the early opening phase of the vocal folds, when the glottis is convergent, supraglottal vortices occur above the vocal folds. During the latter part of the vocal fold closing, when the glottis is divergent, intraglottal vortices are formed between the vocal folds. The intraglottal vortices can alter the vibration of vocal folds, whereas, the supraglottal vortices provide additional sound sources when hitting hard surfaces of the vocal tract or interacting with each other. It was demonstrated in [58] that the level of symmetry in the vocal fold vibration has a strong effect on the glottal energy distribution across the frequency spectrum. It has been postulated that these additional sound sources [15], [56], [123] generate extra harmonics and cross-harmonics in speech.

As indicated in [123] the number of supraglottal vortices is likely to be related to the level of emotional stress. Moreover, the tension of laryngeal muscles responsible for the stiffness of the vocal folds (and hence, the vortices) is controlled by the sympathetic nervous system. Hence, it is likely that different patterns of the glottal wave formation reflect different emotional or mental states of a speaker and therefore contain important cues for the depression recognition in speech.

Fig. 8.1 shows the average normalized area of the autocorrelation envelope for all the speech frames in the TEO-based category in both the depressed (marked with “X”) and control class (marked with “O”). The normalized area measurement was plotted for all the critical bands in the TEO-based category. The normalized area details the strengths of the produced additional harmonics within the critical band, which further indicates the turbulent air flow occurring during the phonation process. This area parameter in TEO has also been documented to provide useful assessment in vocal fold pathology [46].

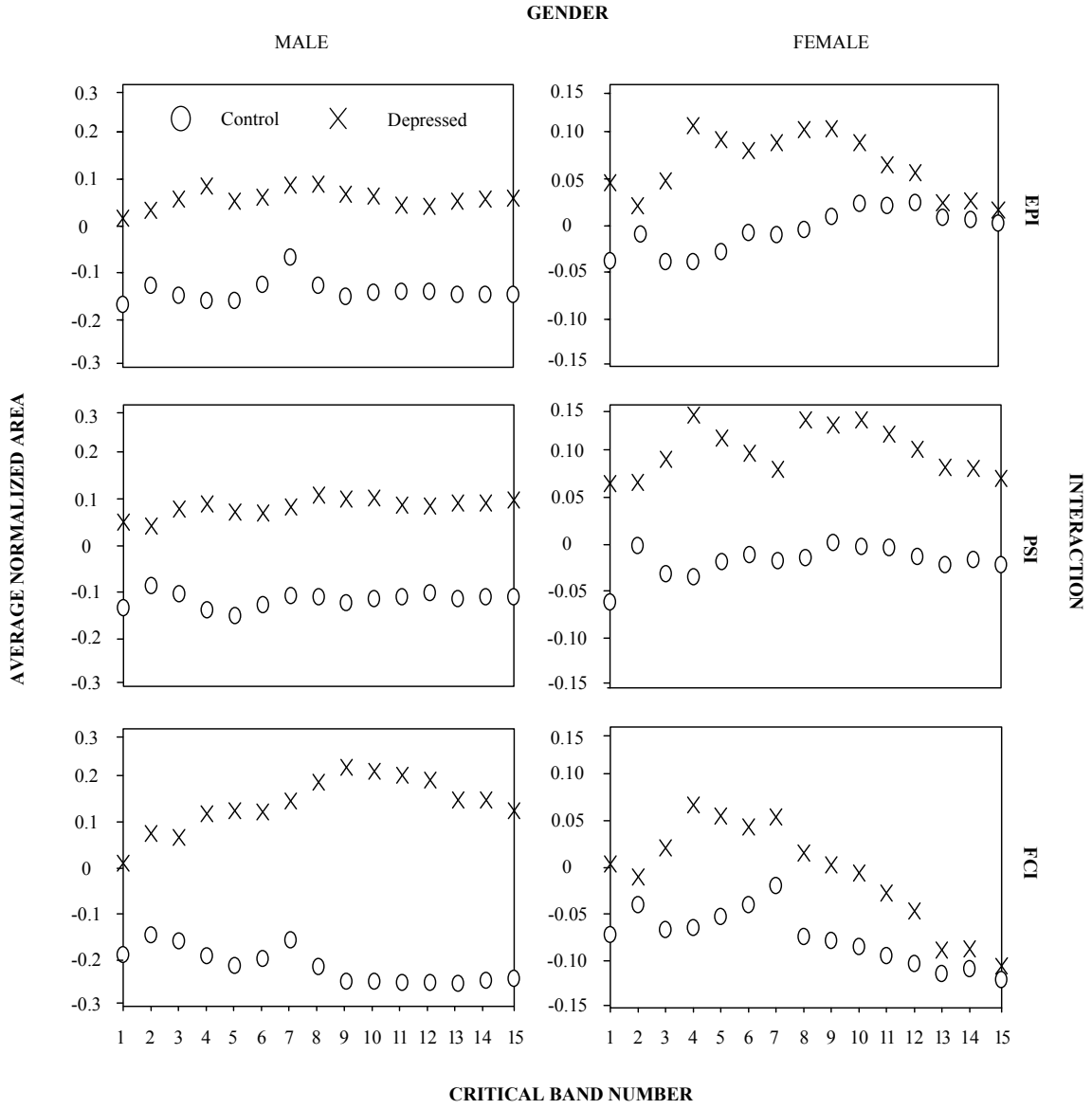

Figure 8.1: Average frames (25msec) normalized area under the autocorrelation envelope for the TEO-based feature category for each of the 15 critical bands in all adolescents within the depressed and control classes.

It is evident from Fig. 8.1 that the average normalized areas of the critical bands for the speeches of depressed participants are higher than the speeches of control participants. This pattern indicates that higher additional harmonics are generated in the depressed speech than in the control speech. Therefore, the result suggests that more

vortices appear in the airflow during the phonation process for the depressed adolescents than for the control adolescents. How the glottal feature, TEO-based feature, and dynamics of the air flow are linked to the regulatory psychophysiological processes occurring during depression remains to be investigated. Though not conclusive, recent studies [32], [52], have suggested that the speech production system show physical manifestations of the psychological difficulties of depressed persons (vocal folds and vocal tract). In such cases, patterns (laminar and vortices) of air flow in the speech production system of depressed participants differ from the airflow of control participants. For example, clinical depression may have a significant effect on vocal fold dysfunction. This could explain why the glottal features are effective in differentiating depressed from control speech. The TEO-based feature detects the presence of the extra harmonics and cross-harmonics generated by the vortices, making it an effective feature for discrimination between depressed and control speeches.

In addition, Table 8.1 and Table 8.2 shows the selection in the top percentage of feature coefficients kept within the TEO-based feature category based on the ranking of their ANOVA F-ratio scores that improved the overall accuracies when compared to using the original number of feature coefficients in the TEO-based feature category.

For the male adolescents, it was observed that the top three feature coefficients (highlighted in bold in Table 8.1) in each interaction contained the feature coefficients of TEO critical-band (CB) 13, TEO CB 14 and TEO CB 15. The frequency range from these bands was 3500Hz to 5500 Hz. This shows that the means of the TEO feature coefficients from those critical bands were likely to be more discriminative between the depressed and control classes. However, for the females no consistency in the top

percentage of feature coefficients in the TEO-based feature category appeared in all the three interactions. This also can be seen by the plots of Fig. 8.1 whereby the TEO critical band frequency ranges consistently shows that it is more discriminative at the higher frequencies for the male adolescents but for the female adolescents, this consistency is not shown in all the interactions.

In general, the plots of the area under the normalized autocorrelation envelope in Fig. 8.1 provided clearer distinction between the depressed and control classes for the male subjects than for the female subjects. This observation indicates that there is a higher variation in the number of additional harmonics between the depressed and control male subjects than between the depressed and control female subjects. Therefore, it appears that the effects of depression on the voice characteristics of male subjects are more profound than the effects on the voice characteristics of female subjects.

These observations are consistent with the results in *Section 7.10* showing that the addition of TEO-based features to the other types of features provided statistically significant (McNemar's test, $p<0.05$) improvement of the classification accuracies only in the case of male subjects. The observed small increase of the classification accuracies for the female subjects was found to be statistically insignificant. It is possible that these differences are related to the fact that there are clear differences between types of depression most frequently exhibited in males and females. One possible reason in why the TEO-based feature (which has proven to do well in stress classification [123]) performs better in males is that it has been documented that women tend to be more open in sharing and discussing problems with their family while men tend to be more reserved

and more aggressive or withdrawn in stressful situations [71]. However, further investigations are needed.

Table 8.1: LIST OF THE TOP FEATURE COEFFICIENTS FROM THE TEO-BASED CATEGORY BASED ON F-RATIO SCORES FROM ANOVA (MALES).

TEO Feature Category		
Interactions	Top % feature coefficients	Top feature coefficients
EPI	10%	**TEO CB15 (F=67.37), TEO CB13 (F=56.33), TEO CB14 (F=46.47),** TEO CB1 (F=40.65), TEO CB2 (F=31.76)
PSI	10%	**TEO CB15 (F=78.63), TEO CB13 (F=63.67), TEO CB14 (F=48.71),** TEO CB15 Δ-Δ (F=20.03), TEO CB1 (F=18.25)
FCI	30%	**TEO CB13 (F=310.28), TEO CB14 (F=274.25), TEO CB15 (F=274.20),** TEO CB11(F=268.24), TEO CB12 (F=265.38), TEO CB10 (F=227.41), TEO CB9 (F=224.05), TEO CB8 (F=155.08), TEO CB6 (F=118.99), TEO CB5 (F=118.75), TEO CB2 (F=109.70), TEO CB7 (74.28), TEO CB4 (F=54.24), TEO CB3 (F=50.28)

*F in brackets indicates F-ratio scores from ANOVA

Table 8.2: LIST OF THE TOP FEATURE COEFFICIENTS FROM THE TEO-BASED CATEGORY BASED ON F-RATIO SCORES FROM ANOVA (FEMALES).

TEO Feature Category		
Interactions	Top % feature coefficients	Top feature coefficients
EPI	10%	TEO CB5 (F=50.83), TEO CB9 (F=32.26), TEO CB6 (F=27.24), TEO CB10 (F=20.36), TEO CB8 (F=16.90)
PSI	10%	TEO CB10 (F=137.61), TEO CB11 (F=136.80), TEO CB12 (F=135.08), TEO CB13 (F=128.56), TEO CB14 (F=113.73), TEO CB9 (F=99.09), TEO CB15 (F=91.93), TEO CB5 (F=80.73), TEO CB8 (F=63.15)
FCI	30%	TEO CB3 Δ (F=52.50), TEO CB2 Δ (F=51.30), TEO CB4 Δ (F=48.75), TEO CB1 Δ (F=46.45), TEO CB1 Δ (F=46.45), TEO CB7 Δ (F=43.69), TEO CB11 (F=40.15), TEO CB6 Δ (F=37.28), TEO CB1 Δ-Δ (F=35.91), TEO CB12 (F=34.20), TEO CB10 (F=31.70), TEO CB7 Δ-Δ (F=30.06), TEO CB3 Δ-Δ (F=28.92), TEO CB5 (F=28.77),TEO CB5 Δ (F=27.67)

*F in brackets indicates F-ratio scores from ANOVA

8.3 Future Direction

Although our study has shown that clinical depression can be detected in adolescents using naturalistic speech samples, clinical depression detection still remains a challenging task due to the large number of potential genetic, psychological, social, cultural and environmental factors that contribute to the development of this condition [108]. In addition, a potential limitation of this study was that the speech may contain some features that are specific to the family context, or that are primarily elicited by parental behavior. Therefore, in future studies, we plan to verify our findings on a different database and also investigate different non-linear approaches for modeling depressive speech characteristics in order to improve discrimination between depressed and control subjects. Further analyses would also be carried out in comparing the different frequency ranges in the critical bands of the TEO-based feature (i.e., low frequency bands, medium frequency bands and high frequency bands).

Another extension of this research that we are heading into is focused on the adolescents' speech samples used in this study that contained the ten labeled emotions manually annotated by trained observers based on the Living in Family Environments system (LIFE) manual. Based on these emotions labeled by the trained observers, studies in the future will attempt to develop an objective approach aimed at recognizing emotions automatically from utterances containing spontaneous speech of the depressed and control adolescents. This is set out to investigate the classification rates in recognizing emotions in adolescents' speech and how they differ between the depressed and control group.

APPENDIX A:

The fundamental rules of probability theory are given as:

$$\textbf{Sum rule :}\ p(X) = \sum_{Y} p(X,Y) \tag{A.1}$$

where *p(X, Y)* denotes the joint probability distribution of *X and Y*.

$$\textbf{Product rule :}\ p(X,Y) = p(Y \mid X)\,p(X) \tag{A.2}$$

where $p(Y \mid X)$ denotes the conditional probability distribution of *Y given X* and *p(X)* is a marginal probability.

From the product rule in Eq. (A.2) and based on the symmetry property given as:

$$\textbf{Symmetry property :}\ p(X,Y) = p(Y,X) \tag{A.3}$$

We can form a relationship between the conditional probabilities called the *Bayes' theorem* and is written in the following form:

$$\begin{aligned} p(Y,X) &= p(Y)\,p(X \mid Y) \\ p(X,Y) &= p(X)\,p(Y \mid X) \\ p(Y)\,p(X \mid Y) &= p(X)\,p(Y \mid X) \\ p(Y \mid X) &= \frac{p(X \mid Y)\,p(Y)}{p(X)} \end{aligned} \tag{A.4}$$

Table A1: DEMOGRAPHIC DATA OF PARTICIPANTS IN OREGON DATABASE

Demographic Category	Depressed	Healthy	Test Statistic
	(n=75)	(n=77)	
Gender			
Male	23	29	χ^2=0.83, ns
Female	52	48	
Age			
Mean (SD)	16.22 (1.11)	16.14 (1.05)	t=0.44, ns
Family Structure			
Dual parent family	47	60	χ^2=4.24, (p<.05)
Single parent family/other	28	17	
Income			
Median	$37,500	$42,500	χ^2=6.37,(p<.05)
Ethnicity			
Caucasian	49	57	χ^2=0.46, ns
African American	2	2	
Asian	0	1	
Native American	1	0	
More than one race	18	15	
Unknown	5	2	

Table A2: LIST OF TOPICS GIVEN IN THE DISCUSSIONS OF THE FAMILY INTERACTIONS

INTERACTIONS	SUB-TOPIC 1 (Duration: 10 mins)	SUB-TOPIC 2 (Duration: 10 mins)
Event Planning Interactions (EPI)	Planning a family vacation trip: • Where are you going to go? • Who are you going to see? • Who should come along with you? • What will you do while you are there?	The good times you had together as a family: • Think what made the time the most enjoyable. • Who was there? • What did you do? • What was funny? • How did you feel?
Problem Solving Interactions (PSI)	Discuss a topic from questionnaire that both parties (teenager & parents) disagree on. For example, "child cleaning up the room": • Try to overcome the solution. • Try to think of any road block that might come into the way of the solution and try to solve it	Discuss another topic from the questionnaire that both parties (teenager & parents) disagree on: • Try to overcome the solution. • Try to think of any road block that might come into the way of the solution and try to solve it.
Family Consensus Interactions (FCI)	Participating in the writing of a book chapter that reflects the shared perspective of the teenager and parents: • Write about the best and worst years of your child's life. • Choose a specific best and worst year and identify what made these years especially good or bad. • Write on the impact that these years have had on your family and what kind of person has your child become today.	Participating in a book chapter (Same as subtopic 1 but with a different theme): • Write about the hardest and rewarding things on parenting your child. • Write about how parenting your child has influenced the person they have become today.

Table A3: CRITERIA TO CODE DIFFERENT EMOTIONS FROM LIFE MANUAL

Code	Emotion	Description
0	Contempt	Communicates a lack of respect for the recipient and suggests that the participant feels somewhat superior to the recipient. The voice tone is sarcastic and disgusted, there are sneering and snorts.
1	Anger	Communicates displeasure. An angry person sounds like he/she is "fed-up". Voice is lowered or raised beyond the limits of normal tone; words usage is abrupt, with one word or syllable being more strongly stressed; short clipped speech; irritation, annoyance, frustration evidenced by changes in the rhythm of speech and the way certain words are stressed.
2	Anxious	Communicates anxiety, nervousness, worry, fear or embarrassment. The voice can be described as tense, fearful, concerned, startled, shocked, hysterical, afraid, uneasy, and worried. There could be speech difficulty, stuttering, or slips of the tongue. Mouth and lips may tremble.
3	Dysphoric	Communicates sadness and depression. Persons that are depressed may appear detached from the ongoing activity, tend to speak slowly, and use a low voice tone. The voice is sad, glum, distressed, discontented, withdrawn, discouraged, despondent, joyless, gloomy or melancholic. There is a low voice tone and slow pace of speech, often with sighing and yawing.
4	Pleasant	Reflects interested and engaged qualities; it is the bridge between neutral and happy, or neutral and carrying. It reflects a change in energy from passive (neutral) to active listening; participant becomes focused on the speaker, is genuinely engaged in the conversation seeking clarification, additional information or elaboration about something that another had just said.
5	Neutral	Occurs when the participant is relatively even-tempered, composed, instructional or reasonable. Neutral is described as a dividing line between negative and positive affects and is generally non-emotional. A neutral voice tone is even, relaxed, without marked stress on individual syllables. In situations where the person's behavior contained a mixture of neutral and any other affect category, the other category was coded. The voice quality is even-tempered, without a trace of dejection, sternness, or sulkiness.
6	Happy	Reflects mood that can be described as glad, silly, playful, funny, hopeful, thrilled, pleased, excited exuberant, cheerful, delighted or enthusiastic. The speech is full of laughter, giggling or smiling. The voice tone is a high pitched or sing-song but not whining. Speech is faster or louder than usual but not angry.
7	Caring	Reflects affection, warmth, support, and liking of the recipient. Participant communicates in a soothing or empathetic manner. The voice tone is soft, warm, affectionate, smiling and soothing.
8	Whine	Expressing cranky, frustrated, agitated mood with "poor me" attitude, often with high pitched, nasal, sing-song voice tone.
9	Belligerence	Provocative, argumentative, contentious or combative affect; when displaying this affect the participant appears to be wanting to start a fight or keep one going. Speech cues for this affect include a rising inflection at the end of end of a challenging questions i.e., "So?", "What are we going to do about it?

.

REFERENCES

[1] ABRAMSON, L. Y., METALSKY, G. I., and ALLOY, L. B., "Hopelessness depression: A theory-based subtype of depression," *Psychological Review*, vol. 96, pp. 358-372, Apr. 1989.

[2] AIRAS M., "TKK Aparat: An environment for voice inverse filtering and parameterization," *Logopedics Phoniatrics Vocology*, vol. 33, pp. 49-64, 2008.

[3] ALKU P., "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication,* vol. 11, pp. 109-118, 1992.

[4] ALKU, P., STRIK, H., and VILKMAN, E., "Parabolic spectral parameter - A new method for quantification of the glottal flow," *Speech Communication,* vol. 22, no. 1, pp. 67-79, 1997.

[5] ALPERT, M., POUGET, E. R., and SILVA, R. R., "Reflections of depression in acoustic measures of the patient's speech," *Journal of Affective Disorders,* vol. 66, pp. 59-69, 2001.

[6] American Psychiatric Association, *Diagnostic and Statistical. Manual of Mental Disorders*, 4th Ed. American Psychiatric Association, Washington, DC., 1994.

[7] ARISTOTLE, *Treatise on rhetoric*. T. Buckley, Trans. Prometheus. Amherst, NY, 1995.

[8] BACHOROWSKI, J. -A. and OWREN, M. J., "Sounds of emotion," *Annals of the New York Academy of Sciences,* vol. 1000, pp. 244-265, 2003.

[9] BARLOW, D. H., "Unraveling the mysteries of anxiety and its disorders from the perspective of emotion theory," *American Psychologist,* vol. 55, pp. 1247-1263, 2000.

[10] BARNEY, A., SHADLE, C. H., and DAVIES, P., "Fluid flow in a dynamic mechanical model of the vocal folds and tract. I. Measurements and theory," *Journal of the Acoustical Society of America,* vol. 105, no. 1, pp. 444-455, 1999.

[11] BATLINER, A., BURKHARDT, F., VAN BALLEGOOY, M., and NÖTH, E., "A taxonomy of applications that utilize emotional awareness," in *Proceedings of the 1st International Language Technologies Conference (IS-LTC '06)*, pp. 246–250, Ljubljana, Slovenia, 2006.

[12] BISHOP C. M., *Pattern recognition and machine learning*. New York: Springer, 2006.

[13] BOSER, B. E., GUYON, I. M., and VAPNIK, V. N., "A training algorithm for optimal margin classifiers," *Proceedings of the fifth annual workshop on Computational learning theory*, New York, NY, USA: ACM, pp. 144-152, 1992.

[14] CAINE, R. N. and CAINE, G., *Making connections: teaching and the human brain*. Menlo Park, Calif.: Addison-Wesley Pub. Co., 1994.

[15] CAIRNS, D. A., and HANSEN, J. H. L., "Nonlinear-analysis and classification of speech under stressed conditions," *Journal of the Acoustical Society of America,* vol. 96, no. 6, pp. 3392-3400, 1994.

[16] CANNON, W. B., *Bodily Changes in Pain, Hunger, Fear and Rage: An Account of Recent Researches Into the Function of Emotional Excitement*. Nabu Press, 2010.

[17] CHANG C. C. and LIN C. J., "LIBSVM: a library for support vector machines," 2001 [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[18] CHILDERS D. G., *Speech processing and synthesis toolboxes*. New York; Chichester: Wiley, 2000.

[19] CHORIN A. J. and MARSDEN J. E., *A mathematical introduction to fluid mechanics*, 3rd Ed. New York: Springer-Verlag, 1993.

[20] CICERO, M. T, *De Oratore*. New York: Oxford University Press, 2001.

[21] COHEN, S., KESSLER, R. C., and UNDERWOOD GORDON, L., Strategies for measuring stress in studies of psychiatric and physical disorders. In: Cohen, S., Kessler, R. C., and Underwood Gordon, L., Editors, *Measuring Stress: A Guide for Health and Social Scientists*, Oxford University Press, New York, NY, pp. 3–26, 1995.

[22] COLLOBERT R., BENGIO S., and BENGIO Y., "A parallel mixture of SVMs for very large scale problems," *Neural Computation,* vol. 14, no. 5, pp. 1105-1114, 2002.

[23] COWAN M., "Pitch and intensity characteristics of stage speech," *Arch Speech*, Suppl. to December issue, 1936.

[24] COWIE, R. and CORNELIUS, R. R., "Describing the emotional states that are expressed in speech," *Speech Communication,* vol. 40, pp. 5-32, 2003.

[25] COWIE, R., DOUGLAS-COWIE, E., TSAPATSOULIS, N., VOTSIS, G., KOLLIAS, S., FELLENZ, W., and TAYLOR, J. G., "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, pp. 32-80, 2001.

[26] CRISTIANINI, N. and SHAWE-TAYLOR, J., *An introduction to support vector machines: and other kernel-based learning methods*. Cambridge; New York: Cambridge University Press, 2000.

[27] DARBY, J. K., and HOLLIEN, H., "Vocal and speech patterns of depressive patients," *Folia Phoniatrica,* vol. 29, no. 4, pp. 279-291, 1977.

[28] DARWIN, C., *The expression of the emotions in man and animals*, 3rd Ed. Harper Collins. London. (US edit.: Oxford University Press. New York.), 1998.

[29] DAVITZ J. R, *The Communication of Emotional Meaning*. New York: McGraw-Hill, 1964.

[30] DELLER J. R., PROAKIS, J. G., and HANSEN, J. H. L., *Discrete time processing of speech signals*. Upper Saddle River, N.J: Prentice Hall PTR, 1999.

[31] DENES P. and PINSON E., *The speech chain: The Physics and Biology of Spoken Language*. 2nd Ed. New York: W.H. Freeman, Company, 1993.

[32] DIETRICH, M., ABBOTT, K. V., SCHMIDT, J. G., and CLARK, A. R., "The frequency of perceived stress, anxiety, and depression in patients with common pathologies affecting voice," *Journal of Voice*, vol. 22, no. 4., pp. 472-488, July 2008.

[33] ELLGRING, H. and SCHERER, K. R., "Vocal indicators of mood change in depression," *Journal of Nonverbal Behavior,* vol. 20, pp. 83-110, 1996.

[34] FAIRBANKS, G. and PRONOVOST, W., "An experimental study of the pitch characteristics of the voice during the expression of emotion," *Speech Monographs*, vol. 6, pp. 87-104, 1939.

[35] FIELD, A. P., *Discovering statistics using SPSS: (and sex, drugs and rock 'n' roll)*. 2nd Ed., London: Sage, 2005.

[36] FLANAGAN J. L., *Speech analysis synthesis and perception.* New York: Springer-Verlag, 1972.

[37] FRANCE, D. J., "Acoustical properties of Speech as Indicators of Depression and Suicidal Risk," Ph.D. Thesis, Vanderbilt University, Tennessee, 1997.

[38] FRANCE, D. J., SHIAVI, R. G., SILVERMAN, S., SILVERMAN, M., and WILKES, D. M., "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Transactions on Biomedical Engineering*, vol. 47, pp. 829-837, 2000.

[39] GABOR, D., "Theory of communication," *IEE J.*, London, vol. 93, pp. 429-457, 1946.

[40] GANCHEV, T., FAKOTAKIS, N., and KOKKINAKIS, G. "Comparative evaluation of various MFCC implementations on the speaker verification task," *in Proc. of the SPECOM-2005*, 2005, pp. 191-194.

[41] GELLHORN, E. and LOUFBOURROW, G. N., *Emotions and emotional disorders*. New York: Hoeber, 1963.

[42] GELLHORN, E., "Motion and emotion: The role of proprioception in the physiology and pathology of the emotions," *Psychological Review*, 71, pp. 457–472, 1964.

[43] GOTLIB, I. H. and HAMMEN, C. L., *Handbook of depression*, 2nd Ed. New York: Guilford Press, 2009.

[44] GUYON I., GUNN, S., NIKRAVESH M., and ZADEH, L., *Feature extraction: foundations and applications*. Physica-Verlag: Springer, 2006.

[45] HANSEN, J. H. L. and BOU-GHAZALE, S., "Getting started with SUSAS: A speech under simulated and actual stress database," *EUROSPEECH-97: Inter. Conf. On Speech Communication and Technology*, vol. 4, pp. 1743-1746, Rhodes, Greece, Sept. 1997.

[46] HANSEN, J. H. L., GAVIDIA-CEBALLOS, L., and KAISER, J. F., "A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment," *IEEE Transactions on Biomedical Engineering,* vol. 45, no. 3, pp. 300-313, 1998.

[47] HANSEN, J. H. L., SWAIL, C., SOUTH, A. J., MOORE, R. K., STEENEKEN, H., CUPPLES, E. J., ANDERSON, T., VLOEBERGHS, C. R. A., TRANCOSO, I., and VERLINDE, P., *The Impact of Speech Under Stress on Military Speech Technology*: NATO Research & Technology Organization RTO-TR-10, vol. AC/323(IST)TP/5 IST/TG-01, Mar. 2000.

[48] HOLLIEN, H., GREEN, R., and MASSEY, K., "Longitudinal research on adolescent voice change in males," *Journal of the Acoustical Society of America,* vol. 96, pp. 2646-2654, 1994.

[49] HOPS, H., BIGLAN, A., LONGORIA, N., TOLMAN, A., ARTHUR, J., "Living in family environments (LIFE) coding system: Reference manual for coders," *Oregon Research Institute*, Eugene, OR, Unpublished manuscript, 2003.

[50] HOPS, H., DAVIS, B., and LONGORIA, N, "Methodological issues in direct observation-illustrations with the living in familial environments (LIFE) coding system," *Journal of Clinical Child Psychology,* vol. 24, no. 2, pp. 193-203, 1995.

[51] HU, H.-T., "An improved source model for a linear prediction speech synthesizer," Ph.D. Thesis, University of Florida, Gainesville, 1993.

[52] HUSEIN, O. F., HUSEIN, T. N., GARDNER, R., CHIANG, T., LARSON, D. G., OBERT, K., THOMPSON, J., TRUDEAU, M. D., DELL, D. M., and FORREST, L. A., "Formal psychological testing in patients with paradoxical vocal folds dysfunction", *Journal of Laryngoscope*, vol. 118, pp 740-747, April 2008.

[53] JAMES, W., "What is emotion?" *Mind*, vol. 9, pp. 188–205, 1884.

[54] JANG R., "Audio Processing Toolbox," 1996 [Online]. Available: http://neural.cs.nthu.edu.tw/jang/

[55] KAISER J. F., "On a simple algorithm to calculate the `energy' of a signal," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing*, pp. 381-384, 1990.

[56] KAISER, J. F., "Some observations on vocal tract operation from a fluid flow point of view," *Vocal Folds Physiology: Biomechanics, Acoustics and Phonatory Control*, 1983.

[57] KHOSLA, S., MURUGAPPAN, S., and GUTMARK, E., "What can vortices tell us about vocal fold vibration and voice production," *Current Opinion in Otolaryngology & Head and Neck Surgery,* vol. 16, no. 3, pp. 183-187, 2008.

[58] KHOSLA, S., MURUGAPPAN, S., PANIELLO, R., YING, J., and GUTMARK, E., "Role of vortices in voice production: Normal versus asymmetric tension," *Laryngoscope,* vol. 119, no. 1, pp. 216-221, 2009.

[59] KHOSLA, S., MURUGUPPAN, S., GUTMARK, E., and SCHERER, R., "Vortical flow field during phonation in an excised canine larynx model," *Annals of Otology Rhinology and Laryngology,* vol. 116, no. 3, pp. 217-228, 2007.

[60] KIM, W. J., "The American academy of child and adolescent psychiatry task force on workforce needs: child and adolescent psychiatry workforce: a critical

shortage and national challenge," *Academic Psychiatry,* vol. 27, pp. 277-282, 2003.

[61] KLEIN, D. N., LEWINSOHN, P. M., SEELEY, J. R., ROHDE, P., "A family study of major depressive disorder in a community sample of adolescents," *Archives of General Psychiatry,* vol. 58, no. 1, pp. 13-20, 2001.

[62] KLERMAN, G. L., "The current age of youthful melancholia - evidence for increase in depression among adolescents and young-adults," *British Journal of Psychiatry,* vol. 152, pp. 4-14, 1988.

[63] KOTLYAR, G. and MOZOROV, V., "Acoustic correlates of the emotional content of vocalized speech," *J. Acoust. Academy of Sciences of the USSR*, vol. 22, pp. 208-211, 1976.

[64] KRING, A. M., & WERNER, K. H., *Emotion regulation and psychopathology*. In P. Philippot & R. S. Feldman (Eds.), The regulation of emotion (pp. 359–385). Hove, UK: Psychology Press, 2004.

[65] KUNY, S. and STASSEN, H. H., "Speaking behavior and voice sound characteristics in depressive patients during recovery," *Journal of Psychiatric Research,* vol. 27, pp. 289-307, 1993.

[66] KURODA, I., FUJIWARA, O., OKAMURA, N., and UTSUKI, N., "Method for determining pilot stress through analysis of voice communication," *Aviation Space and Environmental Medicine,* vol. 47, no. 5, pp. 528-533, 1976.

[67] LADD, D., SILVERMAN, K., TOLKMITT, F., BERGMANN, G., and SCHERER, K. R., "Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect," *Journal of the Acoustical Society of America*, vol. 78, pp. 435-444, 1985.

[68] LANGLIEB, A. M. and DEPAULO, J. R. J., "Etiology of Depression and Implications on Work Environment", *Journal of Occupational & Environmental Medicine*, vol. 50, no. 4, pp. 391-395, April 2008.

[69] LEE, W.-S., ROH, Y.-W., KIM, D.-J., KIM, J.-H., and HONG, K.-S., "Speech Emotion Recognition Using Spectral Entropy," vol. 5315, pp. 45-54, 2008.

[70] LEWINSOHN, P. M., ROHDE, P., and SEELEY, J. R., "Major depressive disorder in older adolescents: Prevalence, risk factors, and clinical implications," *Clinical Psychology Review,* vol. 18, pp. 765-794, 1998.

[71] LIGHTDALE, J. R. and PRENTICE, D. A., "Rethinking sex differences in aggression: aggressive behavior in the absence of social roles," *Personality and Social Psychology Bulletin*, vol. 20, pp. 34-44, 1994.

[72] LINDSLEY, D. B., *Emotion. In S. S. Stevens (ed.) Handbook of Experimental Psychology*, New York: John Wiley & Sons, pp. 473–516, 1951.

[73] LINDSLEY, D. B., *Emotions and the electroencephalogram. In M. L. Reymert (ed.) Feelings and Emotions: The Mooseheart Symposium*. New York: McGraw-Hill, 1950.

[74] LINDSLEY, D. B., *Psychophysiology and emotion. In M. R. Jones (ed.) Nebraska Symposium on Motivation*. Lincoln, NE: University of Nebraska Press, pp. 44–105, 1957.

[75] LINDSLEY, D. B., *The role of nonspecific reticulothalamocortical systems in emotion. In P. Black (ed.) Physiological Correlates of Emotion*. New York: Academic Press, 1970.

[76] LOW, L.-S. A., MADDAGE, N. C., LECH, M., SHEEBER, L., and ALLEN, N., "Content based clinical depression detection in adolescents," in *Proc. European Signal Processing Conf.*, pp. 2362-2365, 2009.

[77] LOW, L.-S. A., MADDAGE, N. C., LECH, M., SHEEBER, L., and ALLEN, N., "Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents," *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 5154-5157, 2010.

[78] MADDAGE N. C., Content based music structure analysis, Ph.D. Thesis, School of Computing, National University of Singapore, Singapore, 2006.

[79] MARAGOS, P., QUATIERI, T., and KAISER. J. F., "On amplitude and frequency demodulation using energy operators," *IEEE Transactions on Signal Processing*, vol. 41, pp. 1532-1550, 1993.

[80] MITCHELL, A., LECH, M., KOKOTOFF, D.M., and WATERHOUSE, R., "Search for high performance direct contact stacked patches using optimization," *IEEE Transactions on Antennas and Propagation,* vol. 51, no.2, pp. 249-255, 2003.

[81] MONROE, S. M. and SIMONS, A. D., "Diathesis-stress theories in the context of life stress research: Implications for the depressive disorders," *Psychological Bulletin*, vol. 110, pp. 406-425, Nov. 1991.

[82] MOORE, E. "Evaluating objective feature statistics of speech as indicators of vocal affect and depression," Ph.D. Thesis, Georgia Institute of Technology University, Georgia, 2003.

[83] MOORE, E., and CLEMENTS, M., "Algorithm for automatic glottal waveform estimation without the reliance on precise glottal closure information," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing*, vol.1, pp. 101-104, 2004.

[84] MOORE, E., CLEMENTS, M. A., PEIFER, J. W., and WEISSER, L., "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," *IEEE Transactions on Biomedical Engineering,* vol. 55, pp. 96-107, 2008.

[85] MORAN, R.J., REILLY, R.B., DE CHAZAL, P., and LACY, P.D., "Telephony-based voice pathology assessment using automated speech analysis," *IEEE Transactions on Biomedical Engineering,* vol. 53, pp. 468-477, 2006.

[86] MOSES, P., *The voice of neurosis*. New York: Grune & Stratton, 1954.

[87] N.H.M.R.C, "Depression in young people: a guide for mental health professionals," *National Health and Medical Research Council*, Canberra, Australia, 1997.

[88] NILSONNE, A., "Acoustic analysis of speech variables during depression and after improvement," *Acta psychiatrica Scandinavica,* vol. 76, pp. 235-45, 1987.

[89] NOLENHOEKSEMA, S. and GIRGUS, J. S., "The emergence of gender differences in depression during adolescence," *Psychological Bulletin,* vol. 115, pp. 424-443, 1994.

[90] OZDAS, A., "Analysis of paralinguistic properties of speech for near-term suicidal risk assessment," Ph.D. Thesis, Vanderbilt University, Tennessee, 2001.

[91] OZDAS, A., SHIAVI, R. G., SILVERMAN, S. E., SILVERMAN, M. K., and WILKES, D. M., "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," *IEEE Transactions on Biomedical Engineering,* vol. 51, pp. 1530-1540, 2004.

[92] PATIL, H. A. and BASU, T. K., "Identifying perceptually similar languages using Teager energy based cepstrum", *Engineering Letters*, 16:1, EL_16_1_22, 2009.

[93] PRESS, W. H., FLANNERY, B. P., TEUKOLSKY, A., and VETTERLING, W. T., *Numerical recipes in C: the art of scientific computing*, 2nd Ed. Cambridge; New York: Cambridge University Press, 1992.

[94] PULAKKA H., "Analysis of human voice production using inverse filtering, high-speed imaging, and electroglottography," Master's thesis, Helsinki University of Technology, Espoo, Finland, 2005.

[95] RABINER, L. R. and SCHAFER, R. W., *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[96] ROSENBERG, A. E., "Effect of glottal pulse shape on the quality of natural vowels," *Journal of the Acoustical Society of America*, vol. 49, pp. 583–590, 1971.

[97] SALISBURY, D. F., "Researchers measure distinct characteristics in speech of individuals at high risk of suicide," *Exploration the Online Research Journal of Vanderbilt University*, October, 19, 2000.

[98] SCHERER, K. R. and ZEI, B., "Vocal indicators of affective-disorders," *Psychotherapy and Psychosomatics,* vol. 49, pp. 179-186, 1988.

[99] SCHERER, K. R., "Expression of emotion in voice and music," *Journal of Voice,* vol. 9, pp. 235-248, 1995.

[100] SCHERER, K. R., "Non-linguistic vocal indicators of emotion and psychopathology," in *Emotions in personality and psychopathology*, C.E. Izard ed New York: Plenum Press, 1979, pp. 495-529.

[101] SCHERER, K. R., "Speech and emotional states," in *Speech evaluation in psychiatry*, J. Darby ed New York: Grune & Stratton, pp. 189-220, 1981.

[102] SCHERER, K. R., "Vocal affect expression: A review and a model for future research," *Psychological Bulletin,* vol. 99, pp. 143-165, 1986.

[103] SCHERER, K. R., "Vocal correlates of emotional arousal and affective disturbance," in *Handbook of Psychophysiology: Emotion and social behavior*, H. Wagner, Ed London: Wiley, 1989, pp. 165-197.

[104] SCHERER, K. R., "Vocal indicators of stress," *Speech evaluation in psychiatry*, New York: Grune & Stratton, pp. 171-187, 1981.

[105] SCHULLER B., BATLINER A., SEPPI D., STEIDL, S., VOGT, T., WAGNER, J., DEVILLERS, L., VIDRASCU, L., AMIR, N., KESSOUS, L., and AHARONSON, V., "The relevance of feature type for the automatic classification of emotional user

states: Low level descriptors and functionals," in *Proc. Interspeech*, pp. 2253–2256, 2007.

[106] SHADLE, C. H., BARNEY, A., and DAVIES, P., "Fluid flow in a dynamic mechanical model of the vocal folds and tract. II. Implications for speech production studies," *Journal of the Acoustical Society of America,* vol. 105, no. 1, pp. 456-466, 1999.

[107] SHATKIN, J. P. and KOPLEWICZ, H. S., "The child and adolescent mental health studies (CAMS) minor at New York university," *Academic Psychiatry,* vol. 32, pp. 438-445, 2008.

[108] SHEEBER, L. B, ALLEN, N. B, LEVE, C., DAVIS, B., SHORTT, J. W., and KATZ, L. F., "Dynamics of affective experience and behavior in depressed adolescents," *Journal of Child Psychology and Psychiatry,* vol. 50, pp. 1419-1427, 2009.

.

[109] SHINWARI, D., SCHERER, R. C., DEWITT, K. J., and AFJEH, A. A, "Flow visualization and pressure distributions in a model of the glottis with a symmetric and oblique divergent angle of 10 degrees," *Journal of the Acoustical Society of America,* vol. 113, no. 1, pp. 487-497, 2003.

[110] SULC, J., "Emotional changes in human voice," *Activitas Nervosa Superior*, vol. 19, pp. 215-216, 1977.

[111] TEAGER, H. M. and TEAGER, "Evidence for nonlinear sound production mechanisms in the vocal tract," *NATO Advanced Study Institute, Series D,* vol. 15, 1990.

[112] TEAGER, H. M. and TEAGER, S. M., "A phenomenological model for vowel production in the vocal tract," *Speech Science: Recent Advances*, pp. 73–109, 1983.

[113] TEAGER, H. M., "Some observations on oral air-flow during phonation," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 28, pp. 599-601, 1980.

[114] THOMAS T. J., "A finite element model of fluid flow in the vocal tract," *Comput. Speech Lang.,* vol. 1, pp. 131-151, 1986.

[115] TONGE, B. J., "Depression in young people," *Australian Prescriber,* vol. 21, pp. 20-22, 1998.

[116] VERVERIDIS D., and KOTROPOULOS C., "Emotional speech recognition: Resources, features, and methods," *Speech Communication,* vol. 48, no. 9, pp. 1162-1181, 2006.

[117] WENGER, M. A., *Emotion as visceral action: An extension of Lange's theory. In M. L. REYMERT (ed.) Feelings and Emotions: The Mooseheart Symposium*. New York: McGrawHill, 1950.

[118] WILLIAMS, C. E. and STEVENS, K. N., ''On determining the emotional state of pilots during flight: an exploratory study," *Aerospace Med*, vol. 40, pp. 1369–1372, 1969.

[119] WILLIAMS, C. E. and STEVENS, K. N., "Emotions and speech: Some acoustic correlates," *Journal of the Acoustical Society of America*, vol. 52, pp. 1238-1250, 1972.

[120] WILSON, E. O., *Sociobiology: The new synthesis.* Cambridge, MA: Belknap, 1975.

[121] WORLD HEALTH ORGANIZATION (WHO) MENTAL HEALTH DEPARTMENT [Online].Available:http://www.who.int/mental_health/management/depression/definition/en/.

[122] YOUNG S., "HTK: The Hidden Markov Model Toolkit V3.4," 1993 [Online]. Available: http://htk.eng.cam.ac.uk.

[123] ZHOU, G. J., HANSEN, J. H. L., and KAISER, J. F., "Nonlinear feature based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing,* vol. 9, pp. 201-216, 2001.