SPEECH EMOTION RECOGNITION USING AUDITORY MODELS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS INSTITUTE
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ENES YÜNCÜ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF COGNITIVE SCIENCE

SEPTEMBER 2013

SPEECH EMOTION RECOGNITION USING AUDITORY MODELS

Submitted by **Enes Yüncü** in partial fulfillment of the requirements for the degree of **Master of Science in Cognitive Science, Middle East Technical University** by,

Prof. Dr. Nazife Baykal _____
Director, **Informatics Institute**

Prof. Dr. Cem Bozşahin _____
Head of Department, **Cognitive Science**

Asst. Prof. Dr. Hüseyin Hacıhabiboğlu _____
Supervisor, **Game Technologies, METU**

Prof. Dr. Cem Bozşahin _____
Co-Supervisor, **Cognitive Science, METU**

Examining Committee Members:

Prof. Dr. Deniz Zeyrek _____
Cognitive Science, METU

Prof. Dr. Cem Bozşahin _____
Cognitive Science, METU

Asst. Prof. Dr. Hüseyin Hacıhabiboğlu _____
Game technologies, METU

Asst. Prof. Dr. Cengiz Acartürk _____
Cognitive Science, METU

Dr. Ceyhan Temürcü _____
Cognitive Science, METU

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: ENES YÜNCÜ

Signature :

ABSTRACT

SPEECH EMOTION RECOGNITION USING AUDITORY MODELS

Yüncü, Enes
M.Sc., Department of Cognitive Science
Supervisor : Assist. Prof. Dr. Hüseyin Hacıhabiboğlu
Co-Supervisor : Prof. Dr. Cem Bozşahin

September 2013, 67 pages

With the advent of computational technology, human computer interaction (HCI) has gone beyond simple logical calculations. Affective computing aims to improve human computer interaction in a mental state level allowing computers to adapt their responses according to human needs. As such, affective computing aims to recognize emotions by capturing cues from visual, auditory, tactile and other biometric signals recorded from humans. Emotions play a crucial role in modulating how humans experience and interact with the outside world and have a huge effect on the human decision making process. They are an essential part of human social relations and take role in important life decisions. Therefore detection of emotions is crucial in high level interactions. Each emotion has unique properties that make us recognize them. Acoustic signal generated for the same utterance or sentence changes primarily due to biophysical changes (such as stress-induced constriction of the larynx) triggered by emotions. This relation between acoustic cues and emotions made speech emotion recognition one of the trending topics of the affective computing domain. The main purpose of a speech emotion recognition algorithm is to detect the emotional state of a speaker from recorded speech signals.

Human auditory system is a non-linear and adaptive mechanism which involves frequency-

dependent filtering as well as temporal and simultaneous masking. While emotion can be manifested in acoustic signals recorded using a high quality microphone and extracted using high resolution signal processing techniques, a human listener has access only to cues which are available to him/her via the auditory system. This type of limited access to emotion cues also reduces the subjective emotion recognition accuracy. A speech emotion recognition algorithm based on a model of the human auditory system is developed and its accuracy is evaluated in this thesis. A state-of-the-art human auditory filter bank model is used to process clean speech signals. Simple features are then extracted from the output signals and used to train binary classifiers for seven different classes (anger, fear, happiness, sadness, disgust, boredom and neutral) of emotions. The classifiers are then tested using a validation set to assess the recognition performance. Three emotional speech databases for German, English and Polish languages are used in testing the proposed method and recognition rates as high as 82% are achieved for the recognition of emotion from speech. A subjective experiment using the German emotional speech database carried out on non-German speaker subjects indicates that the performance of the proposed system is comparable to human emotion recognition.

Keywords: emotions, acoustic, auditory model, auditory filterbank, support vector machine

ÖZ

İŞİTSEL MODELLERİ KULLANARAK OTOMATİK KONUŞMA DUYGU TANIMA

Yüncü, Enes
Yüksek Lisans, Bilişsel Bilimler Bölümü
Tez Yöneticisi : Yrd. Doç. Dr. Hüseyin Hacıhabiboğlu
Ortak Tez Yöneticisi : Prof. Dr. Cem Bozşahin

Eylül 2013, 67 sayfa

Bilişimsel teknolojinin ortaya çıkmasıyla, insan bilgisayar etkileşimi (İBE) basit mantıksal hesaplamaların ötesine geçti. Duygusal bilgi işlemleri insan bilgisayar etkileşimini kullanıcının ihtiyaçlarına göre adapte ederek geliştirmeyi amaçlamaktadır. Bu nedenle görsel, işitsel, dokunsal ve diğer biyometrik sinyalleri yakalayarak duyguları tesbit etmeyi hedeflemektedir. Duyguların insanların tecrübe edinimi ,dış dünya ile etkileşimi ve karar mekanizması üzerinde büyük bir etkisi bulunmaktadır. Duygular insan sosyal ilişkilerinin şekillenmesinde ve hayata dair önemli kararların alınmasında rol oynamaktadır. Bu nedenle duyguların algılaması yüksek düzeyde bir etkileşim için çok önemlidir. Her duygunun onları tanımamızı sağlayan benzersiz özellikleri bulunmaktadır. Aynı söyleniş yada cümlenin ürettiği akustik sinyaldeki değişimin sebebi öncelikle biyofiziksel değişikliklerdir (stres kaynaklı daralma, gırtlak gibi). Akustik sinyalindeki farklılıklar ve duygular arasında ilişki, konuşmadan duygu tanımayı duygusal bilgi işlemleri arasında çok çalışılan bir konu haline getirmiştir. Ana amacı, kayıt edilen bir konuşmadaki duygusal durumu duygu tanıma algoritması kullanarak tespit etmektir.

İnsan işitme sistemi frekansa bağımlı filtreleme ve eşzamanlı maskeleme içeren doğrusal ve edinilmiş bir mekanizmadır. Duygusal konuşma yüksek kaliteli bir mikrofon kullanarak

kaydedilip yüksek çözünürlüklü sinyal işleme teknikleri ile analiz edilebilirken, insan bir dinleyici ancak işitsel sisteminin ona sağladığı verileri kullanabilir. Bu tür duygusal verilere sınırlı erişimi de öznel duygu tanıma doğruluğunu azaltır.İnsan işitme sisteminin bir modelini temel alan bir konuşma duygu tanıma algoritması geliştirildi ve onun doğruluğu bu tez kapsamında değerlendirildi. İşitsel filtreleme tabanlı insan duyma modeli temiz konuşma sinyalleri işlemek için kullanıldı. Elde edilen çıktılardan basit özellikler çıkarıldı ve yedi farklı sınıftaki (öfke, korku, mutluluk, üzüntü, tiksinme, can sıkıntısı ve nötr) duygular için ikili sınıflandırıcı eğitmek için kullanılandı. Geliştirilen sınıflandırıcı daha sonra tanıma performansını değerlendirmek için kullanıldı. Almanca, İngilizce ve Lehçe olmak üzere iki duygusal konuşma veritabanları, önerilen yöntem ile test edildi ve tanıma oranları %82 olarak belirlendi. Almanca veritabanı kullanılarak hazırlanan öznel tanıma testi sonuçlarının, geliştirilen otomatik konuşma duygu tanıma sistemi ile kıyaslanabilir olduğu tesbit edildi.

Anahtar Kelimeler: duygular, akustik, isitsel modeller, isitsel filtre dizgisi, destek vektor makinasi

dedicated to my wife..

ACKNOWLEDGMENTS

First of all, I would like to thank to my supervisors Assist. Prof. Dr. Hüseyin Hacıhabiboğlu and Prof. Dr. Cem Bozşahin. Assist. Prof. Dr. Hüseyin Hacıhabiboğlu has contributed in all phases of my studies and provided great patience, support and encouragement. Prof. Dr. Cem Bozşahin has always supported and guided me to become a cognitive scientist. Under their guidance, I have had a change to develop myself both academically and personally.

I would like to thank to all ISSD family where I worked. They have always watched my back and shown great patience.

Finally, my biggest gratitude is to my newly small family, my wife Afife Yüncü. Ever since I could remember, she always stand my me. Moreover, special thanks to my mom, father and brother for their support and love. I am proud to be your son.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Affective Computing

Humans interact with their environment using many different sensing mechanisms. Human body captures all of this information, evaluates and forms an emotional state in response. Emotions play a crucial role in the whole human experience and they have a huge effect on the human decision making process. They are important part of human social relations and take role in important life decisions.
With the advent of technology, human machine interaction has gone beyond simple logical calculations. As the spread of technological devices increases, the need for high level human-computer interactions requires more natural interfaces. Assigning human like properties to computers like observing, interpreting and generating affective features is known as affective computing (Jianhua Tao, 2005). Affective computing enables computers to detect human emotional state and respond to their users according to their needs. In order to do this, the computer first recognizes the emotion, and then generates a response based on its preset characteristics. This requires an interdisciplinary understanding, involving cognitive science, psychology, sociology and computer science. Affective computing aims to improve human computer interaction and adapt machine responses according to human needs. In recent years, affective computing has been used to evaluate users' pleasant/unpleasant state during interaction. Detecting an unpleasant state during the task and intervening the process is possible with real time affective systems. In human computer interaction, the main task is to keep users' level of satisfaction as high as possible. A computer with affective properties could detect the users' emotion and could develop a counter response to increase user satisfaction.
Affective computing aims to recognize emotions, via capturing cues from body patterns. There are two general measurement namely; physiological and behavioral (Ruiz, 2011). Physiological measurement methods include heart rate, galvanic skin response and blood pressure. Behavioral measures are gathering features from speech, dialog patterns, linguistic, face and body gestures. Speech and gesture recognition are the most popular affective computing topics. Speech and gesture recognition are possible with passive sensors. Moreover, real time processing option and availability of data made them trend topics in this field. In the scope of this thesis, speech emotion recognition task is applied. Acoustic properties speech signals are used to extract the emotional state of a speaker. Speech contains a rich set of information. Both emotions and words are carried though speech. Such properties made speech a valuable source for affective computing. Affective computing do not only enhance human computer interaction, but also improve the computer's ability of decision making (Picard, 1995). In near future, computers could advise human about what to do. A computer with emotions could behave like humans and commit decisions just like humans. In real life case, rational thinking with itself is not enough to make decisions or suggestions. In order to generate human-like

decisions, computers should use their emotions.

1.2 Speech Emotion Recognition

Interpersonal communication is an interaction which involves the exchange of reciprocal ideas and emotions. Gestures and sound are a way of conveying information in a human-to-human interaction. Speech, a special form of sound, is one of the fundamental ways of conveying information between people. Words are not sole component of speech. Acoustic properties of speech also carry important affective features. Emotions exist in every part of the speech. Emotions in speech are transmitted from one communicator to another during an interaction. As a result of exchange of emotions during an ongoing conversation, emotional state of a speaker may easily trigger an interlocutor emotional state resulting in a change in the speech style or tone. Emotional states are transferred and mutually shaped through this process.
Communication involves a source and a receiver. Speaker is the source and the listener is the receiver. In a dyadic conversation, participants are both source and receiver in turns. On the source side, vocal track plays important roles. In the vocal fold, speech is shaped in a way which reflects the emotional state of the speaker. On the receiver side, heard speech signal is exposed to a series of transformations in the auditory system. Auditory system converts the voice in a way, which allows us to perceive content and context. Speech emotion recognition resides on the receiver side which aims at recognizing the underlying emotional state of a speaker.
Acoustic part of speech carries important cues about emotions. Each emotion has unique properties that make us recognize them. Main task of a speech emotion recognition algorithm is to detect the emotional state of a speaker from speech. General automatic speech emotion recognition algorithm composed of two parts which are feature extraction and classification stages. Prosodic and spectral features of speech are the most popular features that are used in speech emotion recognition algorithm. Intonation, pause, stress, pitch and rhythm are prosodic feature examples. Spectral features investigate frequency components of speech signal. Used classification algorithm varies from algorithm to algorithm. Support vector machines, Gaussian mixture models, hidden Markov model and neural networks are the most popular classification algorithms used in speech emotion recognition task. Automatic speech emotion research generally focuses on the selection of right feature set and to detect in which emotion which features are more affective. Different from these studies, in this study, human auditory system is investigated. Our brain investigates the input from our auditory system. There are many auditory models that can simulate the process in our ear. In this study one of the models is investigated and its output is used to extract the emotional state of a speaker. The output of the auditory model is named as modulation transfer function. Yet, there is not any model that is able to tell how our brain evaluate the data from the auditory system and extracts the emotions, this study aims to constitute a machine learning algorithm which recognize emotions using the features extracted from a computational model of human auditory system. In the scope of this study, selected German, Polish and English databases are applied machines. Besides machines, to make comparison, German database is applied to human listeners who do not know any German to measure human emotion recognition rate using only acoustic cues. Comparison results are going to verify the success rate of a computational model of human-like speech emotion recognition algorithm.

1.3 Organization of the Thesis

This thesis is organized as follows. Chapter 2 explains the contribution made to cognitive science. Embodiment and distributed emotions concepts are introduced. The relationship between emotion and cognition is explained. Importance of speech emotion recognition task in a cognitive model of human to human-like machine dialog model is introduced. Chapter 3 presents the background information about emotions, speech emotion recognition algorithms and human and computational auditory system. Chapter 3 starts with the information about emotion models and dimensions of emotion. In automatic speech emotion recognition part, mostly used feature sets, classification algorithms and emotional speech databases are presented. Literature review about previously developed speech emotion recognition algorithms is provided. In addition to them, human auditory system and its computational is exhibited. In chapter 4, details about the developed speech emotion recognition algorithm is presented. Feature extraction and classification task are explained in details. In chapter 5, subjective speech emotion recognition test and the performance of the algorithm is presented. In next chapter, made work is discussed.In chapter 7, conclusion is made and in chapter 8, information about possible future work is explained.

CHAPTER 2

Contributions to Cognitive Science

Providing one comprehensive definition for emotion is a hard task. My favorite definition for emotion is that, emotion is a way of representing ones circumstances, mood or relations to others. Emotion is a way of characterizing states of mind like joy, anger, love, hate derived from natural instinctive state of mind (Gordon, 1990). Emotion has many forms and many different representations. Body reactions like speech, face, walking type provides some cues about emotions.

Recent brain research shows that, there is a part, which plays an important role in expressed and embodied emotions (Gordon, 1990). The amygdala is a small structure in the limbic system, which resides on brains medial temporal lobe. It is revealed that when pictures of threating faces are shown to an individual, some neurons are activated from this cluster (Scheingold, 2010). Moreover, this region has a major role in recognizing emotional responses like facial expressions.

In the literature, there are some different perspectives about theories about emotions. Evolutionary theory is in the middle in Darwinian perspective (Izard, 1984). It is stated that without evolutionary history, emotions cannot be understood. As an example, innate need for survival forces the emotions when we see a bear. Second perspective is Jamessian perspective. It claims that, a bodily change which caused by an outside stimuli is required in order to experience emotions (James, 1994). Our adrenaline production increases involuntarily when we see a bear, which results in fear as an emotion. In cognitivists aspect, physiological changes are experienced immediately and imperceptively when a bear is seen. Socio-constructivist perspective states that emotions are products of culture (Parkinson, 1996). In this aspect, culture imposes us that bear is frightening and when you saw a bear, you should fear.

An interesting approach to emotions is the dimensions of emotions. Emotional space is constituted by two dimensions namely; valence and arousal. Valence represents the level of pleasantness or unpleasantness. Arousal or with its other known name activation is the level of bodily energy. Neutral remains in the middle of two dimensional space. Happy and anger both have high activation level. However, happy has a positive valence value and anger has a negative valence value. Sad has both low level of arousal and valence. Emotion space is important in embodied emotions since, emotional space has a representation in physical level at the human body. To illustrate, when experiencing happy and anger, the amount of energy released will increase due to the quickness of pulse. In high activation level and negative valence level emotions such as fear and anger, we feel shortness of breath and the trembling (Hatsimoysis, 2003). When one is fearful, the eye brow muscles grow tense, whereas if one feels joyful, the muscles in the check will form a smile position.

In this section, arguments about embodied emotions will be discussed. Mind-body relation is induced to body-emotion in this case. Main arguments in embodied emotions are generally about Jamessian perspective. He claims that when we perceive some bodily changes because

of outside stimuli, emotions occur. Our feelings for the same changes are emotions. On the other hand, there are many opposing ideas. They stated that, emotions are more like judgments or thoughts, than perceptions (Hatsimoysis, 2003). On my viewpoint this view is true if we are trained for that case. Such that, in a clash, soldiers emotions are related with their body status.

When a body reaction occurs, unconsciously our brain forces us into an emotional state. It is hard to judge emotions in most of the cases. I defend the idea that emotions are based on perceptions of patterned changes in the body. Emotional dimension is a good proof for these patterned changes. Six basic emotions (anger, disgust, fear, joy, sadness, surprise) have unique body patterns. Anger, disgust, fear and surprise reside on the upper half of the emotional space. All requires a high activation level which means, our heart pulse rate increases. Since all basic emotions have a unique body pattern embodied emotion thesis is supported. Yet, there are some opposing ideas to this approach. In (Hatsimoysis, 2003), some opposing arguments are provided. I have briefly discussed the most interesting ones. Critics point out that some emotions do not involve bodily change at all. Does guilt or loneness has any relation with body movement(Hatsimoysis, 2003)? Long standing emotions correlation with the body changes over time. If someone is in love for a long time, does he always have a high heart rate? Valence level of the lover changes over time. Other critics come from another point. All perceived bodily changes are not emotions. Sport makes people happy. Exercise causes a change in the arousal level, which make them to perceive happier.

In the literature, many emotional modes have been developed. The aim of these models is to generate an artificial agent, who has embodied emotions. These models try to generate emotions using environment and body. These artificial characters should be upset when they lost money just like real humans. In (Bartneck, 2002), success condition of an emotion model is defined such that, generated model should show the right emotion in the right time with right intensity. The OCC (Ortony, Clore, & Collins, 1990) is a complex yet well-known emotional model. The OCC model is able to distinguish 22 different emotional categories. In this model, each emotion has a weight. It is expected that weights should not change rapidly, since a regular persons emotional state do not change rapidly. The OCC model has five phases. The name of first phase is classification. In the classification phase, event or an action is evaluated by the character. Each object or an event has a relation with particular emotions. In this phase, affected emotions are selected result of an event or action. Next phase is the quantification phase. Every event or action requires different emotions to be activated with different intensity levels. In this phase, intensity levels of the activated emotions are determined. Following phases name is interaction. In the interaction phase, shift from one emotional state to other one is modeled. For a person who is surfing on the internet, speed of the internet is important. If the speed of the internet drops this people get angry. If you give this a food that he loves, his anger will fade slowly. Interaction phase models this smooth transition. Next phase is mapping. As mentioned OCC model could distinguish 22 different emotions. In the classification phase, activated emotions were selected. In this phase, first physical state is determined, and then appropriate emotions are selected from the activated emotions. Facial expression could not expose all emotions but some of them. Final phase is the expression phase. In the expression phase, emotions are exposed with all possible channels.

The OCC model is designed to mimic human like emotions. Each OCC model has a character. Consistency is important to generate a stable character. If banana make a character happy, then in the next time should character required to be happy? If a banana is given after one another, then happiness level will drop to neutral level. On the other hand, if banana is given after a certain time again a high level of happiness should be generated.

Distributed emotion is a framework which inherits all aspects of emotion and investigates a

full picture of emotion among people and environment (Hollan, Hutchins, & Kirsh, 2000). We live in a very dynamic society and emotion plays a crucial role to shape these relations. Instead of resizing the concept of emotion to individuals, it must be related with the interactions between individuals (Parkinson, 1996).

Emotion is distributed among people. One group member can fire the emotions of all group members or one members sadness makes other members to feel sad. Mimicking of emotions is observed very frequently in intimate groups. By the time, close friends smile even look similar. Different from mimicking emotions, same event may cause different emotions on different people. One event may make a people feel happy on the other hand, made others sad. Rain is a good example. Slight rain make a farmer happy, on the other hand, made a basketball player unhappy. In (Glazer, 2003), distributed emotion is segmented into 3 titles. Emotion is distributed across members of social groups. Emotions are shared socially. One individuals emotion may distribute among the social group. Emotion is coordinated between external material or environmental and internal structures. People can load their emotions into physical structures and can load them back. Photographs are good examples for this case. When you look at your photographs, you memorize the past event, past emotions you have felt at this event. Emotion is distributed though time. Emotions are time-varying perceptions. When an event occurred, your emotions are generally sharper. People may respond some events overreactions at the beginning of the event. Years later, when you remember the same event, you could assess event as an adventure and you could feel happy. Other tendency is that generally grief, fear turns to rage and anger.

Embodied mind thesis claims the argument that nature of our mind is largely determined by the form of the human body. Embodied emotion argument claims that when we perceive some bodily changes because of outside stimuli, emotions occur. This argument generalizes the embodied emotion argument. We can resize the argument into the form that, emotions are only perceptions. In the big picture, most of it is true. It is fact that, we react unconsciously to the most of the events. This reaction gives birth to some emotions. Some scientific studies also showed that, there is a region in our brain, which controls embodied emotions.Embodied emotion framework is important in level of designing a human-like robot. Without emotions, it is hard to call a robot humanoid. Embodied emotions framework offers a basic model. Embodiment thesis enables the transformation of environmental inputs on body to emotions. We survive in a social society. These social relations are generally shaped by emotional state of ours. In a society, emotions flows though interaction. Emotions mix with each other and turn into other emotional states. Emotions are just like colors. Different from colors, emotions have only two primary units, valence and activation. In a happy society, valence flows one individual to others. Since emotions are distributed over time, valence level always changes. Embodied emotions are directly related with distributed emotions. Each social interaction has a pattern in our body. The concept of bed, have some effects on our body which results in some emotions. Just like this, social interaction generates some concepts and these concepts have body patterns in our body. In a cognitive system, embodied and distributed cognition complement each other.

In a telephone call of a two person, generated emotion model is going to be in closed loop form. What you hear from the speaker has some effects on your body, which generate an emotional state. With this emotional state, you generate a speech. Speech emotion recognition algorithms take part in this task an important role. When you hear something you must convert speech signal into an emotion. In this closed loop cycle, firstly, detected emotion generates a body pattern in your body. Next time you speak, and other listener extracts the emotion. Distributed emotions take part in the transformation of extracted emotions into body patterns. Embodied emotions are results of body patterns. As seen, all concepts are related to

each other.
Although the relationship between emotion and cognition is omitted for a long time, recent studies especially on magnetic resonance imaging (MRI), has provided important cues. Besides being dependent, both have many commonalities as being embodied and distributed. Yet, there are not any common view about emotions, there is a relative agreement upon cognition. Processes related with memory, attention, language and problem solving is known as cognition (Pessoa, 2008). When the distinction between emotion and cognition is projected into brain, affect is being related with unconscious processing and subcortical activity. On the other hand, cognition is related with the conscious processing and cortical involvement. Many studies has shown that emotion has affect on perception, attention, memory and learning. To sum up, emotion has an effect on cognition.
Experiencing emotional expressions evokes increased responses compared to neutral expressions. Emotional visual expressions such as watching a war scene results in an increased activity in visual cortex (Pessoa, 2008). In addition to that, emotional content has an effect on attention. Detection of happy and anger faces are more rapid when compared with neutral ones(Eastwood, 2001). These results could be evaluated such that, amygdala has some effects on cognitive process. It arises two possibilities. First is that, excitation in amygdala has an enhancement effect on visual processing module. Second possibility is that, amygdala results in enhancement in the part of brain which are responsible for attention. Another research has provided interesting information. Emotional faces evoke responses in the amygdala although attention has major on another stimuli (Anderson, Christoff, Panitz, Rosa, & Gabrieli, 2003). This shows that, cognitive and emotional processes occurs independently. Besides attention and perception, in memory event, emotional content has an effect. Studies have shown that humans are better at remembering emotionally arousing information. In a study, subjects are exposed to two videos. First video composed of neutral film clips and the second one contains emotional content (Cahill et al., 1996). After 3 weeks of watching these clips, subjects were better at remembering emotional ones although both clips were taken from the same source. It is stated that, valence dimension do not cause any enhancement in the memory. These studies have shown the relationship between emotion and cognition. Emotional state of a person has some effects on human cognitive system.
At recent years, human to human interaction is being replaced by human machine interaction. Such shift forces machines to behave like humans. In order to accomplish such a task, emotion recognition and generation of a counter emotional state are the crucial part of such devices. Such a system requires an interdisciplinary field which fuses artificial intelligence, psychology, philosophy. Design and development of human like machines are possible with cognitive systems. In the figure 2.1, a human machine dialog flowchart is provided. Given cognitive system is based on an interaction with speech. Speech is one the main source of communication. Speech both carries content and non-verbal elements such as emotions. Therefore, a human like machine should both have speech recognition algorithm and speech emotion recognition algorithm. In the given model, information from speech recognition and speech emotion recognition blocks is combined in artificial intelligence chat box. Chat box will produce an output sentence. In the emotion generator sub block, output sentence emotion will be determined. In the speech synthesizer, generated sentence will be converted to emotional speech.
In the scope of this thesis, speech emotion recognition task is accomplished. Recognition of emotion is crucial in an interaction. In order to design a human like machine, whole stages in the cognitive model should be human alike. In the thesis, a computational model of human emotion recognition system is aimed to be designed. Different from many other speech emotion recognition algorithms, a human auditory model is used. Auditory model outputs

are transformed into feature sets. Auditory model is a computation model of human ear. Yet, the process of emotion recognition in the brain has no model. In order to measure similarity of results, listening test is applied on humans. A German speech dataset is applied both to computers and humans. In the developed computational model is content free. Therefore, human evaluators are selected from the people who do not know any German. Output results of subjective and automatic speech emotion recognition tasks are compared with each other. Comparison of results have provided important information both on the source and receiver sides.

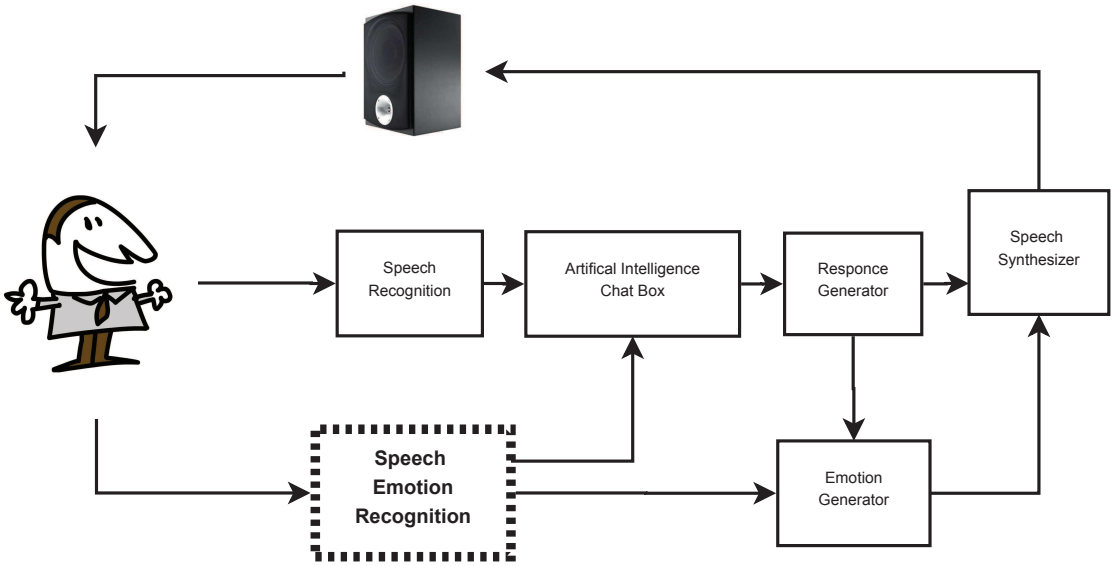

Figure 2.1: Human machine dialog flow chart.

CHAPTER 3

Background

3.1 Introduction

Emotions play a crucial role in modulating how humans experience and interact with the outside world and have a huge effect on the human decision making process. They are an essential part of human social relations and take role in important life decisions. Therefore detection of emotions is crucial in high level interactions. Each emotion has unique properties that make us recognize them. Acoustic signal generated for the same utterance or sentence changes primarily due to biophysical changes (such as stress-induced constriction of the larynx) triggered by emotions. This relation between acoustic cues and emotions made speech emotion recognition one of the trending topics of the affective computing domain. The main purpose of a speech emotion recognition algorithm is to detect the emotional state of a speaker from recorded speech signals.
In this section, background information about speech emotion recognition task is provided. This chapter is organized as follows. In section 3.2, proposed models for emotions are presented. In section 3.3, mostly used features, primary emotions, machine learning algorithms and speech emotion databases are introduced. In section 3.4, both human auditory system and computational model for human auditory system is presented. In this section, details about the process in ears is introduced. In the computational model, evaluation of the auditor models and their correspondence in human auditory system is presented. At the end of the chapter, a table in which presents literature made about the speech emotion recognition is exhibited.
Human auditory system is a non-linear and adaptive mechanism which involves frequency-dependent filtering as well as temporal and simultaneous masking. While emotion can be manifested in acoustic signals recorded using a high quality microphone and extracted using high resolution signal processing techniques, a human listener has access only to cues which are available to him/her via the auditory system. This type of limited access to emotion cues also reduces the subjective emotion recognition accuracy. In the scope of this thesis, biggest contribution is made in developing an algorithm based on a model of the human auditory system. Its accuracy is evaluated and compared with subjective listening tests.

3.2 Models of Emotions

Many different set of emotions are defined in order to conceptualize and put a distinction between emotions. In different times and by different people, emotions are modeled. In this models, some of them claimed some emotions as primary or basic emotions. Some defined emotions as a mixture model instead of discretizing them. As research interest in automatic emotion recognition research increases, detection of a set of emotions become more important. Palette theory was proposed by (Descartes, 1952) in order to describe all the emotions as

a mixture of primary emotions. These primary emotions are anger, disgust, fear, joy, sadness and surprise. These emotions are the most distinct emotions and named as archetypal emotions. In palette theory, other emotions forms from the mixture of primary emotions. Second set of emotions are proposed by (Ekman, 1992). Proposed basic emotions are anger, happiness, surprise, disgust, sadness and fear. The theory of basic emotions are come into view since corresponding facial expression of each basic emotion is unique and universal. Wheel of emotions is another popular theory (Plutchik, 2011). Wheel of emotions is composed of 8 basic emotions and 8 advanced emotions each composed of 2 basic ones. Basic emotions are joy, trust, fear, surprise, sadness, disgust, anger and anticipation. Advanced emotions are constituted by mixing two basic emotions.
Emotion dimensions is another popular theory. Emotion is categorized into three basic dimensions namely; activation, valence and strengthen (Schlosberg, 1954). Activation ranges from sleep to tension (Murray & Arnott, 1993). It is used to refer to the amount of energy required to express a certain emotion. Emotions joy, anger and fear requires increased heart rate, higher blood pressure which results in dryness of mouth and occasional muscle tremor thus higher activation level. On the other hand, sadness results decrease of the heart rate and lower blood pressure, causing a lower activation level (E. Ayadi, Moataz, Kamel, & Karray, 2011). As stated in (E. Ayadi et al., 2011), both happiness and anger requires high activation. Hence, activation itself is not enough to distinguish emotions. As mentioned in (Murray & Arnott, 1993), valence corresponds to pleasantness to unpleasantness, such as the difference between happiness and anger. Strength is also known as control or dominance (Hanjalic, 2006). Strength is useful to distinguish grief and rage. It ranges from no control to full control (Hanjalic, 2006). In (Hanjalic, 2006) emotional dimension curves are constituted with an aim to design a video recommender system. During a football match broadcast, activation changes are easily observed during a break due to foul or cheering of the crowd.
In the scope of this thesis, seven emotion labels are classified using developed speech emotion algorithm. Emotion labels are happiness, anger, fear, disgust, neutral, boredom and sadness.

3.3 Automatic Speech Emotion Recognition

Speech emotion recognition has been a hot research topic during the past 20-15 years within the context of effective computing. Speech emotion recognition is defined as extracting the emotional state of a speaker from his or her speech (E. Ayadi et al., 2011). Emotional features of a speech are embedded in its non-verbal elements. Linguistic content does not have to be affected from emotional state but it may also provide important feedback (Ververidis & Koropoulos, 2006). Many investigations have been made to extract the emotional state of a speaker from speech (Tawari & Trivedi, 2010), (Nwe, Foo, & Silva, 2003), (Petrushin, 2000). Increase in the available computational power led to development of many applications. In the study (Ang, Dhillon, Krupski, Shriberg, & Stolcke, 2002) a ticket reservation system named [SmartKom] is reported. In that work, an automatic speech recognition system which is able to recognize annoyance or frustration is developed. In (Narayanan, 2005),an emotion recognition system is developed for call centers. In addition to those applications, three automatic emotion recognition system is used in a cruise in-car navigation system in which mental state of the driver is monitored in order to increase the safety (Schuller, Rigoll, & Lang, 2004). In another study (Lee, Busso, Lee, & Narayanan, 2009), a model is developed to describe the emotional states and their mutual influence in a dyadic conversation.

3.3.1 Summary of Primary Emotions as Manifested in Speech Signals

3.3.1.1 Anger

Anger requires high energy to be expressed. Definition meaning of the anger is simple extreme displeasure (Murray & Arnott, 1993). In case of anger, aggression increases in which control parameter weakens. Anger is stated to have the highest energy and pitch level when compared with the emotions disgust, fear, joy and sadness (Ververidis & Koropoulos, 2006). The widest observed pitch range and highest observed rate of pitch change are other findings about the emotion label anger when compared with other emotions (Murray & Arnott, 1993). Besides a faster speech rate is observed in angry speeches (Burkhardt & Sendlmeier, 2000).

3.3.1.2 Fear

In emotional dimension, fear has similar features to anger. High pitch level and raised intensity level are correlated with fear (Ververidis & Koropoulos, 2006). It is stated that fear has a wide pitch range. Highest speech rate is observed in fear speeches (Murray & Arnott, 1993). The pitch contour trend separates fear from joy. Although the pitch contour of fear resembles the sadness having an almost downwards slope ,emotion of joy have a rising slope(Ververidis & Koropoulos, 2006).

3.3.1.3 Sadness

In emotional dimension, sadness requires very low energy. In addition, valence degree is negative. Sadness exhibits a pattern that is normal or lower than normal average pitch, a narrow pitch range and slow tempo (Murray & Arnott, 1993). Speech rate of a sad person is lower than the neutral one (Ververidis & Koropoulos, 2006).

3.3.1.4 Joy/Happiness

Joy/happiness exhibit a pattern with a high activation energy, and positive valence. Strength of the happiness emotion may vary. In the emotional state happiness or joy, pitch mean, range and variance increases (Ververidis & Koropoulos, 2006). In (Murray & Arnott, 1993), it is stated that fundamental and formant frequencies increases in case of smile. Moreover, amplitude and duration also increase for some speakers.

3.3.1.5 Disgust

In (Ververidis & Koropoulos, 2006), low mean pitch level, a low intensity level, and a slower speech rate is observed when disgust is compared with the neutral state. Disgust is stated the lowest observed speech rate and increased pause length (Murray & Arnott, 1993).

3.3.1.6 Boredom

Boredom is a negative emotion with negative valence and low activation level same as sad. A lowered mean pitch and a narrow pitch range with a slow speech rate are defined as the properties of a bored expression (Burkhardt & Sendlmeier, 2000).

3.3.2 Acoustic Features Widely Used In Emotion Recognition

As mentioned previously, extraction of speech features is a very important process in speech emotion recognition. Speech features can be divided into several categories. In (E. Ayadi et

al., 2011), speech features are divided into 4 categories; continuous, qualitative, spectral and TEO-based. Continuous features are pitch, energy and formants. Quantitative features are described as voice quality features which are harsh, tense and breathy voices. Most popular acoustic features used in emotion recognition process are outlined below.

3.3.2.1 Pitch features

Pitch is the fundamental frequency of the glottal excitation. Pitch depends on the tension of the vocal folds and subglottal air pressure. Pitch frequency is one of the widely used features in emotion from speech applications. Pitch frequency is also known as the fundamental frequency. The time elapsed between successive vocal fond openings determine the fundamental frequency (Ververidis & Koropoulos, 2006). From pitch features given features could be extracted which are min, max, mean, standard deviation, range at the turn level, slope (mean and max) in the voiced segments, regression coefficient and its mean square error and maximum cross-variation of F0 between two adjoining voiced segments (inter-segment) and with each voiced segment(intra-segment) (Vidrascu & Devillers, 2005).

3.3.2.2 Teager energy operator

Produced number of harmonics due to the non-linear air flow in the vocal tract is another useful acoustic feature. In case of anger, the fast air flow causes nonlinear stress (Teager, 1990). In (E. Ayadi et al., 2011), it was stated that TEO-based features can be used to detect stress in speech.

3.3.2.3 Vocal tract features

Formants are a vocal tract feature. Each formant has its own bandwidth and center frequency (Ververidis & Koropoulos, 2006). Slackened speech can be distinguished from an articulated speech using formant features. Other widely used feature is the energy of a certain frequency which corresponds to the critical bands of the human ear (Ververidis & Koropoulos, 2006).

3.3.2.4 Spectral features

Mel-frequency cepstrum coefficients, linear predictive coding and log frequency power coefficients are the most popular spectral features. Mean and standard deviation of 13 Mel frequency cepstral coefficients (MFCC) are set as discriminating features in many studies. (D. Wu, Parsons, & Narayanan, 2010).

3.3.2.5 Duration features

Mean and standard deviation of the duration of voiced and unvoiced segments, ratio between the duration of unvoiced and voiced segments are (D. Wu et al., 2010) duration features.

3.3.2.6 Energy features

Energy mean, standard deviation, maximum, 25% and 75% quantiles, and the inter quantile distance are the popular energy based features used in speech emotion recognition task (D. Wu et al., 2010).

3.3.3 Machine Learning Algorithms used in Speech Emotion Recognition

There are various types of classification methods employed for using in the task of emotion recognition from speech. Hidden Markov models (HMM), support vector machines (SVM), Gaussian mixture models and artificial neural networks are so far the most popular classifiers (E. Ayadi et al., 2011).
Although Hidden Markov Model has many design issues such as determining the optical number of states, the type of observations and optimal number of observation symbols (E. Ayadi et al., 2011); performance of HMMs for speech emotion recognition task is satisfactory. In (Nwe et al., 2003),empirical assessments of many state numbers are tested. It is claimed that a four state HMM gives the optimal performance to determine the number of states. HMM is generally used to model the time dependency of the system. In (Nwe et al., 2003), it is claimed that frame size of 16 ms and overlapping duration 9 ms gives the best results. The best average rates were 78.5% and 75.5% for Burmese and Mandarin databases, respectively. This compares favorably with respect to human accuracy rate, which was 65.8%.
Gaussian mixture models represent features with joint probability density functions GMMs are more appropriate for speech emotion recognition where only global features such as mean and variance of the fundamental frequency are used (E. Ayadi et al., 2011). Since GMM is based on the assumption that all vectors are independent, it cannot model temporal patterns. Determining the number of mixture model is an important problem with GMMs. In (Schuller et al., 2004), 74.83% average classification accuracy for speaker dependent approach and 89.12% for speaker-dependent approach is obtained using a sixteen component GMMs. This corresponds to a performance that is comparable to HMMs.
Support vector machine is a supervised linear classifier. Determining a way to choose the kernel size is the problem of the SVM. Since there is not any certain way to choose proper kernel size, it is not guaranteed to segment the transformed features correctly (E. Ayadi et al., 2011). SVM classifier was employed in (Schuller et al., 2004). For speaker independent approach, classification accuracy was 76.12. Speaker dependent result was 92.95%.
Another popular classifier is the artificial neural network (ANN). ANNs are known to be effective in non-linear mappings (E. Ayadi et al., 2011). In (Petrushin, 2000), a two layer neural network is used. A classification result for normal state is 55-65%, for happiness is 60-70%, for anger is 60-80% and for fear is 25-50%. Among the other classifiers, ANN provides the worst performance.

3.3.4 Databases

Using a proper database is crucial in order to obtain a good accuracy. Low quality databases result in incorrect conclusions (E. Ayadi et al., 2011). There are different scenarios to obtain a recording for database. Telephone calls (Teager, 1990), a human talking to a fake ASR-machine (wizard of OZ model) are possible scenarios. In (Williams & Stevens, 1972), a comparative study was made between actual and simulated recording. Recording of radio announcer in the evident HINDENBURG is used. Comparative results showed that real-life data is not inconsistent with the data obtained from an actor (Williams & Stevens, 1972). It is also mentioned that acted emotions tend to be more exaggerated than real ones. Therefore, in many databases such as Danish Emotional Speech (Engberg & Hansen, 1996), semi-professional actors were employed. Another problem with the databases is that collections encompass only five or six emotions (Ververidis & Koropoulos, 2006). In some of the databases such as the Berlin Corpus (Burkhardt, 2005), same sentence is uttered by the same person emulating different emotions.
The Danish Emotional Speech (DES) (Engberg & Hansen, 1996) is recorded by 2 male and 2

female semi-professional actors. Speech is expressed in five emotional states: anger, happiness, neutral, sadness, and surprise. The database consists of 2 words (yes, no), 9 sentences, and 2 passages. 20 listeners have verified the emotions in database with a score of 67.
Speech Under Simulated and Actual Stress (You, Chen, Bu, Liu, & Tao, 1997) is recorded in English by 32 speakers. The database contains both simulated speech under stress (simulated Domain) and actual speech under stress (actual domain). 9 male speakers who represent three main USA dialects (General American, Boston, New York) uttered 2 recordings of the same word.
The VAM database (Grimm, Kroschel, & Narayanan, 2008) is recorded from natural human communication in a German TV talk show Vera am Mittag. Corpus consists of 12 hours of recordings. VAM database is an audio-visual speech corpus which is spontaneous and very emotional speech. It is recorded from unscripted, authentic discussions between the guests of the talk-show. The VAM database is composed of two parts. First part is recorded by 19 speakers and has 478 utterances. Second part is recorded by 28 speakers and has 469 utterances. First dataset is evaluated by 17 listeners and second part is evaluated by 6 listeners.
The Berlin emotional speech database (EMO-DB) (Burkhardt, 2005) is one of the most popular dataset in speech emotion recognition task. The database consists of emotional speech in seven emotion categories namely; anger, happy, fear, disgust, neutral, sad and boredom emotions. Database provides a high number of repeated words in different emotions. Corpus is constituted by 5 male and 5 female actresses. Each actor uttered 5 long and 5 short daily German sentences. The corpus was evaluated by 25 listeners. Utterances with a higher recognition rate of 80% are selected. In addition to them, if 60 of the listeners selects an utterance natural, then the sentence remained in the database. The final numbers of utterances for the seven emotion categories are as given; sadness (62), boredom (81), anger (127), fear (69), disgust (46), joy (71) and neutral (79).
Database of Polish Emotional Speech (Cichosz & Slot, 2005) is constituted by four male and four female speakers. Each speaker uttered 5 different sentences in 6 different emotions in Polish language. Emotion labels of the database are anger, boredom, fear, joy, neutral and sadness. Each speaker has recorded speech during one session using a condenser microphone. All speakers were graduate students of Polish National Film, Television and Theater School in Lodz, Poland. The quality of the recorded database is assessed by 50 subjects. 60 randomly generated samples from the database are evaluated bu each listener. Each listener classified the six emotions with an average rate of 72%(ranging from 60% to 84% for different subjects).
Surrey Audio-Visual Expressed Emotion (SAVEE) (Haq, Jackson, & Edge, 2008) database is an audio-visual emotional database which is formed from 4 male actors in 7 different emotions. Database consists of 480 British English utterances in total. Emotion labels are anger, disgust, fear, happiness, neutral, sadness and 'surprise. There are 15 sentences for each of the 7 emotion categories. The sentences were chosen from the standard TIMIT corpus. Database were recorded in a visual media lab with high quality audio-visual equipment. To check the quality of performance, the recordings were evaluated by 10 subjects under audio, visual and audio-visual conditions.Each actor's data were evaluated by 5 native English speakers and 5 of them are selected from the ones whom had lived in UK for more than a year. Audio, visual, and audio-visual cases are evaluated by the subjects at utterance level. The 120 clips from each actor were divided into 10 groups, resulting 12 clips per group. For each evaluator, a different data set was created for each of the audio, visual and audio-visual data. Accuracy of performed test for audio data is 66.5%. For visual data, average accuracy is measured as 88% and for audio-visual corpus measured average accuracy is 91.8%.

3.4 Auditory Models

3.4.1 Human Auditory System

Human auditory system, ears, carries out an important task. It captures speech and transforms acoustic properties of speech into a special form that human brain could extract the information. In this process, mechanical signal is transformed into neural excitations. This transformation enables brain to extract spoken words and nonverbal elements such as emotions (Benesty, Sondhi, & Huang, 2007). Acoustically produced speech signal is a special sound to our ears and our brain; therefore humans are able to extract information from speech very efficiently under adverse listening conditions. Understanding of human auditory system carries importance in many fields. Audio codec engineers take advance of the human auditory system when developing audio compression algorithms. A lossy compression algorithm achieves a compression ratio between 5 to 20 percent (Jaiswal, 2009). In a compression algorithm, perceptually irrelevant part of a sound is discarded. This led us to listen music without losing the quality. The quality of a sound is determined by human auditory system. If a noise in an audio file is not heard, then its quality is still the same on perceptual side. Virtual reality system takes advantage of perceptual properties of human auditory system too. A compressed audio takes place less memory so people could listen music, even on their small devices without the requirement of high memory capacity. Moreover, a compressed audio is easy to stream via internet which an able us to take advance of many features. Besides compression algorithm, advanced audio speakers take advance of human auditory systems. Their design is made in way that, human listeners could perceive the sound better. Research on human auditory system, lead to development of computational models of human auditory systems. Recently, auditory models are being used in speech recognition and speech emotion recognition tasks.
(Benesty et al., 2007) The process of understanding speech is divided into two stages namely; an auditory pre-processing stage and speech pattern recognition. In auditory preprocessing stage, speech sound is transformed into internal representations that brain uses. Internal representations are assembled by our brain, and conveyed information in speech signal is extracted. This is dependent of training, familiarity and attention (Benesty et al., 2007).
Three subparts constitutes the human ear. Outer ear, middle ear and inner ear are named as the peripheral auditory system. The outer ear collects incoming air vibrations. The middle ear transforms these vibrations into mechanical vibrations. The inner ear filters and converts these mechanical vibrations into electrical activity in neurons (*Psychoacoustics*, 2001).

3.4.1.1 Outer Ear

Pinna and ear canal constitute the outer ear. Their purpose is to collect and transmit acoustic energy to the eardrum. It also protects the middle and inner ear. Ear canal has a length about 2 cm which correspond to a resonance at around 3-5 kHz. Therefore ear canal attenuates higher and lower frequencies. This results an increased sensitivity to sounds in 3-5 kHz frequency components of speech.

3.4.1.2 Middle Ear

Middle ear composed of three tiny bones called ossicles (malleus, incus, and stapes). Purpose of the middle ear is to match low impedance of the cochlea. Acoustic signal is transferred from a low impedance environment which is air. On the other hand, cochlea contains a high

impedance environment. If impedance matching is not made, most of the acoustic energy transfer will be lost.

3.4.1.3 Inner Ear

(Plack, 2004) Cochlea is the place where acoustic information contained in sound is transformed into a form that brain could evaluate. In cochlea mechanical vibrations are converted into neural impulses. (Robinson, 2000) These impulses are carried onto the brain via auditory nerve. The spiral shaped structure of the cochlea allows short-time frequency analysis of sound signals.
Cochlea hosts the basilar membrane. (Plack, 2004) Mechanical vibrations cause basilar membrane to move up and down. Each place on the basilar membrane corresponds to different frequencies. Entrance of the basilar membrane is thin and tuned to high frequency components. On the other hand, apex of the cochlea is thick and tuned to low frequency components. Sensitivity to a particular frequency component varies along the basilar membrane. Each place on basilar membrane act just like an auditory filter. Each place on the basilar membrane has a different center frequency and a bandwidth for auditory filters. Even a pure tone excites many places on the basilar membrane. Corresponding place on the BM excites the most, yet excitation diminished as the corresponding place changes. Gammatone filters and ERB filters are approximations of auditory filters to simulate the frequency selectivity of the ear. On basilar membrane, frequency selectivity of position changes logarithmically. (Robinson, 2000) Critical band scale on the basilar membrane is named as Bark scale which is based on a log scale.
(Robinson, 2000) Hair cells lie within the basilar membrane which transduces mechanical energy to neural impulses. There are two types of hair cells which are inner and outer hair cells. Inner hair cells are connected with the motor neurons which carries nerve impulses to the brain. Outer hair cell gets feedback from brain, which provides gain control on the hair cells. Inner hair cells only excites when the basilar membrane moves upwards. Half wave rectification is the corresponding behavior of such excitation. (Plack, 2004) Each excited hair cell requires a certain time between the firings. When a hair is excited, if a second tone with lower amplitude and a slightly higher frequency is played, inner hair cells could not respond therefore sound signal could not be heard. This effect is called as spectral masking. In the presence of the first louder tone, our ear does not have the capability of passing information about the second one.
Modulation filterbank is defined as overlapping band pass filters with different center frequencies such like auditory filterbank. Although any mechanical process related to modulation filterbank is found in ear, (Plack, 2004) it is suggested that there are some neurons in the brainstem which are sensitive to modulation frequencies.
(Benesty et al., 2007) Modulation filterbank allows brain to group together sound elements which belong to same sound source. Grouping sound components which have the same modulation spectrum, allow brain to group sound components which belong to same speaker even if their auditory spectrum is different. Such a property allows a figure background analysis such as cocktail-party phenomenon.

3.4.2 Computational Auditory Model

In the light of the human auditory system, many computational auditory models have been developed which describe the signal processing occurs in the ears. Designed models composed of many stages. An each stage, auditory signal is pre-processed with different filters.
In the model (Dau, Puschel, & Kohlrausch, 1996), auditory model composed of 4 main stages.

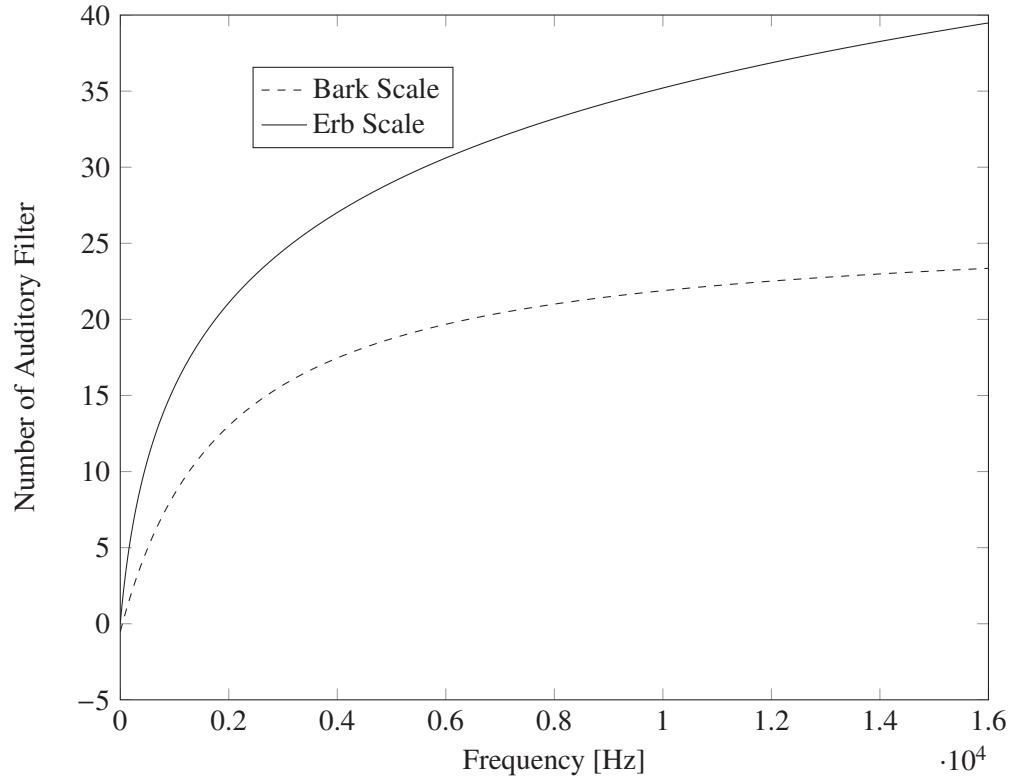

Figure 3.1: Bark Scale and ERB Scale

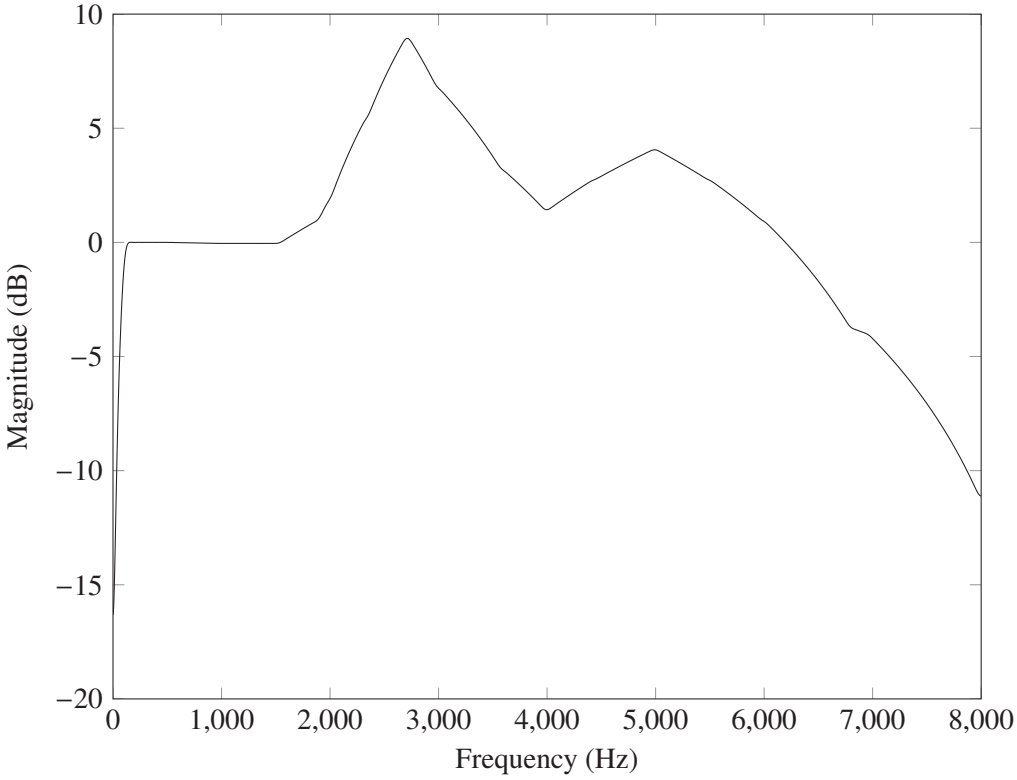

Figure 3.2: Outer and Middle Ear Frequency Response

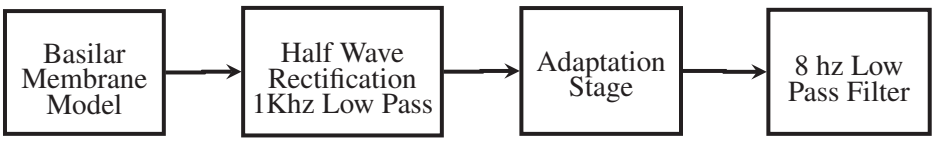

Figure 3.3: Flowchart of the auditory model developed in 1996.

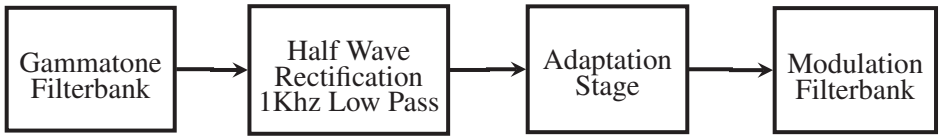

Figure 3.4: Flowchart of the auditory model of developed in 1997.

First stage is basilar membrane model. Auditory filterbanks in cochlea is simulated in linear basilar membrane model. In the second stage, transformation of mechanical oscillators is modeled by an envelope of a signal followed by a 1 kHz low pass filter. In the third stage, adaptation loops are simulated with five feedback loops in chains. In each loop the low-pass filtered output is fed back. Temporal masking phonemana is modeled by these feedback loops. In the final stage, output of the adaptation loops signal is filtered by a low pass filter at 8 Hz. Flowchart of the model is given in figure 3.3.
In the model (Dau et al., 1997), different from the model (Dau et al., 1996), instead of the previosly used linear basilar membrane model, gammatone filterbank is applied. Besides, instead of a 8 Hz low pass filter, modulation filterbank is included. The flowchart of the model is given in figure 3.4. The gammatone filterbank is selected instead of the Strube model since it models the auditory filterbank process better. Frequency response of the gammatone filterbank is given in figure 3.7. After input signal is processed by a bank of gammatone filters, envelope extraction is applied. In the adaptation stage, an extended model is used. In this model, sensitivity for fast temporal variations is increased. Different from 8 hz low pass filter, modulation filterbank is applied. In the modulation filterbank, 12 modulation filters with different center frequencies are applied to the output of adaptation stage signal. Modulation filterbank is included to the model in order to analyze the amplitude changes of the envelope signal. Frequency response of the modulation filterbank is given in the figure 3.8 .
In the model (Jepsen, Ewert, & Dau, 2008), auditory model (Dau et al., 1997) is further extended. The flowchart of the extended model is given in figure 3.5. In this model, outer middle ear transformation is included. Outer-middle ear transformation frequency response is given in figure 3.2. Output signal is processed by dual resonance non-linear filterbank (DRNL) instead of gammatone filter. DRNL is developed to reflect the nonlinearities of basilar membrane (Poveda & Meddis, 2001). Block diagram of the DRNL is given in figure 3.6. Different from the model (Dau et al., 1997), an expansion stage is included after the envelope extraction block. A first order low pass filter with a cut-off frequency of 150 Hz is included before applying modulation filterbank in order to simulate the decreasing sensitivity to sinusoidal modulation as a function of modulation frequency.

3.5 Summary

In this chapter background information about speech emotion recognition is provided. Commonly used features employed in speech emotion recognition is presented. Primary emotions and corresponding distinct features for emotions are illustrated. Used machine learning algorithms are presented. PCA and SVM are the machine learning algorithms that in this thesis

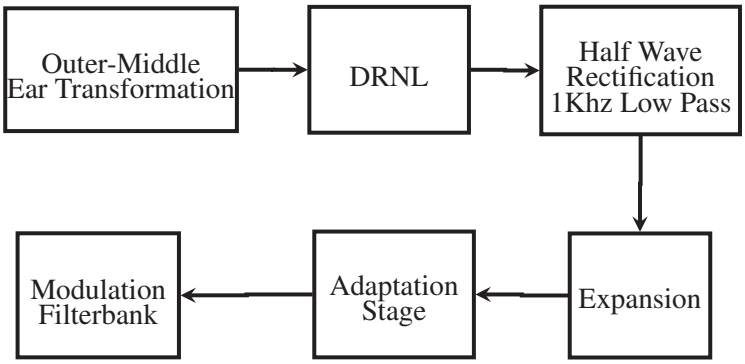

Figure 3.5: Flowchart of the auditory model developed in 2008.

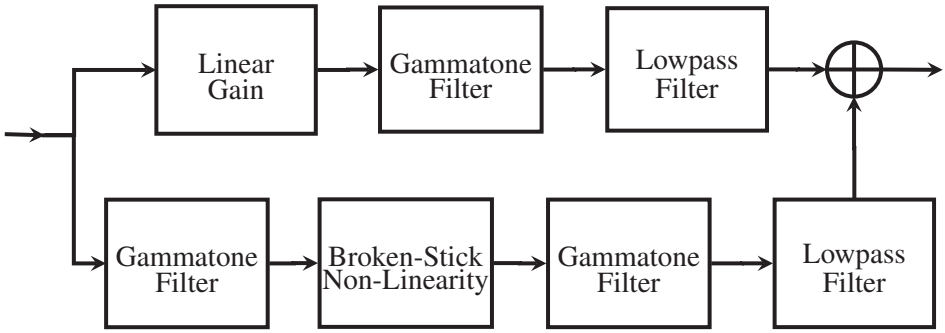

Figure 3.6: Dual Resonance Nonlinear Filter.

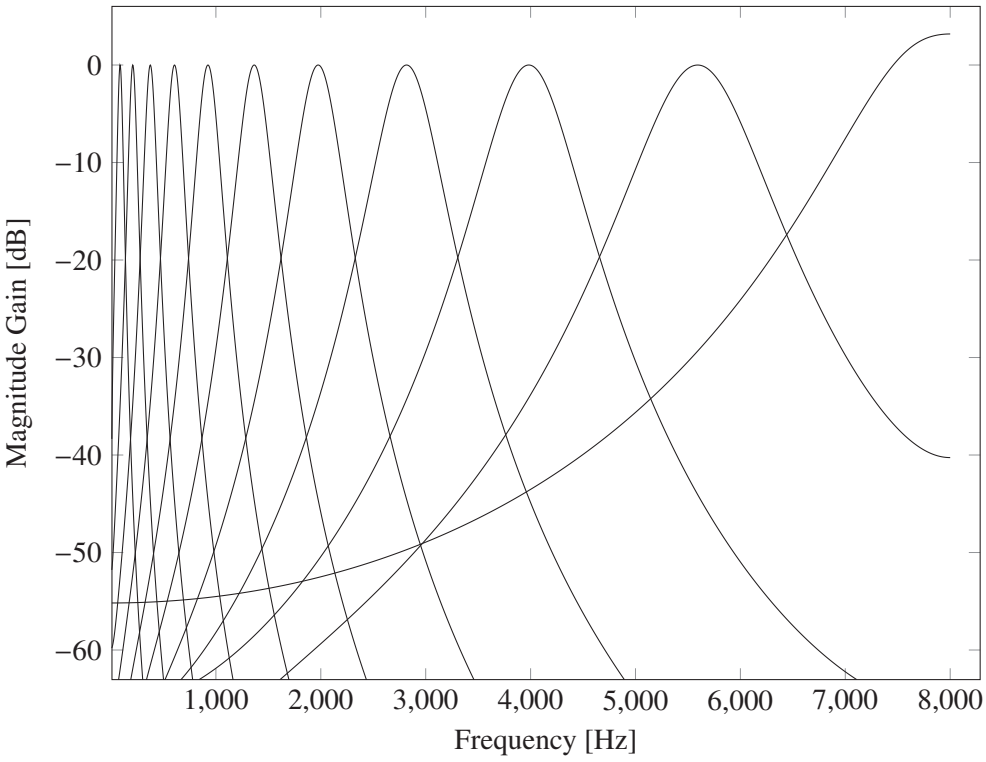

Figure 3.7: Frequency response of gammatone filterbanks

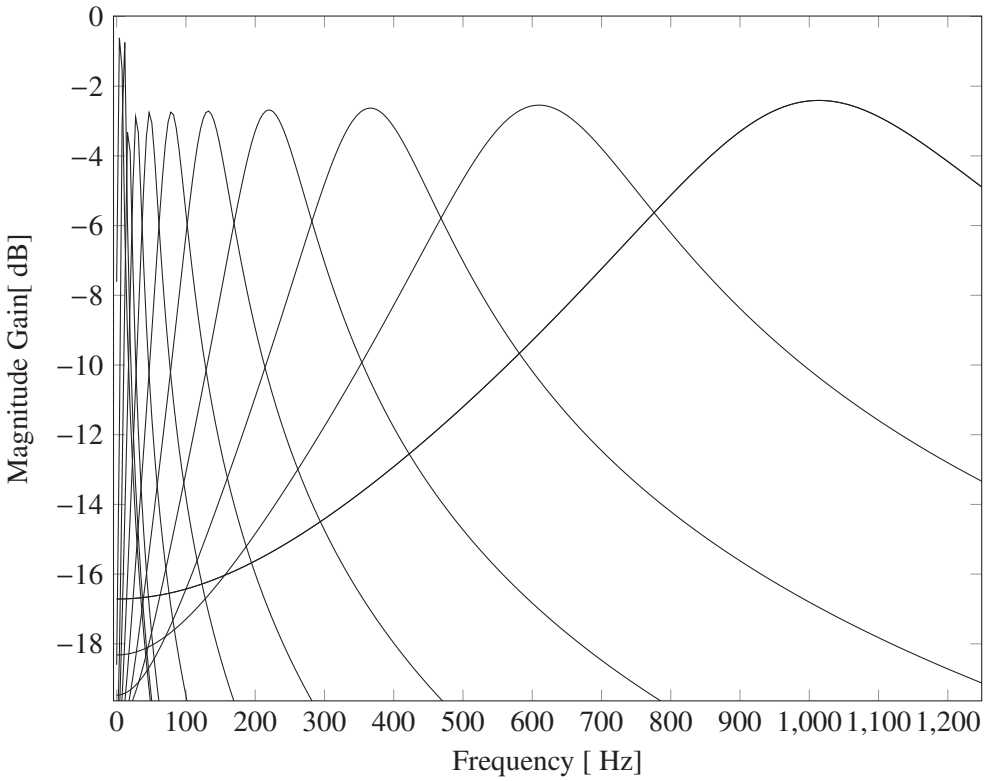

Figure 3.8: Frequency Response of Modulation Filterbank

employed. Popular emotional speech databases are introduced. EMO-DB, Polish database and SAVEE are used to test the developed speech emotion recognition algorithm. Human auditory system is presented. Transformations occurring in auditory system is exhibited. Corresponding computational model for human auditory system is explained. Evolution of computational auditory models are presented. Computational model is used in feature extraction for the classification of the emotions.
In table 3.1, previous work made in speech emotion recognition topic is provided. In the table, accuracy of the results, machine learning algorithms, database and feature set is given. This table provides a detailed comparison between the developed algorithms.

Table 3.1: Automatic Speech Emotion Recognition Background Work

Reference	Corpus	Feature Set	Classification Method	Emotions	Accuracy
(Petrushin, 2000)	140 utterances per emotional state at 22-kHz/16 bit.	Pitch, vocal energy, frequency spectral features, formants, speech rate and pausing	Neural Network	Happiness, anger,sadness, fear, and neutral	70%
(Schuller, Rigoll, & Lang, 2003)	From five speaker 5250 samples in German and English	Low level contours	Hidden Markov Models	Surprise, joy,anger, fear, disgust,sadness, neutral	77.8%
(M. M. H. E. Ayadi, Kamel, & Karray, 2007)	Berlin emotional speech database	12 mel frequency cepstrum coefficient MFCC, 12 delta coefficients, 0th cepstral coefficient, and the speech energy.	Gaussian mixture vector autoregressive model	Anger, boredom, fear, happiness,sadness, and neutral	76%
(Nogueiras, Moreno, Bonafonte, & no, 2001)	Spanish INTERFACE Database	Instant values and contours of pitch and energy.	Hidden semi continuous Markov models	Anger, disgust, fear, joy, sadness and surprise, neural	70%
(Nwe et al., 2003)	Six Burmese and six Mandarin speakers generated 720 utterances	Logarithmic Frequency Power Coefficients	Hidden Markov Model	Anger, Disgust, Fear, Joy, Sadness and Surprise	77.1%
(Tawari & Trivedi, 2010)	Berlin emotional speech database	Speech intensity,pitch and speaking rate, MFCC	Support Vector Machine	Surprise, joy,anger, fear, disgust, sadness, neutral	84%
(Ang et al., 2002)	Collected over the telephone and sampled at 8 kHz.	Duration and speaking rate features, pause features, pitch features,energy features, and spectral tilt features	Decision tree	Neutral, annoyed, frustrated, tired, amused	71.7%
(S. Wu, Falk, & Chan, 2009)	Berlin emotional speech database	Spectro-temporal features, trajectories of pitch and intensity,	Support Vector Machine	Surprise, joy,anger, fear, disgust, sadness, neutral	88.6%

Continued on next page

Table 3.1 – continued from previous page

Reference	Corpus	Feature Set	Classification Method	Emotions	Accuracy
(Iliou & Anagnostopoulos, 2010)	Berlin emotional speech database speaker dependent	133 features from pitch, mel frequency cepstral coefficients, energy and formants.	probabilistic neural Network	Surprise, joy,anger, fear, disgust, sadness, neutral	94%
(Cichosz & Slot, 2007)	Berlin emotional speech database (independent)	Regression parameters of pitch and mean energy in low frequency sub-bands	Decision Tree	Surprise, joy,anger, fear, sadness, neutral	72%
(Casale, Russo, Scebba, & Serrano, 2008)	Berlin Database of Emotional Speech	15 log energy coefficient, the 12 cepstral coefficients C1 C12, the pitch period, and the voicing class.	Support Vector Machine	Surprise, joy,anger, fear, disgust, sadness, neutral	92%

CHAPTER 4

Speech Emotion Recognition Using Auditory Models

In the previous section, human auditory system and computational auditory model is presented. As mentioned previously, human auditory system composed of two stages which are auditory preprocessing stage and speech pattern recognition. Preprocessing stage occurs in our ears. On the other hand speech pattern recognition process is made in the brain. Speech emotion recognition task is handled in this place too.
Although computational model of auditory system represent the output of the ears, there are very little cues about the process in our brain. In the scope of this thesis, this process is tried to be modeled using pattern recognition algorithms. General pattern recognition algorithms have two stages which are feature extraction and classification stages. Feature extraction part is the most crucial part of pattern recognition task. Extracted features are expected show similarities for same class, on the other hand, for different class case; they are expected to be very distinctive. In this thesis, auditory model outputs are used in feature extraction part. Bodily changes affect the speech, and human ear is the preprocessor of this signal. Since emotion processing is related with the perceptual phenomena, it is expected that, in a true computational model, algorithm will provide more accurate results. In this point of view, computational auditory models are compared with each other. It is expected that, accuracy of the results should exhibit an increase when (Jepsen et al., 2008) is compared with (Dau et al., 1997) since (Jepsen et al., 2008) is an updated version of (Dau et al., 1997).
In classification task, principal component analysis (PCA) and support vector machines (SVM) are employed to classify extracted features. Distribution of features is important in selection of classification method. In the classification stage, emotions are tried to segment into seven classes.

4.1 Feature Extraction Using Auditory Model Outputs

Used auditory model composed 6 main stages (Jepsen et al., 2008). First stage, outer middle ear transformation stage attenuates the frequency component below 1 kHz and above 5 kHz. In the second stage, 31 channel auditory filterbank is applied to the signal. Therefore at the end of this stage, 31 signals which are processed with DRNL filters are obtained. Each DRNL filters' center frequency changes with the channel. Center frequency of DRNL filters are scaled in ERB scale. Each signal subject to an envelope, expansion and adaptation processes. In the modulation filterbank, each of 31 signals is exposed to 12 channel modulation filterbank. An example of the processed signal is exhibited in figure 4.1. In (a) original speech signal "ay" is plotted. In(b) output signal of DRNL stage is given. In this plot, signal is processed by 10th channel of DRNL filter. As seen from the plot, since DRNL is a non-linear low pass filter, signal has lost some of its frequency components. In (c) envelope of the signal is plotted. As mentioned in the previous section, envelope extraction process is the half-wave

rectification followed by a low pass filter. As observed from the plot, signal resides on the upper side of 0 as a result of half-wave rectification. In (d), adaptation stage output is plotted. In (e) 4th channel modulation filter is applied to signal which is output of adaptation stage plotted in (d).
As mentioned in computational auditory model, there are 31 auditory filters with different center frequencies. Each of filtered signals is additionally processed by 12 channel modulation filterbank. Condition for applying a modulation filter to input signal, center frequency of modulation filter should not be greater than quarter of the center frequency of the auditory filterbank. Therefore first channel of the auditory filterbank can only be filtered by four modulation filterbank. 10th channel can be only filtered by first 8 modulation filterbank. At the end of this process, total of 283 modulation filtered signal exists. Each signals length is the same with the input signal. Each signal has a modulation and auditory filterbank channel which are going to be represented by one extracted feature value. Selected feature sets are:

- Mean of the signal
- Standard deviation of the signal

In feature extraction part, each signals mean is calculated. At the output, 283 length feature vector is obtained. Besides mean of the signal, standard deviation of the signal is also calculated. Fusion of mean and variance value constitutes the feature vector which is going to be used in classification method. In the figures 4.2, 4.4, 4.6, visualization of the anger, neutral and sad emotions' mean of the auditory model outputs are visualized. In anger emotion case, extracted feature in the modulation channels of 4 to 12, mean values deviation is not much. When neutral emotion in figure 4.4 is compared with sad emotion in figure 4.6, mean values of sad emotion in the modulation channel 4 to 6 is lower compared with neutral ones.
In the figures 4.3, 4.5, 4.7, visualization of the anger, neutral and sad emotions' standard deviation of the auditory model outputs are visualized. In anger emotion case, second modulation channel is high and nearly linear compared to others. On the other hand, neutral emotions second modulation channel drops as the acoustic channel increases. In sad emotion, second modulation channel value is high in low acoustic channels and drops instantly.
In the visualization of features, each emotion with each feature composed of 4 different plots. These plots correspond to different speaker with different sentences. Yet, in same emotion and same feature, their characteristics are similar. On the other hand, each emotion's features have specific characteristic. These properties made the extracted features very distinctive for speech emotion recognition task. Besides, recognition task is carried out speaker and text independent.

4.2 Automatic Speech Emotion Recognition Algorithm

4.2.1 Speech Emotion Recognition Using PCA

Principal component analysis is a variable reduction method, which transforms the variables that are highly correlated into a smaller number of uncorrelated variables. The number of principle components equal to or less the number of sample. It is based on the mechanism that if the dataset is going to be reduced into k dimensions then find k number of orthogonal eigenvector by maximizing the variance. Largest variance is named as first principal component. Preceding components holds the next highest variance. In this speech emotion recognition task, principal component analysis is used because of its simplicity and detection the most discriminating features within a vector. In this content, first five principle compo-

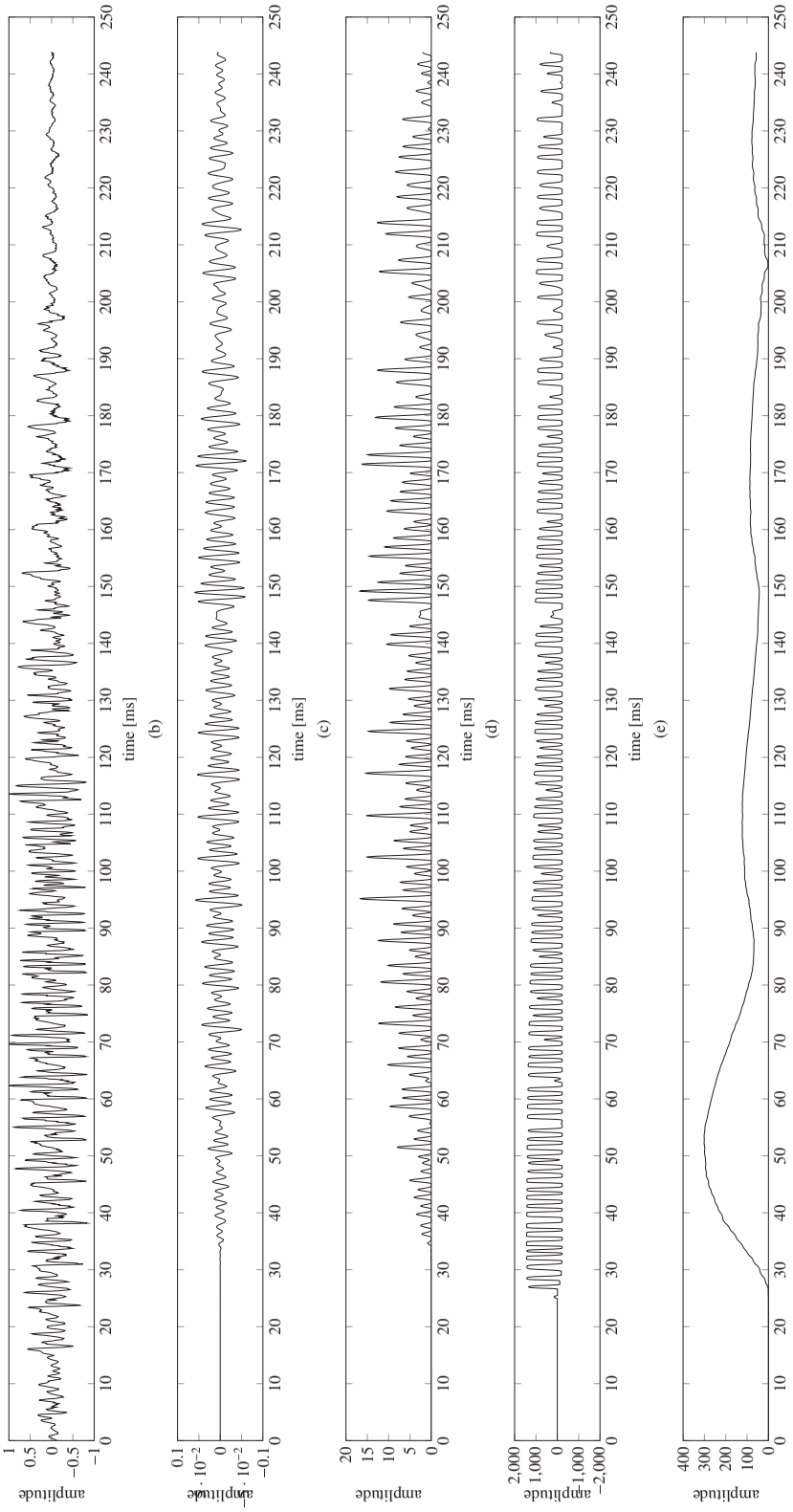

Figure 4.1: Audio Model Stages

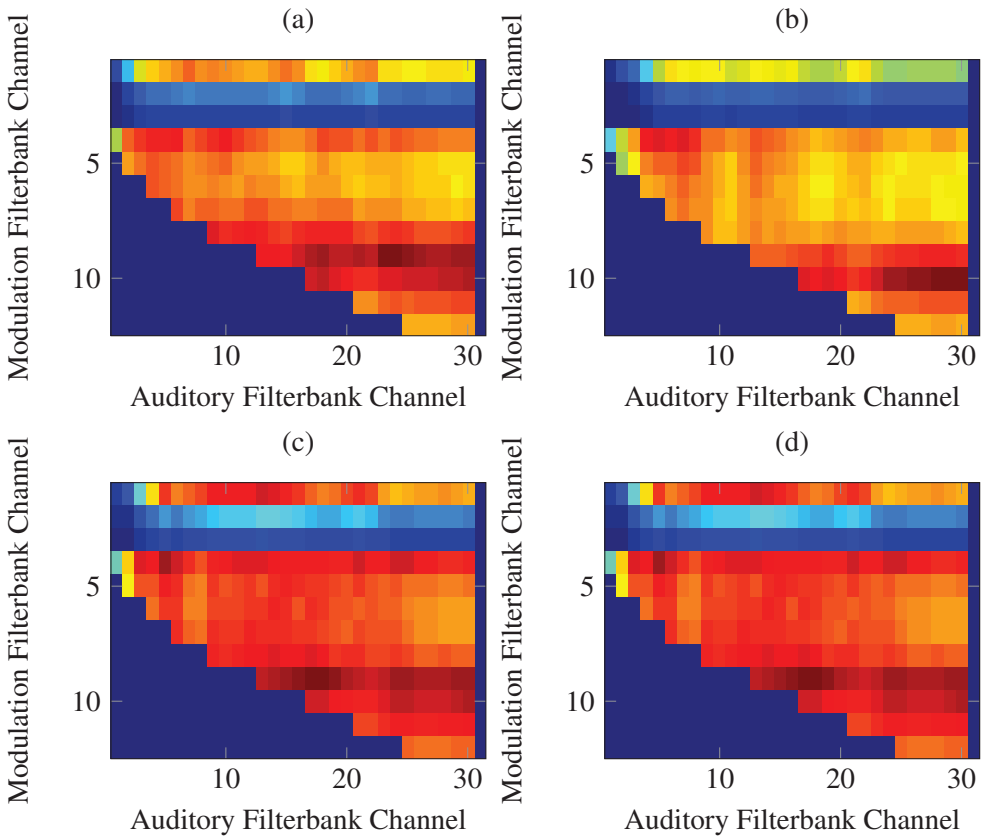

Figure 4.2: Anger speeches' mean of auditory model output
In the given figure, modulation filtered signals mean is projected into 'modulation filterbank channel' and 'auditory filterbank channel' space. Figures in 'a' to 'd' belongs to the emotion of anger. All speech signals are taken from Berlin emotional speech database. In (a) record is taken from speaker 3 while uttering "Das will sie am Mittwoch abgeben" which means "She will hand it in on Wednesday." In (b) same sentence is uttered by speaker 8. In(c) speaker 3 uttered "Heute abend könnte ich es ihm sagen." which means "Tonight I could tell him." In 'd' same sentence in 'c' is uttered by speaker 8.

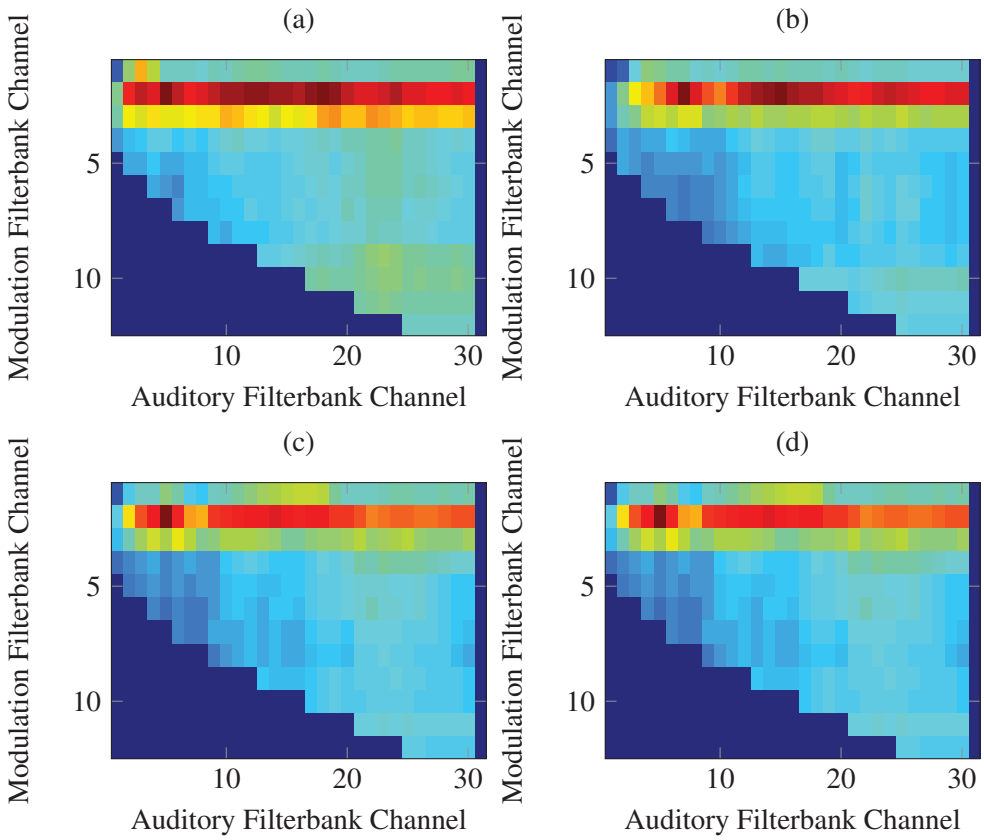

Figure 4.3: Anger speeches' standard deviation of auditory model output
In the given figure, modulation filtered signals standard deviation is projected into 'modulation filterbank channel' and 'auditory filterbank channel' space. Figures in 'a' to 'd' belongs to the emotion of anger. All speech signals are taken from Berlin emotional speech database. In (a) record is taken from speaker 3 while uttering "Das will sie am Mittwoch abgeben" which means "She will hand it in on Wednesday." In (b) same sentence is uttered by speaker 8. In(c) speaker 3 uttered "Heute abend könnte ich es ihm sagen." which means "Tonight I could tell him." In (d) same sentence in (c) is uttered by speaker 8.

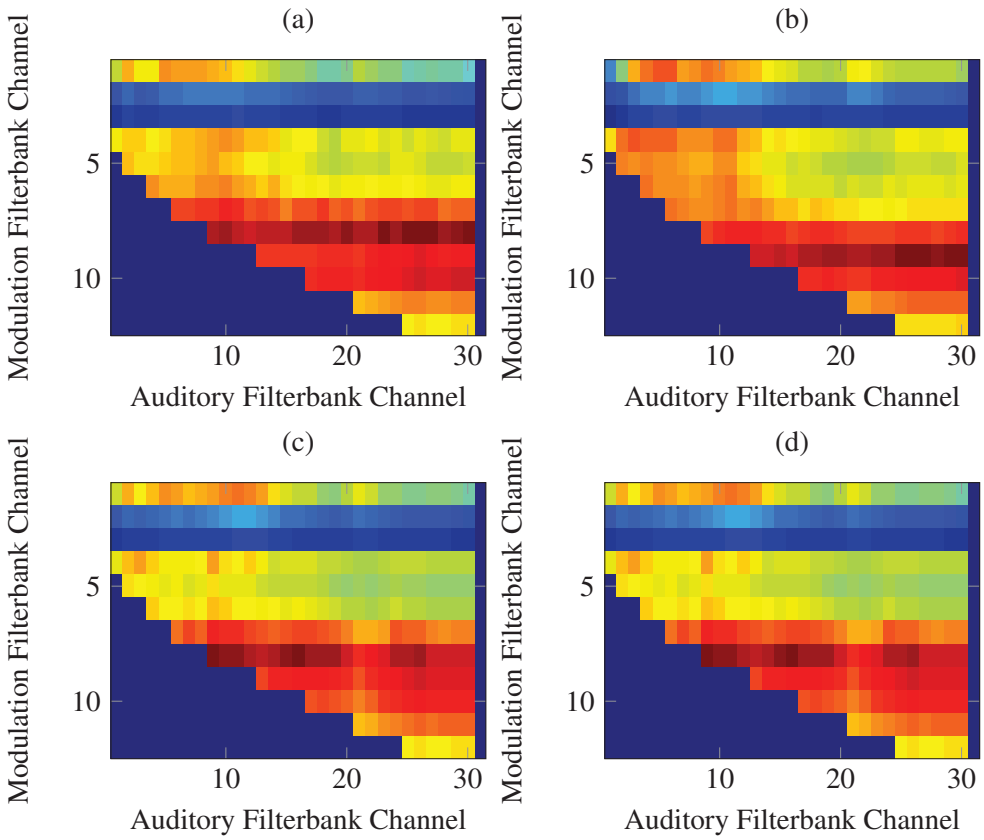

Figure 4.4: Neutral speeches' mean of auditory model output
In the given figure, modulation filtered signals mean is projected into 'modulation filterbank channel' and 'auditory filterbank channel' space. Figures in 'a' to 'd' belongs to the emotion of neutral. All speech signals are taken from Berlin emotional speech database. In (a) record is taken from speaker 3 while uttering "Das will sie am Mittwoch abgeben" which means "She will hand it in on Wednesday." In (b) same sentence is uttered by speaker 8. In(c) speer 3 uttered "Heute abend könnte ich es ihm sagen." which means "Tonight I could tell him." In (d) same sentence in (c) is uttered by speaker 8.

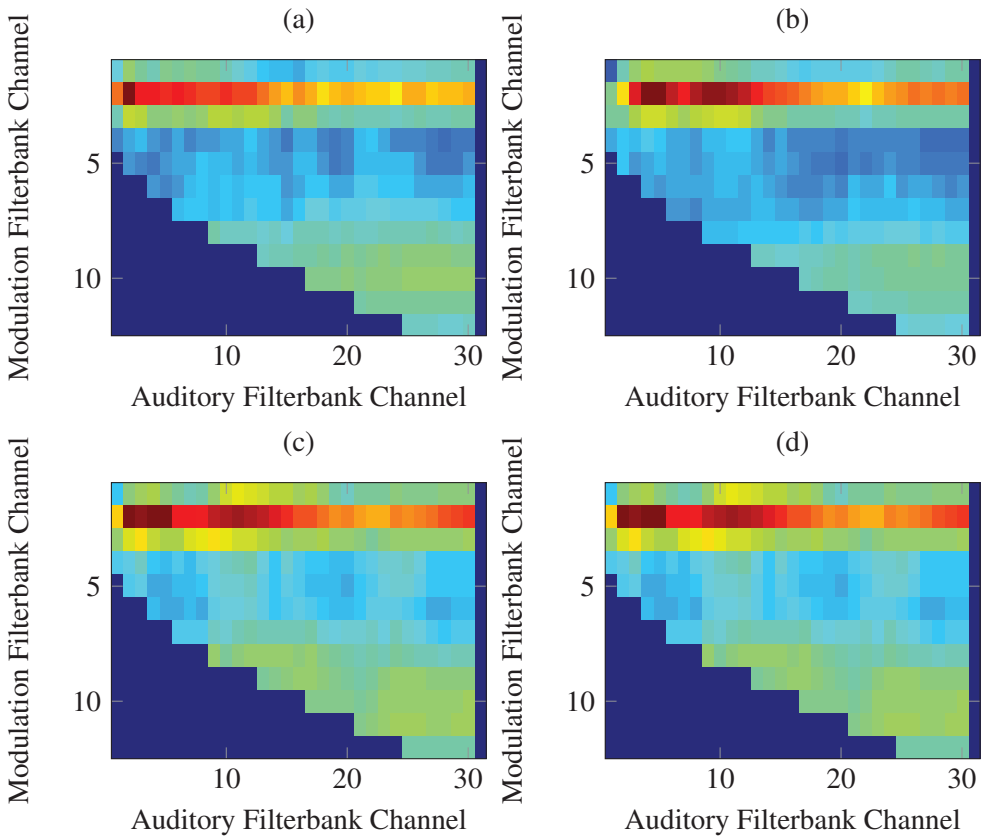

Figure 4.5: Neutral speeches' standard deviation of auditory model output
In the given figure, modulation filtered signals standard deviation is projected into 'modulation filterbank channel' and 'auditory filterbank channel' space. Figures in 'a' to 'd' belongs to the emotion of neutral. All speech signals are taken from Berlin emotional speech database. In (a) record is taken from speaker 3 while uttering "Das will sie am Mittwoch abgeben" which means "She will hand it in on Wednesday.". In (b) same sentence is uttered by speaker 8. In(c) speaker 3 uttered "Heute abend könnte ich es ihm sagen." which means "Tonight I could tell him." In (d) same sentence in (c) is uttered by speaker 8.

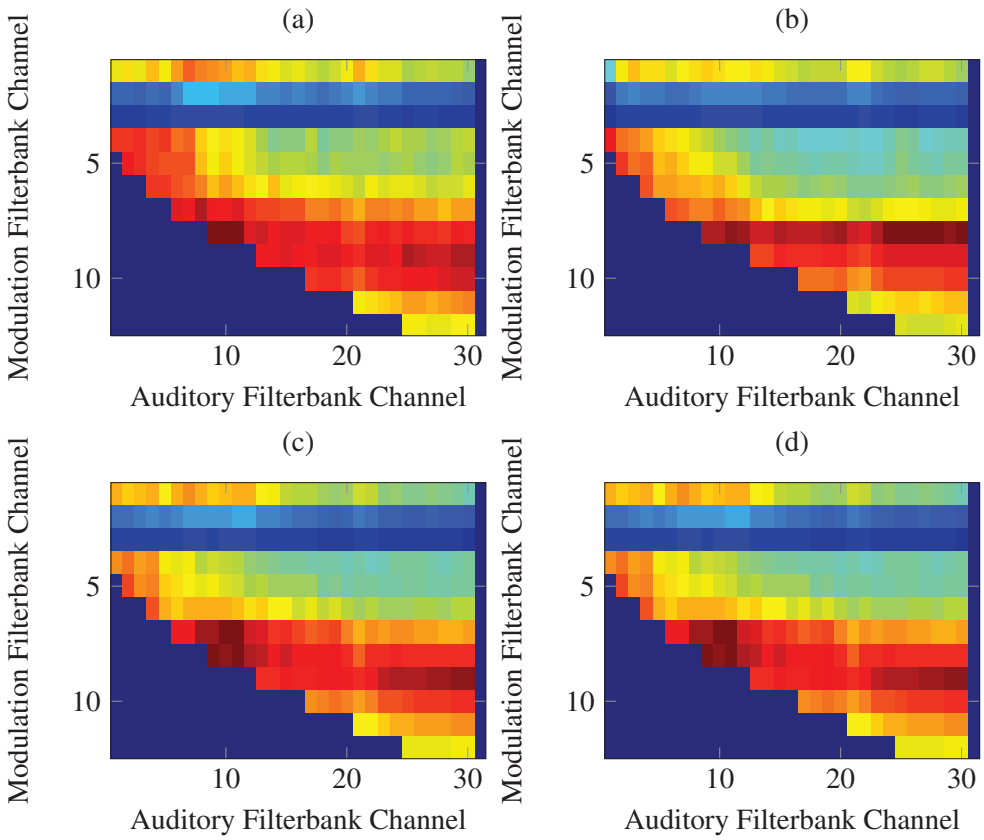

Figure 4.6: Sad speeches' mean of auditory model output
In the given figure, modulation filtered signals mean is projected into 'modulation filterbank channel' and 'auditory filterbank channel' space. Figures in 'a' to 'd' belongs to the emotion of sad. All speech signals are taken from Berlin emotional speech database. In (a) record is taken from speaker 3 while uttering "Das will sie am Mittwoch abgeben" which means "She will hand it in on Wednesday.". In (b) same sentence is uttered by speaker 8. In(c) speaker 3 uttered "Heute abend könnte ich es ihm sagen." which means "Tonight I could tell him." In (d) same sentence in (c) is uttered by speaker 8.

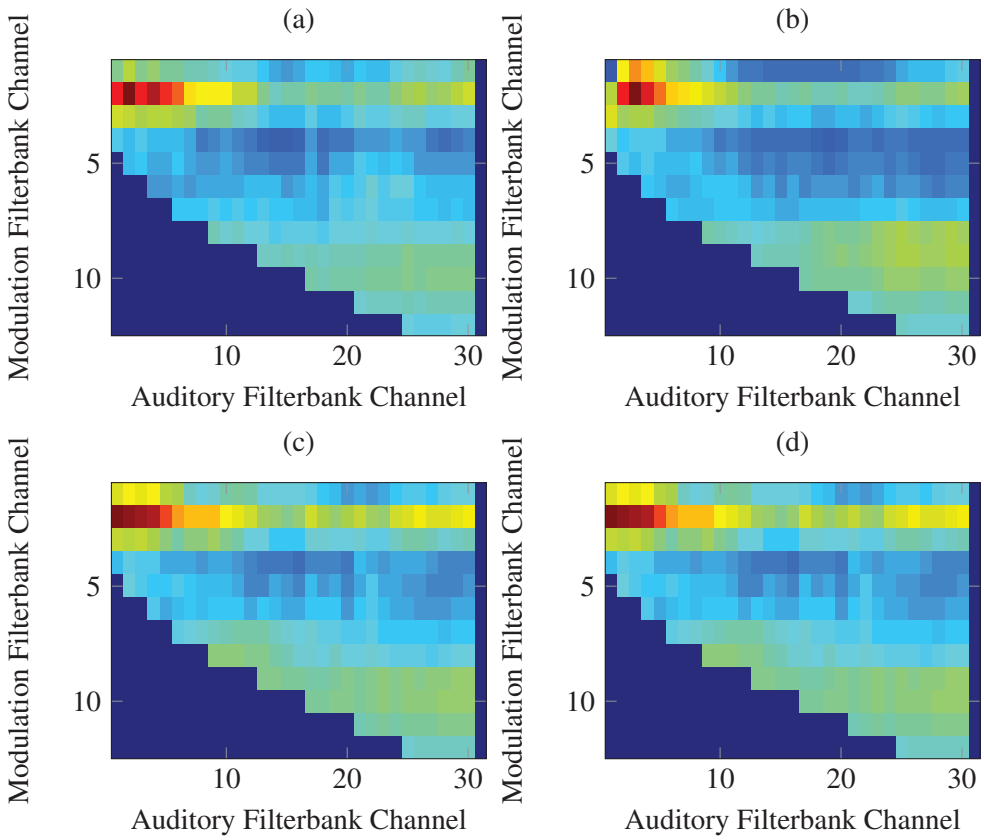

Figure 4.7: Sad speeches' standard deviation of auditory model output
In the given figure, modulation filtered signals standard deviation is projected into 'modulation filterbank channel' and 'auditory filterbank channel' space. Figures in 'a' to 'd' belongs to the emotion of sad. All speech signals are taken from Berlin emotional speech database. In (a) record is taken from speaker 3 while uttering "Das will sie am Mittwoch abgeben" which means "She will hand it in on Wednesday.". In (b) same sentence is uttered by speaker 8. In(c) speaker 3 uttered "Heute abend könnte ich es ihm sagen." which means "Tonight I could tell him." In (d) same sentence in (c) is uttered by speaker 8.

nents holds the most of data therefore, the dimension of vector is reduced to 5. The process of PCA is given below:

1. Extract feature vector from each sample.
2. Extract mean vector by averaging feature vectors.
3. Extract normalized vector by subtracting each extracted vector from the mean vector.
4. Extract covariance vector by multiplying inverse of normalized vector with normalized vector.
5. Extract eigenvector of covariance matrix.
6. Extract Eigenvectors of covariance matrix by multiplying of Eigenvector with normalized vector.
7. Extract Projection vector by multiplying inverse of Eigenvectors of covariance matrix with the normalized vector for each sample.

In classification, a binary tree is constructed. In each step, seven emotions are segmented into two groups. The reason of using a binary tree shape is because of the some disadvantage of PCA. Maximization of variance may result in loss of data. Therefore some distinctive properties of seven emotions may not be observed. Instead of classifying whole emotions in one step; in each branch principal components are re-calculated using the emotions which are going to be classified in this branch. Second reason is that, creation of a binary tree is necessary in order to use more superior classifiers such as SVM which is a binary classifier.
In order to generate a binary decision tree, Euclidean distances of principal component of each emotion is used. Since projected vectors of all samples may lose some of the discriminating features, in a binary tree most discriminating features are handled. In figure 4.8 (a), 7 principal components are given. Each principal component belongs to one emotion. For each emotion there are 40 samples. Each samples principal component is calculated then their mean is taken. In 4.8 (c), only happy and anger samples principal components are extracted. When (a) and (c) is compared, in (c), principal component of anger and happy is more distinctive than anger and happy emotions principal component in (a). In the table 4.1, extracted principal components of seven emotions' Euclidean distance to each other is given.
In the first stage, 7 emotions (anger, fear, happy, disgust, sad, neutral and boredom) are divided into 2 segments. First segment is named as excited emotions which are anger, happy, fear, disgust. Non excited emotions are sad, neutral and boredom. The reason why emotions labels are selected as excited and non-excited is that in emotion space in activation domain sad, boredom and neutral resides on the below half of the domain and other emotions resides on the upper half of emotion space.
To classify emotions into excited and non-excited segments, leave one out method is used in order to increment the number of train samples and to test all samples. In the method at first, projected images of train samples of both excited and non-excited emotions are calculated. Then test sample is first centered then projected using the previously calculated inverse of eigenvectors of test samples. Mean of projected vectors of excited and non-excited emotions are calculated separately. Then Euclidean distance to the first five projected vectors is found. Sample emotion belongs to the segment with the smallest distance. Results for the segmentation of excited and non-excited emotions are given in table 4.2 Overall segmentation result is estimated as 84.6%.
In the second stage, if emotion belongs to a non-excited segment, then second binary tree

	Anger	Happy	Disgust	Fear	Neutral	Boredom	Sad
Anger	0	0.9062	3.7920	3.6624	5.0299	6.9354	9.4076
Happy	0.9062	0	3.0660	2.8515	4.1760	6.1509	8.6490
Disgust	3.7920	3.0660	0	1.0200	2.1450	3.2064	5.6940
Fear	3.6624	2.8515	1.0200	0	1.6723	3.4045	5.8608
Neutral	5.0299	4.1760	2.1450	1.6723	0	2.6001	4.8822
Boredom	6.9354	6.1509	3.2064	3.4045	2.6001	0	2.6121
Sad	9.4076	8.6490	5.6940	5.8608	4.8822	2.6121	0

Table 4.1: Euclidean distances of each projected vector of emotions.

Figure 4.8: Principal components

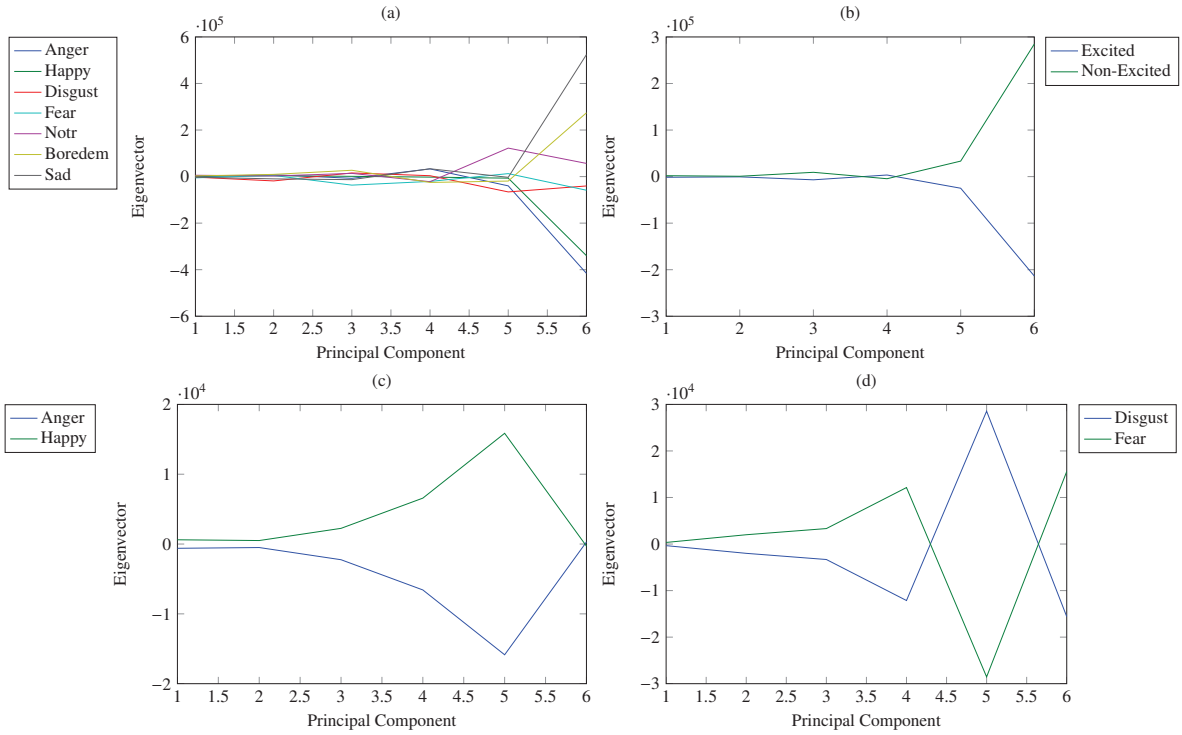

Table 4.2: Segmentation results of excited and non-excited emotions using PCA.

	Excited	Non excited
Excited	268	37
Non excited	45	185
Rate(%)	85.6	83.3

Table 4.3: Segmentation results of neutral and sad-boredom emotions using PCA.

	Notr	Sad-Boredom
Notr	62	26
Sad-Boredom	17	117
Rate(%)	78.5	81.8

is applied. In this classification case, emotion labels are again divided into two labels. First label is sad-boredom and second label is neutral (no emotion) case. In each branch, principal components are re-calculated. In this branch, mean normalization is applied using samples belong to the emotions of neutral, sad and boredom. As in the excited and non-excited classification, same procedure applies. Mean of projected images of neutral emotions are found and mean of projected images of sad-boredom emotions are calculated. Each samples extracted principal components Euclidean distance to each branches principal component is calculated. If distance to neutral emotion is smaller, then the sample's emotion is neutral, if not then move to the next branch to find whether the sample's emotion is sad or boredom. Results are given in the table 4.3.
Whether a sample is sad or boredom is detected using the same steps mentioned above. First calculate sad and bored emotions principal components, then extract the mean of principal components for both sad and boredom emotions. For the test sample, apply mean normalization then extract projected vector. Euclidean distance is used to find the distance to the each emotion. Segmentation result is provided in the table 4.4.
In the excited emotion side, there are 4 emotions which are anger, happy, fear and disgust. Projected vector distance illustrated in 4.8 (a), has shown that, distance between anger and happy is closer and distance between fear and disgust is closer. Therefore branch segments emotions into two labels which are happy-anger and fear-disgust. Success rates are provided in the table4.5.
Fear and disgust is separated using the procedure. Their success rates are given in the table 4.7. Happy and anger results are provided in the table 4.6. Happy and anger results are weak compared to separation of other emotions. Therefore, different classification methods for the segmentation of happy and anger emotions are searched. Extracted features and classification method for happy and anger emotions is demonstrated in the next section. Generated binary tree is visualized in figure 4.9. As seen from the figure, general algorithm requires the training of 6 different principal components.

4.2.2 Segmentation of Happy and Anger Using Spectral Features

Obtained classification results of happy and anger with PCA are low when compared with other emotions. This led to search for new strong features for happy and anger segmentation. Spectral shape features are selected to classify happy and anger. Features extracted from short time Fourier transform are named as spectral shape features. (Peeters, 2004) Spectral centroid

Table 4.4: Segmentation results of sad and boredom emotions using PCA.

	Boredom	Sad
Boredom	63	6
Sad	18	56
Rate(%)	77.8	90.3

Table 4.5: Segmentation results of anger-happy and fear-disgust emotions using PCA.

	Anger-Happy	Fear-Disgust
Anger-Happy	170	25
Fear-Disgust	28	90
Rate(%)	85.9	78.3

Table 4.6: Segmentation results of anger and happy emotions using PCA.

	Anger	Happy
Anger	78	25
Happy	49	46
Rate(%)	61.4	64.8

Table 4.7: Segmentation results of fear and disgust emotions using PCA.

	Fear	Disgust
Fear	54	8
Disgust	15	38
Rate(%)	78.3	82.6

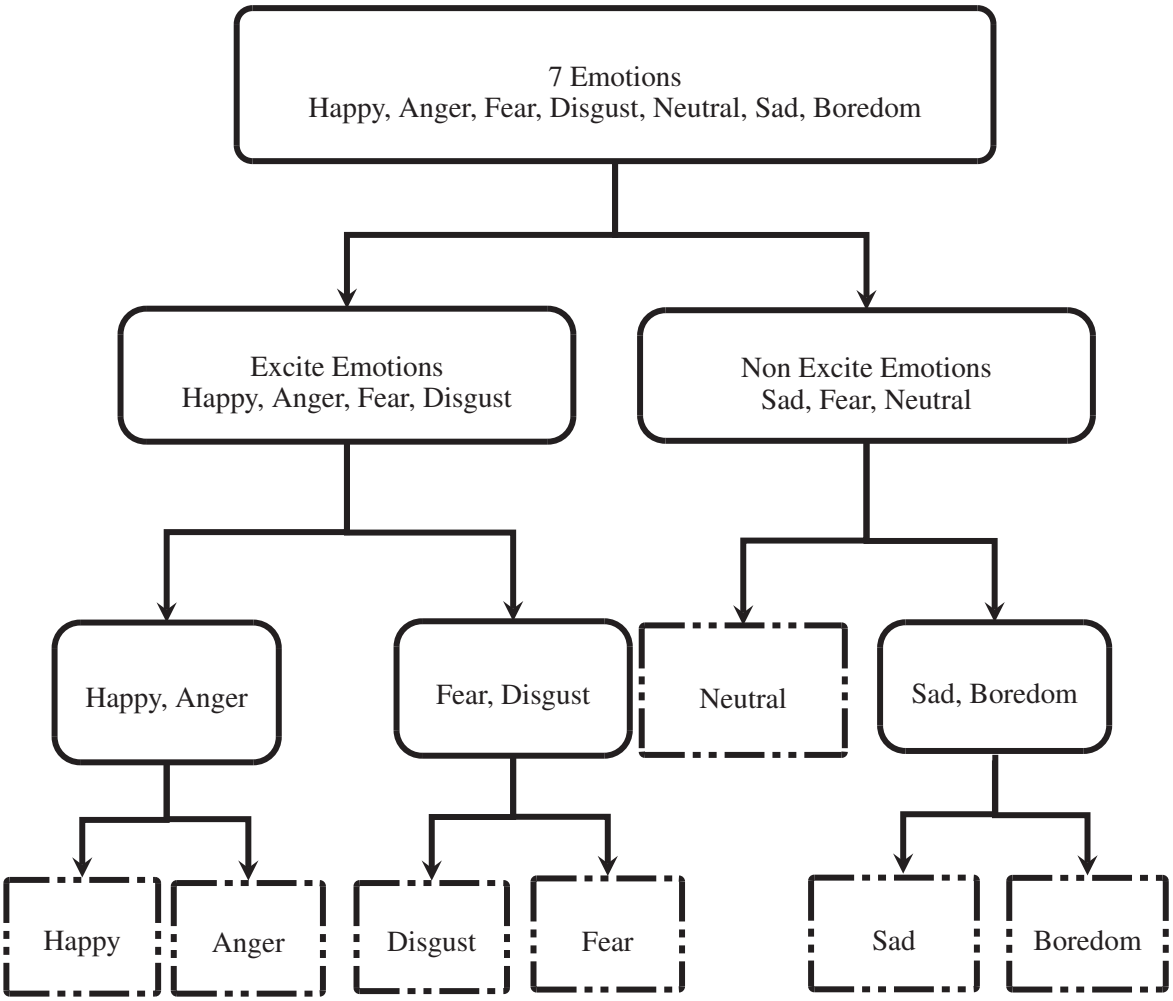

Figure 4.9: General algorithm flow chart.

and spectral mean are selected as a feature set in this classification task. Each speech signal is segmented into frames with length of 60 ms and overlap 20 ms. Extracted spectral features' distributions over frames are analyzed. Brightness of sound is the common term used for spectral centroid. Spectral centroid of a signal is the midpoint of its spectrum. Centroids are computed by taking average of the frequencies weighted by their amplitudes. Extraction of spectral centroid is applied to each overlapping window. Calculation of spectral centroid is as follows:

$$SpectralCentroid = \frac{\sum_{k=1}^{N} kF[k]}{\sum_{k=1}^{N} F[k]} \tag{4.1}$$

Spectral centroid features are originally proposed by for speech recognition systems. Spectral centroids have found a wide usage in the area of speech recognition systems because of its effectiveness. Spectral centroids have already been used for speech recognition systems, speaker recognition systems and cognitive load classification tasks (Le, Ambikairajah, Epps, Sethu, & Choi, 2011).

In addition to spectral centroid, spectral mean is also used. In estimation of spectral mean, each frame's spectrum's mean is taken. Spectrum mean is equal to averages of amplitudes in frequency domain. In this thesis, at first, each frames' energy is normalized to 1 then its spectral mean is calculated.

After calculation spectral mean and spectral mean for each frame, using statistical calculation, for each utterance two feature is obtained. Coefficient of variation is calculated for centroid vector. Standard deviation is calculated for mean vector. Coefficient of variation (CV) is equal to ratio of standard deviation to mean. CV is the variation as a percentage of mean. CV is useful in cases such as; standard deviation depends on the change of mean. In order to comparison two data sets with different mean and different units, CV provides more comprehensive analysis. In this content, centroid vectors coefficient of variance is used as a feature vector. Second feature is extracted by taking the mean of spectral means.

Discrimination of features are first inspected on same speaker same sentence case. Results have shown that for the mentioned case, CV of centroid is higher for anger compared to happy. Reverse case is observed for standard deviation of spectral mean. Distribution of features in two dimension space is given in figure 4.10. To classify happy and anger emotions, SVM is used. Hyperplane is drawn in the figure. To test the generated features, Berlin Emotional Speech Database is employed. Total of 196 happy and anger sample exists in the dataset. Half of the dataset is used for training. Other halves of the samples are used for classification. Since extracted features for training is not separable, quadratic kernel is used for the training of SVM. Highest success rate is obtained when quadratic kernel is used. When polynomial kernel is used, success rates drop. Auto scale is enabled in order to normalize the data points at their mean, and scaled them to have unit standard deviation, before training. Although non-linear kernels are used, still extracted features are not separable. With high order kernels, hyperplane is able to separate data, yet the success rate drops since train number is low. High order kernels could not constitute a hyperplane which fits to the data, when train number is not enough.

When half of the sample is used for training and other half is used for classification, success rate of classification is measured 77%. When leave one out method is used, success rate is measured as 76%.

4.2.3 Speech Emotion Recognition Using SVM

In developed speech emotion recognition algorithm, features are extracted using a state-of-art computational auditory model. Extracted features are classified using a generated binary tree

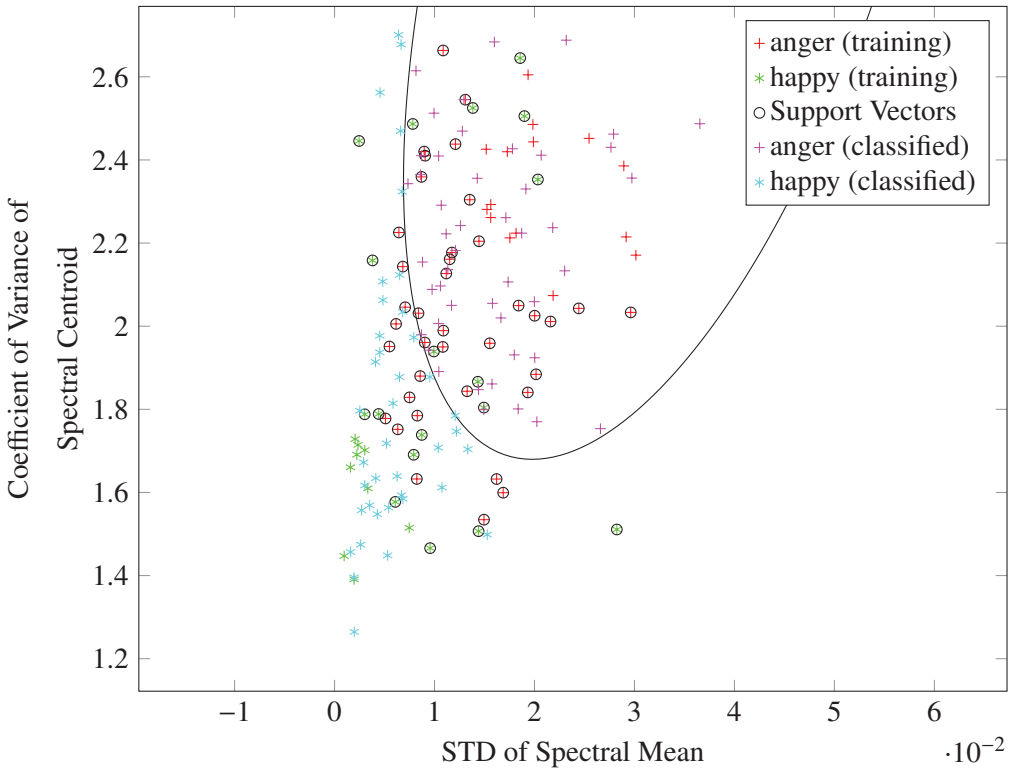

Figure 4.10: Spectral Features Classification with SVM

as given in the figure. In each branch, audio samples are classified into two segments. Segmentation is implemented using principal component analysis. In segmentation algorithm, audio samples first projected, then their distance to test sample is measured using Euclidean distance.

Used computational auditory model generated total of 286 modulation filtered signals. From each modulation filtered signal, 2 features are extracted which are signal's mean and signal's standard deviation. When two features are fused, total length of the feature vector becomes 572. At first, extracted feature vector size is reduced using component analysis. PCA transforms correlated variable into smaller number of uncorrelated variable. PCA is not a classification method, in fact a variable reduction method. On the other hand, SVM is a binary classification algorithm which separates the data finding a hyperplane that maximizes margins. In this section, given binary tree is applied with SVM instead of PCA.

Binary tree given in figure 4.1 is process by six different SVM's. Each SVM is trained separately. SVM kernels were set as linear kernel. First branch is the classification of excited and non-excited emotions. Table results are derived with leave-one-out method on Berlin Speech Emotion Test. In table 4.8 segmentation results of excited, non-excited emotions is provided. Although PCA success rate was measured as 84.67%, classification with SVM has a higher success rate with 97.16%. Second branch is segmentation of neutral and sad-boredom emotions. In table 4.9, success rate of neutral, sad-boredom is given. Third branch is the segmentation of sad and boredom emotions. Rates are given in table 4.10. Segmentation of excited emotions is processed in two levels. First segmentation is happy-anger and fear-disgust. Results are given in table 4.11. Success rate of fear-disgust is given in table 4.13. Final branch is the segmentation of happy and anger emotions. Classification rate is given in table 4.12 which is equal to 87%. Compared with other algorithms, segmentation

Table 4.8: Segmentation results of excited and non-excited emotions using SVM.

	Excited	Non excited
Excited	302	4
Non excited	11	218
Rate(%)	96.4	98.2

Table 4.9: Segmentation results of neutral and sad-boredom emotions using SVM.

	Notr	Sad-Boredom
Notr	71	12
Sad-Boredom	8	131
Rate(%)	91.2	91.6

with PCA has a success rate of 62%. In the previous section, spectral features are extracted. Extracted features are classified with a success rate of 76%. These results have shown that, SVM provides higher classification accuracy. In addition to that features extracted from auditory model are discriminant to segment happy and anger than extracted short time spectral features.

Table 4.13: Segmentation results of fear and disgust emotions using SVM.

	Fear	Disgust
Fear	66	2
Disgust	3	44
Rate(%)	95.6	95.6

4.3 Summary

In this chapter, developed speech emotion recognition algorithm is presented. How computational auditory models outputs are used to extract features is explained. Two simple features are extracted from each modulation filtered signal. Visualization of the features in modulation filterbank channel and auditory filterbank channel is provided. Extracted features are at first classified using PCA. To classify emotions a binary tree is generated. To construct binary tree, extracted projected vector distance in Euclidean space is used. With the same manner, train emotional speeches projected vectors are used to classify emotions in each branch. In the tables, classification accuracy of each branch is presented. Overall classification accuracy directly depends on the classification made on each branch. Since classification rate of happy and anger emotion is very low, different feature sets are search. Different from extracting features from auditory model, short time spectral features are investigated. Spectral centroid and spectral mean of the speech signals has proved to be distinctive in the segmentation of happy and anger emotions. Coefficient of variation of spectral centroid vector and standard

Table 4.10: Segmentation results of sad and boredom emotions using SVM.

	Boredom	Sad
Boredom	78	2
Sad	3	60
Rate(%)	96.2	96.7

Table 4.11: Segmentation results of anger-happy and fear-disgust emotions using SVM.

	Anger-Happy	Fear-Disgust
Anger-Happy	186	12
Fear-Disgust	12	103
Rate(%)	93.9	89.5

Table 4.12: Segmentation results of anger and happy emotions using SVM.

	Anger	Happy
Anger	118	16
Happy	9	55
Rate(%)	92.9	77.46

deviation of energy normalized spectral means are used as features. Extracted features are classified by a quadratic kernel SVM. Extracted results are higher when compared with the results obtained when PCA is used. Results have shown that, spectral features provides important cues for the classification of happy and anger.
In the next section, SVM is used to classify emotions. Since SVM is a binary classification algorithm, previously generated binary tree is employed. SVM is used with linear kernel. Since in the binary tree there are six branches, seven different SVMs are trained. In tables, success rate of classification of each branch is provided. Results provides an important feedback about the that process in the auditory system generates important cues for emotions. Support vector machine is successfully classify emotions even though simple features such as mean and standard deviation is used to extract features from auditory model. When PCA and SVM is compared in classification task, accuracy rate of SVM is higher in each branch. Besides, extracted features from auditory model is more distinct when compared with short time spectral features although SVM is used in both classification task of happy and anger emotions. In the results chapter, performance of the overall algorithm with SVM is provided. Algorithm is tested on three emotional databases whose languages are German, English and Polish.

CHAPTER 5

Results

In the results section, subjective speech emotion recognition performance is presented. Besides, subjective emotion recognition performance, developed automatic speech emotion recognition algorithms rates are exhibited. Developed algorithm is tested on three different databases. Algorithms' speaker and language dependence/independence cases are measured and exhibited in this section. In the comparison part, subjective performance and automatic recognition performance are compared.

5.1 Subjective Speech Emotion Recognition Performance

In subjective speech emotion recognition test, using only acoustic properties, human listeners recognition performance is measured. Listening test speech samples are selected from German EMO-DB. In order to test the recognition using only acoustic properties of speech, listeners are selected such that non of the listeners do not know any German. As a result, evaluators are only exposed to acoustic properties of speech samples. In that way, human listeners and developed algorithms performance is compared under equal conditions. Extracted recognition rates has provide a comparison for the automatic recognition rates. Test was carried on 4 male and 6 female listeners. Age of male listeners are 23, 23, 20 and 19. Age of female listeners are 49, 25, 20, 20 and 17. Listening test was applied in a quiet environment using the same test components which are a computer and a headphone. In listening test, each evaluator is requested to detect the emotion of 10 speech samples from 7 emotions. Selected emotion labels are happy, anger, fear, sad, disgust, neutral and boredom. Test is applied multiple chose style. In the multiple choices, besides emotion labels, "can't decide" choice is included in case listeners could not decide on one of the choices. It is allowed that, during the test, a speech sound could be listened as long as listener desire. It is not allowed that, without selecting a choice, listener could not pass to the next speech sound. Snapshot of the listening test is provided in B.1. In listening test, total of 70 emotional speech sound is tested. Duration of test varied six to ten seconds. Subjective recognition results are given in table 5.1. Subjective test results have shown that, total recognition rate is 58.4%. Anger has the highest emotion rate with 87.0%. On the other hand, recognition of happy has the lowest rate of 44.4%. Highest score of a listener acquired is 78.57 %. Measured lowest recognition rate of a listener is 37.14%. Each listeners' recognition rates for each emotion and total recognition rates are given in table 5.2.
In subjective emotion recognition test, context independent emotion recognition performance of human listeners is measured. Among seven emotions, anger has proven to be the most distinguishable emotion. On the other hand, happy is measured as the most difficult emotion to discriminate. Recognition rate of female listeners is higher (62.1%) compared to male listeners(52.8%). Recognition rate of boredom, happy and sad emotions are discriminable higher

Table 5.1: Subjective test results.

	Anger	Boredom	Disgust	Fear	Happy	Sad	Neutral	Precision(%)
Anger	**87**	0	6	14	13	0	3	70.3
Boredom	2	**58**	13	1	8	22	12	50.0
Disgust	3	3	**47**	4	2	1	4	73.44
Fear	0	1	10	**52**	10	5	2	65.00
Happy	3	2	5	14	**40**	0	6	57.14
Sad	2	13	15	7	5	**67**	5	58.77
Neutral	1	30	3	3	9	4	**58**	53.70
Cant Decide	2	3	1	5	3	1	10	
Rates (%)	87.0	52.7	47.0	52.0	44.4	67.0	58.0	**58.4**

Table 5.2: Each listeners' recognition rates.

	Anger	Boredom	Disgust	Fear	Happy	Sad	Neutral	Rates (%)
F1	0.9000	0.8182	0.5000	0.3000	0.4444	0.7000	0.8000	0.44
F2	0.6000	0.5455	0.4000	0.3000	0.5556	0.7000	0.4000	0.50
F3	1.0000	0.7273	0.6000	1.0000	0.2222	0.7000	0.5000	0.68
F4	1.0000	0.5455	0.5000	0.4000	0.2222	1.0000	0.6000	0.61
F5	0.9000	0.2727	0.4000	0.6000	0.3333	0.5000	0.4000	0.40
F6	1.0000	0.5455	0.5000	0.8000	0.6667	1.0000	1.0000	0.78
M1	1.0000	0.6364	0.5000	0.6000	0.3333	0.7000	0.6000	0.62
M2	0.9000	0.2727	0.7000	0.6000	0.7778	0.5000	0.7000	0.62
M3	0.8000	0.5455	0.3000	0.3000	0.7778	0.6000	0.2000	0.50
M4	0.6000	0.3636	0.3000	0.3000	0.1111	0.3000	0.6000	0.37

for females. Recognition performance of male subjects are given in table 5.4. Recognition performance of female subjects are given in table 5.3.

5.2 Automatic Speech Emotion Recognition Performance

Developed speech emotion recognition algorithm is applied on three emotional databases which are EMO-DB, SAVEE and Polish database. In that way, algorithm is tested on three different languages namely, German, English and Polish. Given results in tables are extracted using the algorithm given in section 4.2.3.

Results in table 5.5 is obtained using the speech samples in EMO-DB. In EMO-DB, total of 535 emotional speech samples exists. In order to test the algorithm, leave one sample out method is applied. Each time, 534 speech samples are used for training and 1 speech sound is used to test. All speech sounds in EMO-DB has become the train sample. Each SVM is re-trained in binary tree every time test sample changes. In this task, 7 emotions are classified with a average recognition rate of 82.9%. Sad emotion has highest recognition rate of 91.9% and happy emotion has the lowest recognition rate of 63.4%.

When automatic speech emotion recognition algorithm is tested on Polish database with leave one speech sample out method, average recognition rate is measured as 71.3% as given in table 5.6. In the segmentation of Polish emotion speech database, 6 emotions(anger, fear, happy, boredom, sad, neutral) are segmented. Anger and fear emotions have the lowest recognition rate of 62.5%. Neutral emotion has the highest recognition rate of 82.5%.

Table 5.3: Female listeners subjective test results.

	Anger	Boredom	Disgust	Fear	Happy	Sad	Neutral	Precision(%)
Anger	**53**	0	4	8	8	0	2	70.6
Boredom	0	**41**	6	0	2	11	10	58.5
Disgust	3	1	**28**	3	1	0	2	73.6
Fear	0	1	7	**31**	7	2	1	63.2
Happy	3	1	4	8	**26**	0	3	57.7
Sad	1	10	10	4	3	**47**	4	59.4
Neutral	0	12	0	3	7	0	**35**	61.4
Cant Decide	0	0	1	3	0	0	3	
Rates (%)	88.3	62.1	46.6	51.6	48.1	78.3	58.3	**62.1**

Table 5.4: Male listeners subjective test results.

	Anger	Boredom	Disgust	Fear	Happy	Sad	Neutral	Precision(%)
Anger	**34**	0	2	6	5	0	1	70.8
Boredom	2	**17**	7	1	6	11	2	36.9
Disgust	0	2	**19**	1	1	1	2	73.8
Fear	0	0	3	**21**	3	3	1	67.7
Happy	0	1	1	6	**14**	0	3	56.0
Sad	1	3	5	2	2	**20**	1	57.14
Neutral	1	18	3	0	2	4	**23**	45.10
Cant Decide	2	3	0	2	3	1	7	
Rates (%)	85.0	38.6	47.5	52.5	38.9	50.0	57.5	**52.8**

Table 5.5: Automatic speech emotion results using EMO-DB with leave one speech sample out.

	Anger	Fear	Happy	Disgust	Sad	Neutral	Boredom	Rate (%)
Anger	**116**	2	9	0	0	0	0	91.3
Fear	1	**53**	8	2	1	3	1	76.8
Happy	15	8	**45**	1	0	1	1	63.4
Disgust	1	2	2	**37**	0	0	4	80.4
Sad	0	0	0	0	**57**	3	2	91.9
Neutral	0	1	0	2	2	**68**	6	86.1
Boredom	0	0	0	1	3	9	**68**	83.9
Precision	87.2	80.3	70.3	86.0	90.5	80.9	82.9	82.9

Table 5.6: Automatic speech emotion results using Polish database with leave one speech sample out method.

	Anger	Fear	Happy	Sad	Neutral	Boredom	Rate (%)
Anger	**25**	3	8	2	2	0	62.5
Fear	7	**25**	1	2	3	2	62.5
Happy	11	1	**26**	0	2	0	65.0
Sad	0	5	0	**31**	0	4	77.5
Neutral	1	2	0	4	**33**	0	82.5
Boredom	0	2	2	4	1	**31**	77.5
Precision	56.8	65.8	70.3	72.1	80.5	83.8	71.3

Third database is an English database named as SAVEE. Leave one speech sample out method is used in the recognition task. Results are given in table 5.7. Overall recognition rate is measured as 73.81%. Six emotions are segmented which are anger, fear, happy, sad, neutral and disgust. Highest recognition rate is obtained on emotion label neutral and lowest emotion recognition rate is obtained on disgust emotion with the given rates 87.5% and 60.0% respectively.

In order to measure the speaker independence, leave one speaker out method is applied. EMO-DB is constituted from ten speakers' speeches. In Polish database, eight speakers exists. In SAVEE database, there are four speakers. In leave one speaker out method, train speech sounds are selected from the speech sounds which do not belong to the train samples speaker. In table 5.8, speaker independent results are provided. Recognition rate of EMO-DB

Table 5.7: Automatic speech emotion results using SAVEE database with leave one speech sample out method.

	Anger	Fear	Happy	Disgust	Sad	Neutral	Rate (%)
Anger	**46**	3	7	4	0	0	76.6
Fear	6	**38**	11	0	4	1	63.3
Happy	6	10	**39**	5	0	0	65.0
Disgust	5	2	2	**36**	8	7	60.0
Sad	0	4	0	5	**46**	5	76.6
Neutral	2	1	0	7	5	**105**	87.5
Precision	70.7	65.5	66.1	63.1	73.2	88.9	73.81

Table 5.8: Automatic speech emotion results using EMO-DB with leave one speaker out method.

	Anger	Fear	Happy	Disgust	Sad	Neutral	Boredom	Rate (%)
Anger	**109**	2	15	0	0	0	0	85.8
Fear	2	**43**	13	0	2	8	1	62.3
Happy	27	10	**31**	1	0	2	0	43.6
Disgust	1	3	4	**33**	1	1	3	71.7
Sad	0	1	0	0	**51**	7	3	82.2
Neutral	0	2	0	3	2	**64**	8	81.0
Boredom	0	1	0	8	5	11	**56**	69.14
Precision	78.4	68.2	49.2	73.3	83.6	68.8	78.8	72.3

Table 5.9: Automatic speech emotion results using Polish Database with leave one speaker out method.

	Anger	Fear	Happy	Sad	Neutral	Boredom	Rate (%)
Anger	**25**	2	9	2	2	0	62.5
Fear	9	**19**	2	6	1	3	47.5
Happy	14	3	**20**	0	3	0	50.0
Sad	1	4	0	**20**	6	9	50.0
Neutral	1	4	1	6	**25**	3	62.5
Boredom	0	4	1	3	6	**26**	65.0
Precision	50.0	52.7	60.61	54.5	58.14	63.41	56.25

is measured as 72.13%. Speaker independent results for Polish database is given in table 5.9. Recognition rate is measured as 56.25% which is low compared to recognition rates given in table 5.6. Highest recognition rate is measured on boredom emotion with a rate of 65.0% and lowest recognition rate is measured in the segmentation of fear with a rate of 47.4%.

Third recognition method other than leave one sample out and leave one speaker out is the test of algorithm on speaker dependent level. In table 5.10, speaker dependent results are given for EMO-DB. In speaker dependent test, train speech sound samples are selected from the same speaker of the test speech sample. Measured speaker dependent recognition rate is 76.26%. Speaker dependent test is applied on all 10 speakers of EMO-DB.

In this recognition test, auditory model's performance is tested using leave one speech sample out. Although classification is the same with others, auditory model (Dau et al., 1997) used in

Table 5.10: Speaker dependent automatic speech emotion results using EMO-DB

	Anger	Fear	Happy	Disgust	Sad	Neutral	Boredom	Rate (%)
Anger	**104**	4	15	2	0	2	0	81.8
Fear	4	**51**	6	6	0	1	1	73.9
Happy	17	6	**45**	1	0	1	1	63.3
Disgust	3	9	5	**24**	0	1	4	52.17
Sad	0	0	0	0	**57**	0	5	91.9
Neutral	0	2	0	2	1	**68**	6	86.0
Boredom	0	2	1	1	8	10	**59**	72.8
Precision (%)	81.3	68.9	62.5	66.7	86.4	81.9	77.6	**76.26**

Table 5.11: Automatic speech emotion results using EMO-DB using the auditory model of (Dau et al., 1997)

	Anger	Fear	Happy	Disgust	Sad	Neutral	Boredom	Rate (%)
Anger	**90**	10	25	2	0	0	0	70.9
Fear	4	**56**	5	2	0	2	0	81.2
Happy	24	9	**37**	1	0	0	0	52.1
Disgust	2	2	1	**39**	0	1	1	84.8
Sad	0	0	0	8	**52**	1	1	83.9
Neutral	0	2	1	5	2	**62**	7	78.9
Boredom	0	3	0	15	5	5	**53**	65.4
Precision	75.0	68.3	53.6	54.2	88.1	87.3	85.5	72.7

Table 5.12: Automatic speech emotion results using Polish database using the auditory model of (Dau et al., 1997)

	Anger	Fear	Happy	Sad	Neutral	Boredom	Rate (%)
Anger	**25**	6	8	1	0	0	62.5
Fear	6	**23**	1	7	3	0	57.5
Happy	9	1	**27**	0	3	0	67.5
Sad	0	2	0	**29**	0	9	72.5
Neutral	1	2	1	6	**29**	1	72.5
Boredom	0	2	0	8	5	**25**	62.5
Precision	60.9	63.9	72.8	56.9	72.5	71.4	65.83

feature extraction method is different. With this test, auditory model's performance is tested. The test is applied both on EMO-DB and Polish databases. Performance on EMO-DB using the auditory model (Dau et al., 1997) is given in table 5.11. Recognition rate is measures as 72.7% which is low when compared with the model (Jepsen et al., 2008). Recognition rates for Polish database is given in table 5.12 which have a rate of 65.8%.
In another recognition test, EMO-DB and Polish databases are fused and tested with leave one speech sample out method. When two languages are mixed, recognition performance is measured 73.4% for seven emotions. Results are given in table 5.13. In order to measure the language independence, Polish database is used for test, and EMO-DB is used for training. However, recognition rates were very close to chance level.

Table 5.13: Automatic speech emotion results using fusion of EMO-DB and Polish databases with leave one speech sample out method

	Anger	Fear	Happy	Disgust	Sad	Neutral	Boredom	Rate (%)
Anger	**129**	12	22	0	1	3	0	77.2
Fear	13	**72**	10	2	3	8	1	66.1
Happy	24	10	**71**	3	0	2	1	63.9
Disgust	0	1	4	**33**	2	1	5	71.7
Sad	0	3	0	1	**83**	7	8	81.4
Neutral	1	6	1	3	6	**90**	12	75.6
Boredom	0	4	0	4	8	10	**95**	78.5
Precision	77.7	66.6	65.7	71.7	80.6	74.4	77.9	73.4

5.3 Speech Emotion Recognition Performance Comparison

Extracted subjective and automatic speech emotion recognition results are compared which are tested using EMO-DB. Four different conducted test cases is compared. Results and test cases are given in table 5.14. Listening test carried on human listeners have the lowest recognition rate. In all test cases anger emotion have a recognition rate above 80%. Happy emotion have the lowest recognition rate in three test cases. In speaker dependent case, recognition of happy has second lowest rate. Although neutral emotion have a recognition rate above 80% in all automatic recognition algorithms, human neutral recognition rate have a rate of 58%. Recognition of sad emotions is one of the most distinguishable emotion in all test cases.

Table 5.14: Speech emotion recognition performance comparison

	Anger	Fear	Happy	Disgust	Sad	Neutral	Boredom	Rates
Subjective P.(%)	87.0	52.0	44.4	47	67	58.0	52.7	58.4
Leave one speech out(%)	91.3	76.8	63.4	80.4	91.9	86.1	83.9	82.9
Leave one speaker out(%)	85.8	62.3	43.6	71.7	82.2	81.0	69.14	72.3
Speaker dependent(%)	81.8	73.9	63.3	52.17	91.9	86.0	72.8	76.26

CHAPTER 6

Discussion

In the scope of this thesis, automatic speech emotion recognition algorithm is developed. Developed emotion recognition algorithm aims to classify seven emotions that are anger, happy, fear, sad, neutral, disgust and boredom. Computation model of auditory system (Jepsen et al., 2008) is used in order to extract features. Extracted features are first projected using component analysis. Using principal component analysis, each emotional speech samples Euclidean distance to each other is estimated. According to their distance in Euclidean space, a binary tree is constructed. Each time, new branch is constructed, projected vectors are re-calculated. Each time, their distance of the projected vectors is calculated in Euclidean space. According to these distances, branches are constructed. Besides constructing binary tree, principal component analysis is used to classify emotions. Each test speech sound, classified using the binary tree. In each branch, sample signals' distance is found then segmented.

In each branch classification results are provided in tables. The branch which happy and anger emotions are classified has a segmentation rate of 62.6%. Since the recognition rate is slightly above the chance level, new feature set is searched for the segmentation of happy and anger emotions. As features short time spectral features are extracted. As short time spectral features, spectral centroids and spectral mean is extracted. As first step, same speaker same sentence case is tested. Each sentences happy and anger emotions short time spectral centroids and spectral means are extracted. When happy and angers features are compared, it is observed that mean of the spectral centroids is higher in anger emotion of the same sentence. On the other hand, standard deviation of the centroid of the happy emotions is higher compared to anger emotions. Yet, both mean and standard deviation of the centroid requires normalization. When ratio of mean to standard deviation (coefficient of variation) is taken anger and happy speech signals coefficient of variation of centroid becomes distinguishable. In addition to centroid, spectral mean is taken. When energy normalization is applied to each short time window, distribution of spectral statistic features for anger and happy emotions becomes distinctive. Most distinctive feature is shown to be the standard deviation of the spectral mean. In order to segment happy and anger emotions, these two features coefficient of variation of spectral centroid vector and standard deviation of spectral mean vector is classified with support vector machines. The classification result is estimated as 76%.

In the constructed binary tree, the branch in which happy and anger is segmented is very special. When binary tree is inspected all branches are compatible with the activation dimension of emotion space except happy and anger. Happy and anger emotions lie on the same activation level, yet on different valence level. In this study, it is inspected that distribution of spectral features provides important cues in valence domain.

As a second method, speech emotion recognition algorithm is implemented with support vector instead of principal component analysis. For each branch, support vector machines are trained. When segmentation results of PCA and SVM is compared, in each branch SVM

overwhelms PCA in classification performance. Besides that, it has been shown that features extracted from auditory model carries information in valence domain. Although the classification of happy anger with PCA has a rate of 62%, with SVM, classification rate is 87% which is even higher than classification using spectral features. In the overall classification algorithm, SVM with linear kernel overwhelm the classification performance using PCA. When PCA and SVM is compared, SVM provides better classification performance. When PCA and SVM are compared, PCA is a unsupervised data reduction method. On the other hand, SVM is a supervised binary classifier. SVM estimates the best hyperplane which separates two classes using training set. One advantage of SVM is that, distribution of the data is not important. Since SVM is supervised, position of the data is known in the feature space. In unsupervised methods, distribution of the feature vectors should fit to a distribution. In the classification task carried out using PCA, first dimension reduction is applied using PCA, then Euclidean distance of projected vectors is used the measure distance of each test sample to train sample. In this approach, since Euclidean distance is used, distribution of the feature vectors should be such that, in feature space, each class should form a n dimensional sphere. Instead of Euclidean distance, when different distance measures such as Mahalanobis distance are used. Yet, performance of classification hasn't increased. This shows that in feature space, distribution of feature vectors is not suitable to classify. On the other hand, higher classification result with SVM shows that, there is a hyperplane which separates feature vectors.

In the results section, developed auditory model based speech emotion recognition algorithm performance rates are provided. Algorithm is tested with three different speech emotion databases whose languages are German, Polish and English. On databases 5 different cases tested. In the first case, leave one speech sound out method is applied. In this method, in the training except the test speech sound, for training all remaining speech samples are used. Recognition rates of the German, Polish and English databases are 82.9%, 71.3% and 71.8% respectively. Although obtained results do not prove the speaker independence or language independence, these results prove that developed SER algorithm is not specific for any language or speaker. Developed algorithm is working for these languages in cases algorithm is trained with the given dataset.

In the second case, EMO-DB and Polish databases are tested with the auditory model (Dau et al., 1997) based speech emotion recognition algorithm. In this case, current auditory model's performance is compared with the one (Jepsen et al., 2008). In this test, it is expected that, if the current auditory model simulate the human ear better, then it is expected that in this test performance rates should be lower. As expected, results drop in both databases. In EMO-DB, performance rate drops to 72.7% and performance of the Polish database result dropped to 65.83%. These results have shown that, extracted features from newly auditory model are more discriminating.

In the third case, speaker independence of the algorithm is tested. In order to test speaker independence, leave one speaker out method is implemented. Results extracted on EMO-DB are 72.3%. On the other hand, results extracted from Polish and English databases are 56.25% and 48.2% respectively. These low results may be a result of insufficient number of speakers in the databases. Number of speakers in the EMO-DB is 10. Number of speaker in the Polish Database is 8 and 4 in the English Databases.

In the fourth case, speaker dependent case is tested. 76.24% recognition rate is obtained when the developed algorithm is tested on EMO-DB. When speaker dependent case is compared with leave one speech sample out method, dependent case has lower rate. Low result may be results of low number of train samples. In EMO-DB, total of 535 samples and 10 speakers exist which means a speaker have nearly 53 speech sentences. For speaker dependent case, although number of train samples is low, sufficient recognition performance is obtained.

In the fifth case, language independence is tested. When EMO-DB is used for training and Polish database is used for test samples, performance of the recognition algorithm is very close to the chance level. On the other hand, when EMO-DB and Polish database is fused and leave one speech sound out method is applied, 73.4% recognition rate is obtained. This result shows that, extracted features for emotion recognition task are discriminating for different emotions and have similarities for same emotions. In the classification part, support vector machine generate a hyperplane which separates two classes. This hyperplane is generated in a way that maximizes the distance between the support vectors. Generated hyperplane changes as the train samples changes. When Polish database and EMO-DB is fused, generated hyperplane separate both EMO-DB emotions and Polish emotions. In the scope of this information, with sufficient number of speaker and language, SVM could generate a hyperplane which fits to all speakers and languages.

Automatic speech emotion recognition test has provided some insight in the source side of a speech. Automatic emotion recognition rates have shown that in the speaker side, generated speech carries unique features for seven emotions. This strengthens the embodied emotions thesis. Some bodily changes results some changes in the generated speech. Even for the same speaker same sentence case, generated speech sample shows distinctive difference which supports the embodiment thesis.

Comparison of subjective and automatic speech emotion recognition performances has shown that automatic emotion recognition is better. This result shows that although information in the speech is available, human emotion recognition part is not able to capture or evaluate necessary data. Besides low recognition rate for listeners, variation between the recognition rates of the listeners is very high (44.4%, 78.57 %). Low recognition rate and difference in the recognition rates may be result of two possibilities. First possibility is that, in the preprocessing system, process show some distinctions and causes loss of information. Filters in the human auditory system are not unique at all. Especially, auditory filterbank responses differ from person to person. It is a known fact; people who are liable to music have more accurate acoustic filterbank responses. If the frequency response of the filterbanks is not accurate enough, loss of information such as emotions is possible. Second possibility is that, the part in the human brain which is responsible for emotion recognition is got affected and its recognition performance may degenerate. Emotional or mental state may have an effect on the recognition performance. Pleased or unpleased state of the listener may increase or decrease the attention of the listener. As mention in chapter two, emotions has an effect on the cognitive process. Different emotional state may result in different recognition rates. In the subjective test, it is requested that, after listening an emotional speech, listeners should select an emotion label. Besides, emotional tags, cannot decision chose is included too. In such a multiple choices, attention has a high importance and recognition rates directly related with it. On the other hand, some cognitive processes have lost its importance such as memory. In that case, increases in the attention will probably increase the recognition rate. Reward or a punishment will probably increase the success rates since it could increase the attention of the listener. Besides reward or punishment, activation and valence domain may have an effect on the recognition performance. In the high activation and high valence level such as happy, recognition performance may increase. Emotional states such as sad or boredom may decrease the recognition rate.

To more generalize the emotion recognition task, in the made subjective test, emotional speech samples are exhibited as a snapshot. Besides, although emotions are analogue, discretized labels are provided in the choices. On real life case, emotional stimuli occur as a result of something. An event triggers an emotional state and by the time, emotional state changes from one emotion to other. On the other hand, in the made subjective test, listeners are subject to short

uncorrelated emotional stimuli. Since we live in a social environment, interactions shape our emotions. Therefore, human emotion recognition performance should be measured in a social event, such as a telephone dialog. As mentioned in the distributed emotions, one event may cause different emotional states in different people. To further develop the test, instead of uncorrelated emotional stimuli, a book chapter may be offered as an emotional stimulus. Instead of discretized emotional labels, listener could select the emotional label from a two dimensional map which represents the emotional dimensions. In that way, real corresponding emotional state may be extracted. Such a study, in further be used in a dialog model or emotion generation stage.

CHAPTER 7

Conclusions

As technology evolves, interest in human like machines increases. Technological devices are spreading and user satisfaction increases importance. A natural interface which responds according to user needs has become possible with affective computing. The key issue of affective computing is emotions. Any research which is related with detection, recognition or generating an emotion is affective computing. User satisfaction or un-satisfaction could be detected with any emotion recognition system. Besides detection of user satisfaction, such systems could be used to detect anger or frustration. In such cases, user could be restrained like driving a car. In emotion detection tasks, speech or face emotion detections are the most popular ones. Easy access to face or speech data made them very popular.

Speech carries a rich set of data. In human to human communication, via speech information is conveyed. Acoustic part of speech carries important info about emotions. In this thesis, automatic speech emotion recognition algorithm which only relies on acoustic features is developed. Different from other speech emotion recognition algorithm which uses, prosodic or short time spectral features, developed algorithm is based on an auditory model. In this work, human speech emotion process is tried to be simulated. More importantly, to test human speech emotion performance using only acoustic part, listening test is constituted with German speeches on listener who do not know any German. Therefore, extracted results provide a comparison between subjective and automatic speech emotion recognition task. Selected auditory model composed of 6 main stages. In the auditory model outer-middle ear transformation, auditory filterbank, envelope, expansion, adaptation and modulation filterbank stages exist. Each stage has an equivalence in the human auditory system. Short time process is replaced by auditory filterbanks . Therefore, the need for short time processing is disappeared. Since process after the auditory system is unknown, generating a model which simulates the speech emotion recognition in humans is not possible. Therefore a subjective test result provides important cues about the process.

Computational auditory model is used to generate a feature set. To convert auditory model output to feature set, simple transformation methods such as mean and standard deviation is used. Extracted features are classified into 7 discrete emotions using classification algorithms. Extracted maximum accuracy rate of 82% shows that, human auditory system is specialized for emotions recognition. Besides, when different auditory models are used to extract features, maximum accuracy is obtained using this model. This profs the correctness of the used auditory model.

In the classification part, using extracted features, 7 emotions are segmented. To construct a binary tree, PCA is used. From each training sample, projected vectors are constituted then first five principal components mean is calculated for each emotion. Distance between each emotion is found using the Euclidean distance between each emotion. In binary tree, seven emotions are segmented into two labels which are named as excited and non-excited emo-

tions. Non-excited emotions, sad, neutral and boredom are segmented into two labels which are neutral and sad-boredom emotions. In the excited emotions branch, anger-happy is one branch and fear-disgust is another branch. Main reason of the constructing a binary is to maximize the uncorrelated variables between classes. When seven emotions distance is estimated, distance between happy and anger is lower compared, when only happy and anger emotions distance is found. In addition to that, to classify emotions with binary classifiers such as SVM a binary tree is required. To classify emotions, since there are 6 branches, six times projected images are calculated for each branch. When SVM is compared with PCA, accuracy of SVM surpass for each branch. Since classification accuracy of happy and anger emotion is lower compared to other branches, other features are investigated. Short time spectral features are tested, to segment happy and anger. Spectral centroid and spectral mean provides distinctive features. These features are classified with SVM, yet, when output of the auditory model is classified with SVM, higher results are obtained.
To test the overall accuracy of the developed algorithm, English, German and Polish datasets are used. Algorithm with the SVM's Overall performance is tested. Five different cases tested with the databases. Algorithms speaker and language dependence, independence is tested in these scenarios. Besides these five scenarios, subjective test is implemented on 10 listeners. Yet, automatic recognition surpassed the subjective performance of human evaluators. Maximum accuracies are obtained when Berlin emotional speech database is used. For speaker dependent case accuracy of the results was 76.26% and for speaker independent case 72.3% accuracy is obtained.

CHAPTER 8

Future Work

In future work, performance of the generated algorithm could be improved. In feature extraction part, extracted features from auditory model may be enhanced. Instead of using mean and standard deviation, more complex methods could be used to extract features from auditory model output. Besides, modulated signals are not the only output generated by auditory model. Human auditory system transmit to the brain, phase information of the first three auditory filterbank output.

Results have shown that when leave speech sample out method is implemented, highest accuracy rates are obtained for all three databases when compared with speaker dependent and independent cases. In leave one speech sample out method, all speakers are included in the training part. This shows that, there is hyperplane which can classify all seven emotions. SVM selects the hyperplane from many choices which maximizes the margins. Since in speaker independent case, number of training samples is low, SVM selects the hyperplane accordingly. On the other hand, when leave one speech sample out method is implemented, generated hyperplane also segments the training samples in speaker independent case. To overcome this issue, and to generalize algorithm into speaker or language independence, a normalization in features could be searched. Besides, generated algorithm could be tested with noisy data, which fits to the real life data. In that case, algorithm could be extended to real life case. Since SVM is a binary classifier, binary decision tree s generated. Yet generated binary tree may not fit to the all languages. Therefore, instead of a binary classifier, multi class classifiers may increase success rate and could work properly for many languages. Besides multi class classifiers, using ensemble learning many different models could be fused. In the scope of this thesis, features are extracted using only the auditory model. Advantage of ensemble learning is that, developed many different speech emotion recognition algorithms could be fused to obtain more healthier results.

As a further study, cognitive model given in figure 2.1 could be completed. Emotion generation block should be next part of the study. In this block, using recognized emotions, and content, counter emotional state could be detected.

REFERENCES

Anderson, A. K., Christoff, K., Panitz, D., Rosa, E. D., & Gabrieli, J. D. E. (2003). Neural correlates of the automatic processing of threat facial signals. *J Neurosci*, *13*, 5627–5633.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., & Stolcke, A. (2002). Prosody-based automatic detection of annoyance and frustration in human-computer dialog. *ICSLP*, *3*, 2037–2040.

Ayadi, E., Moataz, Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition features, classification schemes and databases. *Pattern Recognition*, *44*, 572–587.

Ayadi, M. M. H. E., Kamel, M. S., & Karray, F. (2007). Speech emotion recognition using gaussian mixture vector autoregressive models. *ICASSP*, *4*, 957–968.

Bartneck, C. (2002). *emuu-an embodied emotional character for the ambient intelligent home* (Unpublished doctoral dissertation). Technische Universiteit Eindhoven.

Benesty, J., Sondhi, M. M., & Huang, Y. (2007). *Springer handbook of speech processing*. Springer.

Burkhardt, F. (2005). A database of German emotional speech. *Proc Interspeech*, 3–6.

Burkhardt, F., & Sendlmeier, W. F. (2000). Verification of acoustical correlates of emotional speech using formant-synthesis. *ISCA Workshop on Speech and Emotion*, *4*, 151–156.

Cahill, L., Haier, R. J., Fallons, J., Alkire, M. T., Tang, C., Keator, D., ... Mcgaush, J. L. (1996). Amygdala activity at encoding correlated with long-term, free recall of emotional information. *National Academy Science*, *93*, 8016–8021.

Casale, S., Russo, A., Scebba, G., & Serrano, S. (2008). Speech emotion classification using machine learning algorithms. *IEEE International Conference on Semantic Computing*, 158–165.

Cichosz, J., & Slot, K. (2005). Low dimensional feature space derivation for emotion recognition. *Interspeech*.

Cichosz, J., & Slot, K. (2007). Emotion recognition in speech signal using emotion extracting binary decision trees. *ACII*.

Dau, T., Kollmeier, B., & Kohlrausch, A. (1997). Modeling auditory processing of amplitude modulation. i. detection and masking with narrow-band carriers. *The Journal of the Acoustical Society of America*, *102*, 2892–2905.

Dau, T., Puschel, D., & Kohlrausch, A. (1996). A quantitative model of the effective signal processing in the auditory system. i. model structure. *Journal of The Acoustical Society of America*, *99*, 3615–3622.

Descartes, R. (1952). *Descartes phililosophical writings*.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*, 169–200.

Engberg, I. S., & Hansen, A. V. (1996). Documentation of the Danish emotional speech database(des). *Aalborg University, Denmark,*.

Glazer, C. S. (2003). *Looking closely at emotional expression in an online course: A case study of distributed emotion* (Unpublished doctoral dissertation). the University of Texas.

Gordon, R. M. (1990). *The structures of emotions: Investigations in cognitive philosopgy*. Cambridge University Press.

Grimm, M., Kroschel, K., & Narayanan, S. (2008). The vera am mittag german audio visual emotional speech database. *IEEE International Conference on Multimedia and Expo*.

Hanjalic, A. (2006). Extracting moods from pictures and sounds:towards truly personalized tv. *IEEE Signal Processing Magazine*, *23*, 90–100.

Haq, S., Jackson, P., & Edge, J. (2008). Audio visual feature selection and reduction for emotion classification. *AVSP*, 185–190.

Haq, S., & Jackson, P. J. (2012). *Svm tutorial classification, regression and ranking* (G. Rozenberg, T. Back, & J. N. Kok, Eds.). Springer.

Hatsimoysis, A. (2003). Philosophy and the emotions. *Royal Institute of Philosophy Supplements*.

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction*, *7*, 174–196.

Iliou, T., & Anagnostopoulos, C. N. (2010). Classification on speech emotion recognition - a comparative study. *International Journal On Advances in Life Sciences*, *2*, 18–28.

Izard, C. E. (1984). *Emotion-cognition relationships and human development* (C. E. Izard, J. Kagan, & R. B. Zajonc, Eds.). Cambridge University Press.

Jaiswal, R. C. (2009). Audio video engineering.

James, W. (1994). The physical basis of emotion. *Psychological Review*, *101*, 205–210.

Jepsen, M., Ewert, S., & Dau, T. (2008). A computational model of human auditory signal processing and perception. *The Journal of the Acoustical Society of America*, *124*, 422.

Jianhua Tao, T. T. (2005). Affective Computing: A Review. *Lecture Notes in Computer Science*, *3784*, 981–995.

Le, P. N., Ambikairajah, E., Epps, J., Sethu, V., & Choi, E. H. C. (2011). Investigation of spectral centroid features for cognitive load classification. *Speech Communication*, *53*, 540-551.

Lee, C., Busso, C., Lee, S., & Narayanan, S. (2009). Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions. *Interspeech*, 1983–1986.

Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. *Journal of the Acoustical Society of America*, *93*, 1097–1108.

Narayanan, S. S. (2005). Toward detecting emotions in spoken dialogs. *Speech Communication*, *13*, 293–303.

Nogueiras, A., Moreno, A., Bonafonte, A., & no, J. B. M. (2001). Speech emotion recognition using hidden markov models. *Eurospeech*.

Nwe, T. L., Foo, S. W., & Silva, L. C. D. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication*, *41*, 603–623.

Ortony, A., Clore, G. L., & Collins, A. (1990). *The cognitive structure of emotions*. Cambridge University Press.

Parkinson, B. (1996). Emotions are social. *British Journal of Psychology*, *87*, 663–683.

Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the cuidado project.

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nat Rev Neurosci*, *2*, 148-158.

Petrushin, V. A. (2000). Emotional recognition in speech speech signals: Experimental study, development and application. *ICSLP*.

Picard, R. W. (1995). Affective Computing. *MIT Technical Report*.

Plack, C. J. (2004). Auditory perception.

Plutchik, R. (2011). The nature of emotions. *American Scientist*.

Poveda, E. A. L., & Meddis, R. (2001). A human nonlinear cochlear filterbank. *The Journal of the Acoustical Society of America*, *110*, 3107–3118.

Psychoacoustics. (2001). `www.music.miami.edu/programs/mue/research/mescobar/thesis/web/Psychoacoustics.htm`. (Accessed: 2013-08-21)

Robinson, D. J. M. (2000). *The human auditory system.* `http://www.mp3-tech.org/programmer/docs/human_auditory_system.pdf`. (Accessed: 2013-08-19)

Ruiz, N. (2011). *Cognitive load measurement in multimodal interfaces* (Unpublished doctoral dissertation). The University of New South Wales.

Scheingold, L. (2010). Understanding and expressing feelings.

Schlosberg, H. (1954). Three dimensions of emotion. *Psychological Review*, *61*, 81–88.

Schuller, B., Rigoll, G., & Lang, M. (2003). Hidden markov model based speech emotion recognition. *ICASSP*, *2*, 1–4.

Schuller, B., Rigoll, G., & Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. *IEEE international conference on acoustics speech and signal processing*, 577–580.

Tawari, A., & Trivedi, M. M. (2010). Speech emotion analysis: Exploring the role of the content. *IEEE Transactions on Multimedia*, *12*, 502–509.

Teager, H. (1990). Some observations on oral air flow during phonation. *IEEE Trans. Acoust. Speech Signal Process*, *28*, 599–601.

Ververidis, D., & Koropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, *48*, 1162–1181.

Vidrascu, L., & Devillers, L. (2005). Detection of real-life emotions in call centers. *Interspeech*.

Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: some acoustical correlates. *Journal of the Acoustical Society of America*, *52*, 1238–1250.

Wu, D., Parsons, T. D., & Narayanan, S. S. (2010). Acoustic feature analysis in speech emotion primitives estimation. *Interspeech*.

Wu, S., Falk, T. H., & Chan, W.-Y. (2009). Automatic recognition of speech emotion using long-term spectro-temporal features. *Digital Signal Processing*, 1–6.

You, M., Chen, C., Bu, J., Liu, J., & Tao, J. (1997). Getting started with susas: a speech under simulated and actual stress database. *EuroSpeech*, *4*, 1743–1746.

APPENDIX A

Appendix A

A.1 Support Vector Machine

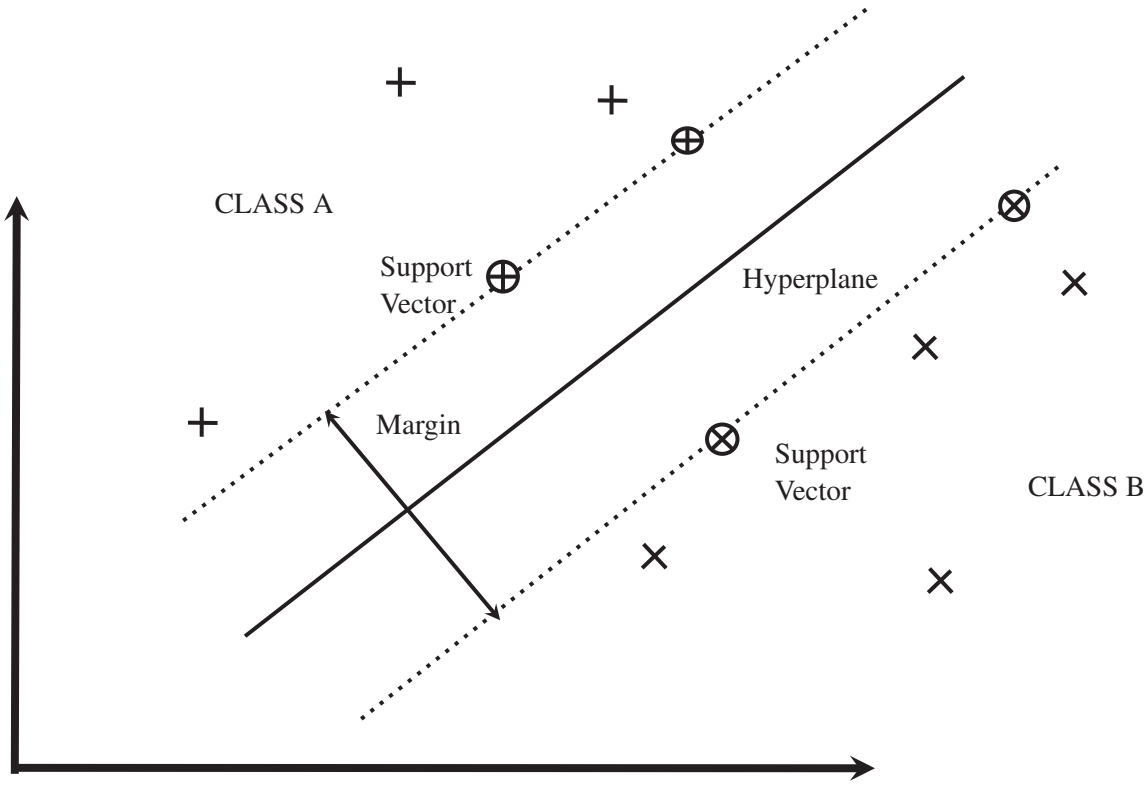

Figure A.1: A linear Support Vector Machine

Support vector machines are relatively new binary classification algorithm which has gain interest in many recognition tasks such as face recognition, speech recognition,written digit recognition, speaker identification . SVM classifier finds the best hyperplane which separates one class from other type of class. The best hyperplane is measured such that, a margin in which distance to support vectors are maximized. In figure A.1, support vector, hyperplane and margin is illustrated. Summation of the shortest distance to the separating hyperplane to the support vectors is named as the margin. Support vectors are the points which are closest

to the separating hyperplane (Haq & Jackson, 2012).
Assume there are d dimensional feature vector x for each training sample. Each feature vector x_i belongs a class $y_i = \pm 1$. To segment these two class dataset, there is a hyperplane. The equation of the hyperplane is given in A.1. In the training process, weight vector w and bias b will be calculated. Finding w and b that minimize $\|w\|$ for all data points x_i,y_i is the the question which defines the best hyperplane.

$$F(x) = \mathbf{w} \cdot \mathbf{x} - b \tag{A.1}$$

In order to classify the training set, equation F must satisfy the conditions given in A.2. If training set x_i belongs to class $y_i = 1$, equation F must return positive value. Else if training set x_i belongs to class $y_i = -1$, equation F must return negative value.

$$\begin{aligned} w \cdot x_i - b > 0 \quad & if \quad y_i = 1 \\ w \cdot x_i - b < 0 \quad & if \quad y_i = -1 \end{aligned} \tag{A.2}$$

D is called linearly separable, if there exist a linear function F which satisfies the condition given in A.2. These conditions are revised into form A.3. To maximize the margin, conditions in A.2 is revised into form A.3. If D is linearly separable, w and b could rescaled to satisfy A.3.

$$y_i(w \cdot x_i - b) \geq 1, \forall (x_i, y_i) \in D \tag{A.3}$$

Distance from the hyperplane to a vector x_i is equal to $\frac{|F(x_i)|}{\|\mathbf{w}\|}$. Distance of the closest feature vector to the hyperplane is $\frac{1}{\|\mathbf{w}\|}$ which is equal to the margin. The closest vector to hyperplane is called as support vector as given in figure A.1.

$$minimize : Q(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|^2 \tag{A.4}$$

$$subject\ to \ : y_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1, \forall (x_i, y_i)\ \epsilon\ D \tag{A.5}$$

In order to maximize margin, $\|\mathbf{w}\|$ should be minimized as given in equation A.4 at the same time must satisfy the condition given in A.5. These two arises a optimization problem that must be solved.

$$J(w, b, \alpha) = \frac{1}{2}\mathbf{w} \cdot \mathbf{w} - \mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i x_i \tag{A.6}$$

In order to solve the optimization problem, method of Largrange multipliers are used as given in the equation A.6.

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i x_i \tag{A.7}$$

$$0 = \sum_{i=1}^{m} \alpha_i y_i \tag{A.8}$$

Differentiating J(w,b,a) with respect to w and b and setting the results equal to zero, following two conditions are obtained as given in A.7 and A.8. For any $\alpha_i \geq 1$, when this equation is substituted given optimization problem is reduced to A.9.

$$Q(a) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \mathbf{x_i} \cdot \mathbf{x_j} \tag{A.9}$$

It is possible that, in some cases, separating dataset into two classes with a straight line is not possible. In that case, generated hyperplane is such that some variables which do not fit to the

hyperplane are omitted. Soft margin SVM detects the omitted variables while maximizing margin. To deal with this problem, slack variables are used to measure the degree of misclassification. This results in an optimization problem in which while keeping the distance of misclassified samples' distance minimum, margin should be maximized. With another parameter degree of how much the separating hyperplane could allow samples to reside on the wrong side is adjusted. As the parameters value increases, separating hyperplane allows more misclassified samples.

In another case, separating data with a linear hyperplane is not possible. In that case, instead of a linear hyperplane, high order kernels are used. First, data is mapped to a higher dimension in which data is separable with a linear hyperplane. Then, with dimension reduction, linear hyperplane is mapped to a higher order kernel.

APPENDIX B

Appendix B

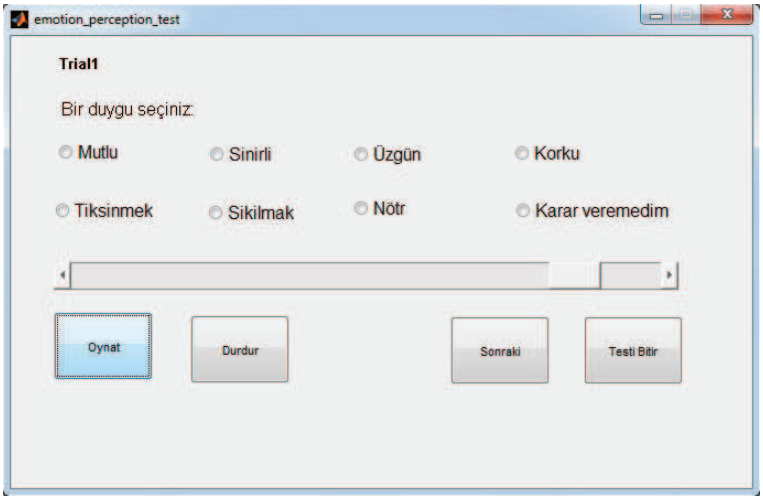

Figure B.1: Subjective test interface

TEZ FOTOKOPİSİ İZİN FORMU

ENSTİTÜ

Fen Bilimleri Enstitüsü	☐
Sosyal Bilimler Enstitüsü	☐
Uygulamalı Matematik Enstitüsü	☐
Enformatik Enstitüsü	X
Deniz Bilimleri Enstitüsü	☐

YAZARIN

Soyadı : Yüncü
Adı : Enes
Bölümü : Bilişsel Bilimler

TEZİN ADI (İngilizce) : Speech Emotion Recognition Using Auditory Models

TEZİN TÜRÜ : Yüksek Lisans X Doktora ☐

1. Tezimin tamamından kaynak gösterilmek şartıyla fotokopi alınabilir. X
2. Tezimin içindekiler sayfası, özet, indeks sayfalarından ve/veya bir bölümünden kaynak gösterilmek şartıyla fotokopi alınabilir. ☐
3. Tezimden bir (1) yıl süreyle fotokopi alınamaz. ☐

TEZİN KÜTÜPHANEYE TESLİM TARİHİ :