CLASSIFICATION OF EMOTIONS IN VOCAL RESPONSES

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

BY

ECE ÇAĞLAYAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF
MASTER OF SCIENCE
IN
MEDICAL INFORMATICS

SEPTEMBER 2017

CLASSIFICATION OF EMOTIONS IN VOCAL RESPONSES

Submitted by **ECE ÇAĞLAYAN** in partial fulfillment of the requirements for the degree of **Master of Science in Medical Informatics Department, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin _____
Dean, **Graduate School of Informatics**

Assoc. Prof. Dr. Yeşim Aydın Son _____
Head of Department, **Health Informatics**

Assoc. Prof. Dr. Tolga Esat Özkurt _____
Supervisor, **Health Informatics**

Examining Committee Members

Assoc. Prof. Dr. Yeşim Aydın Son _____
Health Informatics Dept., METU

Assoc. Prof. Dr. Tolga Esat Özkurt _____
Health Informatics Dept., METU

Asst. Prof. Dr. Aybar Can Acar _____
Health Informatics Dept., METU

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

Asst. Prof. Dr. Reza Zare Hassanpour _____
Computer Engineering Dept., Çankaya University

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: ECE ÇAĞLAYAN

Signature :

ABSTRACT

CLASSIFICATION OF EMOTIONS IN VOCAL RESPONSES

Çağlayan, Ece
M.Sc., Department of Medical Informatics
Supervisor : Assoc. Prof. Dr. Tolga Esat Özkurt

September 2017, 77 pages

Emotion is a relatively short-term conscious experience characterized by intense mental activity and high level of pleasure or dissatisfaction. During a dialogue, a person feels the emotion in the other person voice and chooses accordingly how to react. Within the scope of this thesis, it is investigated whether we can distinguish the emotional content of a response from the speech signals regardless of the semantics. Accordingly, audio recordings containing six basic and neutral emotions were played to the participants severally. Since the aim is to measure the effect of the acoustic structure rather than semantic structure we took account of German voice recordings from the Berlin emotional speech database. In this respect, meaningful Turkish sentences comprising neutral words were shown on the screen randomly as the next step of the experiment. Participants were expected to read these sentences with their emotional reaction to the previous voice record. Audio recordings of the participants were taken. Thus, an artificial ”dialogue” was reproduced. To our knowledge, this is the first research of classification of emotional responses to an emotional audio record. In our study, 30 basic features were extracted from speech records of 21 subjects who participated in our experiment and their emotional responses to audio records were classified using an artificial neural network. By this way, it is considered that the measurement of the acoustic response to a particular emotion can be classified. After the statistical analysis, it has been shown that the response given for the anger can be classified in reasonable rate. In addition to classifying the responses to emotional audio records, we foresee that classification performance for emotional responses can be increased.

Keywords: emotions, emotional speech, emotional reaction classification, acoustic features, machine learning

ÖZ

SÖZEL TEPKİLERDEKİ DUYGULARIN SINIFLANDIRILMASI

Çağlayan, Ece
Yüksek Lisans, Sağlık Bilişimi Bölümü
Tez Yöneticisi : Doç. Dr. Tolga Esat Özkurt

Eylül 2017, 77 sayfa

Duygu, yoğun zihinsel aktivite ve yüksek derecede zevk veya hoşnutsuzluk ile karakterize edilen nispeten kısa süreli bilinçli bir deneyimdir. Kişi diyalog sırasında karşısındakinin sesindeki duyguyu hisseder ve nasıl tepki vereceğini ona göre seçer. Bu tez kapsamında, insanlarda gerçekleşen bu yeteneğin makine öğrenme yöntemleri kullanarak sınıflandırılıp sınıflandırılamayacağı araştırılmıştır. Bu doğrultuda katılımcılara, altı temel duygu ve nötr duyguyu ayrı ayrı içeren ses kayıtları dinletilmiştir. Cümlenin anlamsal bütünlüğündense akustik değerlerin etkisinin ölçümü istenildiğinden, Berlin duygusal konuşma veri tabanından alınan Almanca ses kayıtları dikkate alınmıştır. Nötr kelimeler içeren anlamlı Türkçe cümlelerden rastgele bir tanesi deneyin sonraki adımı olarak ekrana yansıtılmıştır. Katılımcılardan bu cümleleri, bir önceki ses kaydına karşı duydukları tepkiyle okumaları beklenmiştir. Bu sırada katılımcıların ses kaydı alınmıştır. Böylece, yapay bir "diyalog" ortaya konmaktadır. Bildiğimiz kadarıyla, çalışmamız duygusal konuşmalara verilen cevapların duygusal niteliğini sınıflandıran ilk çalışmadır. Çalışmada 21 katılımcıya uygulanan deneylerin ses kayıtlarından 30 temel öznitelik çıkarımı yapılmış ve yapay sinir ağı kullanılarak duygusal seslere tepkiler sınıflandırılmıştır. Bu yolla kişilerin belli duygular karşısındaki akustik tepkilerinin ölçümü yapılabilmiştir. Yapılan istatiksel analizlerin sonunda kızgınlık için verilen tepkinin makul oranda sınıflandırılabileceği gösterilmiştir. Çalışmamız, duygusal seslere verilen tepkilerin sınıflandırılmasına ek olarak verilen tepkilerin sınıflandırma başarısını arttırabileceğini öngörmektedir.

Anahtar Kelimeler: duygular, duygusal konuşma, duygusal tepki sınıflandırması, akustik özellikler, makine öğrenimi

to my family ...

ACKNOWLEDGMENTS

First of all, I would like to express my special thanks of gratitude to my supervisor Assoc. Prof. Dr. Tolga Esat ÖZKURT who has contributed in all phases of my studies and provided great patience, support and encouragement. His observations and inspirational criticism were very valuable.

Besides my supervisor, I would like to thank Assist. Prof. Dr. Didem GÖKÇAY for providing an access to TUDADEN databases used in this work.

I would also like to thank all of my professors and colleagues from Informatics Institute for their help during my graduate studies. I have always consulted with them about my projects. Their feedback helped me through this project and writing up this thesis.

Finally, I would also like to thank my family members and friends who encourag e me a lot to achieve my goal. They are always patient and helpful during my thesis.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ASR	Automatic Speech Recognition
SER	Speech Emotion Recognition
LPC	Linear Predictive Coding
LPCC	Linear Predictive Cepstral Coefficients
LFPC	Log Frequency Power Coefficients
MFCC	Mel Frequency Cepstral Coefficient
MEDC	Mel Energy Spectrum Dynamic Coefficients
DWT	Discrete Wavelet Transform
WVD	Wavelet Vaguelette Decomposition
ASSESS	Automatic Statistical Summary of Elementary Speech Structures
LD	Linear Discriminant
GVQ	Generative Vector Quantization
TEO-FM-Var	TEO decomposed FM Variation
TEO-Auto-Env	Normalized TEO Autocorrelation Envelope Area
TEO-CB-Auto-Env	Critical Band Based TEO Autocorrelation Envelope Area
SVM	Support Vector Machine
HMM	Hidden Markov Model
NN	Neural Networks
ANN	Artificial Neural Network
GMM	Gaussian Mixture Model
K-NN	K- Nearest Neighbor
DNN	Deep Neural Network
DFT	Discrete Fourier Transform
DCT	Discrete Cosine Transform
ConvNets	Convolutional Neural Networks
RBM	Restricted Boltzmann Machines
DBN	Deep Belief Networks
CTC	Connectionist Temporal Classification

CHAPTER 1

INTRODUCTION

Nowadays, people spend considerable part of their daily lives online by being on the Internet and social media frequently (Van den Eijnden et al. 2016). As technology evolved, people have become more involved with the intelligent machines. For this reason, human-computer interaction is becoming important. This interaction became the focus of many workspaces, especially computer science and behavioral science (Vošner et al. 2016). As the machines gained more human characteristics, the realization of human-computer interaction increased.

Emotions are part of human life and always spice up the life. In addition, emotions are the factors that affect relationships with other people (Hayley et al. 2017). Emotional intelligence is having knowledge and ability about emotions. More precisely, individuals, whose emotional intelligence is high, are more successful in solving practical problems (Afshar and Rahimi 2014). For instance, they perceive the facial expressions better, express emotions appropriately, feel comfortable in perceiving the implicit messages. Individuals like this generally have no difficulty in guessing how they will behave while living their emotions. In addition, a person, who has high emotional intelligence, is more sensitive to others and more capable of empathy (Alloway et al. 2016).

While emotions play such a significant role in human communication, studying humanoid machines with emotional characteristics is expected. In other words, it has become inevitable that emotional intelligence gains importance, as artificial intelligence becomes so important in computers. In order to humanize the machines, it was known that the only one sense (vision) that a person owns was not sufficient, and analysis the speech data (auditory sense) was researched, additionally.

Today's humanoid machines are able to understand what we want to say and analyze how we feel. Besides that, you expect a human being whom you communicate with to react emotionally like you do. So that you can establish a realistic dialogue with machines. Therefore, teaching how machines should respond to certain emotions became an utmost important topic (Terzis et al. 2012; Robinson et al. 2011). It is under consideration that the responses to emotional speech can be modeled in a certain pattern. However, it is not yet clear how this model will be created. In this thesis, studies on the analysis of this model are made. If the reactions given to an emotion have a certain model structure, this model can also be used to teach how should machines react.

The intelligent personal assistant can be an example of the use of this emotional pattern. This assistant, worked by the voice command answers the question we ask, helps us find what we are looking for on the Internet, or gives us advice. While doing these tasks, the desired reality is that the assistant can speak with emotional tones as a human does. For the sake of argument, we had an imaginary accident and we got panic. We could use this assistant to call for an ambulance. As it would detect the panic in our tone and make us relax using a calming tone. This type of application may particularly be useful for people with chronic illnesses such as asthma or the old people who require special care (Lugović et al. 2016).

In addition to this example, perception of emotional speech can have many uses in health care applications. These are recognizing the emotions of patients after the treatment, monitoring the rehabilitation patients. Psychologists who provide personal counseling may use them as the resource for patient's emotional state. The intelligent assistant might help for the companion, and health conditions can always be monitored.

The intense working conditions of parents, the challenging conditions of life have left the today's children more alone, which has made children more interactive with technology. Emotionally-deprived machines led the majority of new generation children to have behavioral and personality disorders (Taylor 2012). People with the personality disorders often lack emotional experience and have difficulty understanding the emotions of others. Machines with emotional intelligence will be able to overcome this difficulty (Coffey et al. 2017). On the other hand, emotional intelligence can be converted into a new therapy for the patients who suffer from birth such as autism (Schafer et al. 2016).

People express their emotions in verbal or nonverbal ways. In this thesis, the verbal expression is examined. Three basic questions about the classification of emotions are researched: 1) Can responses to emotions be classified? 2) How distinct are the classification accuracies for emotions? and 3) Can features from emotional responses contribute to the classification of emotional state?

The content of first hypothesis in our work is that vocal responses can be classified using artificial neural networks. For this purpose, an experiment was applied made with the help of participants. A voice record contains one of seven emotions listened to participants each trial. It was expected that they reacted vocally to a voice record. The non-Turkish audio records were played to minimize the cognitive impression on the verbal response. Again with the same aim, the screens are displayed with semantically neutral Turkish words. Participants were asked to respond emotionally by reading them. The acoustic, prosodic and statistical features for both emotional audio records and their emotional responses were extracted and classified using an artificial neural network, separately. The percentages of classification obtained for each emotion are compared with each other. As in the other research, classification of the new dataset, that occurred after the features of the responses were added to the voice records, was made. Statistical analysis was conducted to determine whether or not the features of emotional responses contributed to the audio records' classification. Literature research shows that such a study has not been done to this day.

This thesis contains 4 more chapters. Briefly summarizing the contents of chapters;

literature review is presented in the second chapter. A review of databases, features and classification methods for the emotional speech and dialogue analysis is realized. In the third chapter, the experimental setup and methods used in our study are explained. Analysis results are given in the fourth chapter. In the final chapter, results are interpreted and discussed, while limitations of our study are explained briefly.

CHAPTER 2

LITERATURE REVIEW

2.1 Human Speech System

A vast majority of human beings use speech almost every day to express their thoughts. Nowadays, people are interacting not only with people but also with computers. For this reason, human-computer interaction is a trending research area that aims to optimize the virtual communication. In order to do this, firstly, how the brain processes the speech structure must be examined. This knowledge will lead to the development more of ”human-like” machines.

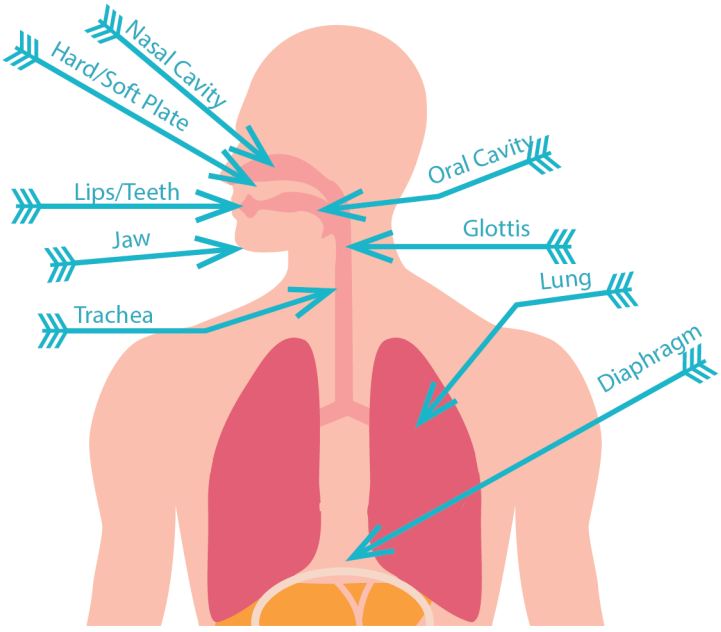


Figure 2.1: An overview of the human speech production system.

Speech is not just a wind coming from one’s mouth and reaching to the others’ ears, as there are detailed processes behind speech mechanism (Figure 2.1). The speech production mechanism contains three basic functions: motor control, articulatory motion and sound generation (Honda 2003). Motor control is a brain function that summarizes the thoughts that have been intended to be said and stimulates nerves of the

speech production organs. Changing the shape of the speech production organs to make the voice is associated with the articulatory motion part. The third function is the part that the air comes out of the mouth which turns into the acoustic waves in the space and reaches the ear.

2.1.1 Speech Production

Speech is a communication system based on a physical response that helps a person establish and maintain a balanced relationship with himself and his social environment. The conversion of thoughts into voice is called speech production. At the beginning of a human speech, vocal cords are stretched. The air is pushed towards the larynx from the lungs. The air vibrates the cords, and by this way, quasi-periodic wave (pitch impulse) is generated. The frequency of the periodic signal is called the pitch or fundamental frequency (F0). A change in the pitch frequency occurs while speaking.

After that, the pressure impulse vibrates the air in the oral and nasal cavity to form a particular sound. Both oral and nasal cavity acts as a resonator and helps to produce the sound wave. The frequency of resonators is called formant frequency. For creating different sounds, jaw, tongue, velum, lips or mouth should move to change the shape of the oral or nasal cavity.

Both pitch and formant frequency vary according to the gender or age. An adult male has a pitch frequency between 100 and 146 Hz, whereas for an adult female it is between 188 and 221 Hz (Gelfer and Mikos 2005).

2.2 Definition of Emotion

Emotion is a complex psychophysiological function of the central nervous system resulting from the interaction of an individual with biochemical (intrinsic) and environmental influences. Emotion constitutes an important element of the human soul and is in connection with all other forms of mental activity. It is the main factor that determines the sense of personal health and plays a central role in the daily life of a person (Cabanac 1992).

2.2.1 Emotional Model

Russell (1980) suggests a circular chart in which every basic emotion represents a bi-directional entity in the same emotional continuity. The suggested directions are arousal and valence fitting into the two axes of the chart. The valence dimension indicates how positive or negative the emotion varies from unpleasant to pleasant. The arousal reflects how emotion is actively simulated. These two axes intersect and divide the graph into four quadrants as positive/negative and low/high combinations. As Russell (1980) argues, it is possible that different emotional labels can be drawn on

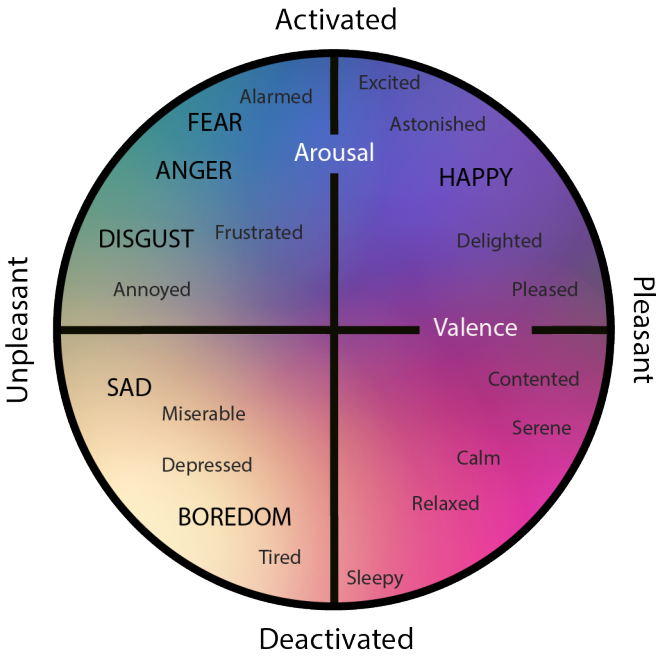


Figure 2.2: 2D Valence-Arousal plane. Emotion distribution is determined by the arousal levels (low to high) and their valence (unpleasant to pleasant).

the two-dimensional plane in various positions according to the valence and arousal (Figure 2.2).

2.2.2 Types of Emotion

Number of emotion types and classification of them are still in debate. In this section, emotions used in this thesis research were briefly explained.

Anger

Anger has the highest energy amongst fear, sadness, happiness, and disgust (Ververidis and Kotropoulos 2003)

Fear

Having high pitch and intensity level, this emotion correlates with anger. Pitch property lets the separation it from happiness, with which some other properties are similar (Ververidis and Kotropoulos 2003).

Sadness

Having very low energy, negative valence degree and lower average pitch, the speech rate of this emotion is lower than neutral (Murray and Arnott 1993).

Happiness

Positive valence and high energy are key properties of happiness. In addition, it is

stated that fundamental and formant frequencies increase in case of smiles (Ververidis and Kotropoulos 2003). In some cases, amplitude and duration may also increase for some speakers (Murray and Arnott 1993).

Disgust

When compared to neutral state, low mean pitch level, a low-intensity level, and a slower speech rate are observed (Ververidis and Kotropoulos 2003). Lowest speech rate and pause length increases are observed (Murray and Arnott 1993).

Boredom

Similar to sadness, boredom is a negative emotion with negative valence and low energy level. Lowered mean pitch and a narrow pitch range with a slow speech rate are observed (Murray and Arnott 1993).

Neutral

Features of all emotions are compared with neutral as its features are taken as the reference.

2.3 Automatic Speech Recognition (ASR)

ASR can be defined as the transcription of voice recordings spoken independently from the computer. ASR converts real-time talking to readable texts (Stuckless 1994). In summary, ASR allows a computer to identify words that a person is speaking on a microphone or a phone and translate them to produce a written text.

It is the result of more than 50 years of research that a machine can understand the spoken speech fluently. Although the machines cannot recognize the whole speech, ASR technology is used routinely in a range of applications and services, in every environment.

If the system is trained by the voice of a single speaker, then a much wider vocabulary can be used. Achieving an accuracy greater than 90% might be possible. Commercially available ASR systems often require only a short training of a speaker, and at regular speed, continuous conversation with high accuracy can be achieved with a large vocabulary.

2.4 Emotional Speech Recognition

Emotional speech recognition aims to identify the emotional state of a human from the voice. The emotional state is an essential factor in human communication. The primary objective of automatic speech recognition is the quality of human-machine interaction.

First emotional speech recognition started with the using of statics of the acoustic features (Bezooijen 1984; Tolkmitt and Scherer 1986). Additionally, the stress of speech

was questioned through the neutral sentences (Hansen and Cairns 1995). These methods were improved by iterative algorithms (Cairns and Hansen 1994; Womack and Hansen 1996). Currently, researchers focus on hybrid approaches of classifiers that increase the classification efficiency. With this approach, real-time applications are becoming even more advanced such as call center and medical diagnosis in therapy (Petrushin 1999; Lee et al. 2004; France et al. 2000).

Although there are lots of developments in recognition methods, emotion recognition still faces many difficulties. In addition to three sub-difficulties described by Schuller et al. (2009) in accuracy, classification, and features; noise, reverberation and feature selection also cover these difficulties.

Table 2.1: Table of emotional speech recognition review

No	Reference	Database	Feature Extraction	Feature Classification	Results
1	Pan et al. (2012)	Berlin, Chinese	Energy,Pitch, LPCC,MFCC, MEDC	SVM	Best Accuracy (MFCC + MEDC + Energy)
2	Morales-Perez et al. (2008)	SES (Spanish)	Gabor Transform, DWT, WVD, LPC, Raw Data	Confusion Matrix	LPC < WVD < DWT < GABOR < RAW DATA < MIXED The best results is 80.66%
3	Demircan and Kahramanlı (2014)	Berlin	MFCC, LFPC	K-NN	Classification success is found 50%. LFPC is better choice.
4	McGilloway et al. (2000)	Their DB: 5 passage x 5 emotion x 40 subject	ASSESS features	LD, SVM, GVQ	50% correct classification for 5 emotions
5	Ingale and Chaudhari (2012)	300 emotional real life situations	pitch, energy, duration, formant, LPCC, MFCC	GMM, K-NN, HMM,SVM and ANN	HMM classifier is observed as 76.12% for the speaker dependent
6	Wu et al. (2009)	Their case-study	Intonation Groups-based features	Minimum Classification Error (MCE), GMM	83.94% for the inside test 60.13% for the outside-open test
7	Yang et al. (2012)	Emotional Prosody Speech and Transcripts Corpus (EPST)	pitch, energy, pitch difference, energy difference, first four formants	SVM	They achieved accuracy of 80.5%

After reviewing the previous research and examining correlations between fundamental speech features and emotion classes, one may observe that that F1 formant frequency for anger, sadness, disgust, and fear is greater than for F1 formant frequency happiness (Wu et al. 2009).

Pan et al. (2012) extracted energy, pitch, LPCC, MFCC, MEDC features from Berlin Database (Burkhardt et al. 2005) and their own SJTU Chinese database and used

SVM with the combination of features. They found that the combination of MFCC, MEDC and energy gave the most accurate result.

Yang et al. (2012) used the same classification method, i.e., SVM on the Emotional Prosody Speech and Transcripts Corpus (EPST) (Liberman et al. 2002) and achieved 80.5% accuracy of correct classification for 6 emotions. More detailed information on the recognition of emotional speech is given in Table 2.1.

In Fayek et al. (2017), they investigated the application of end-to-end deep learning to Speech Emotion Recognition (SER) and explored how each of these architectures can be employed in this task. Various deep learning architectures were explored on a SER task. Experiments conducted illuminate how feed-forward and recurrent neural network architectures and their variants could be employed for paralinguistic speech recognition, particularly emotion recognition. Convolutional Neural Networks (ConvNets) demonstrated better discriminative performance compared to other architectures. The proposed SER system which relies on minimal speech processing and end-to-end deep learning, in a frame-based formulation, yields state-of-the-art results on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database (Busso et al. 2008) for speaker-independent SER.

2.4.1 Databases Used in Emotional Speech Recognition

In this section, general characteristics of databases are presented and also detailed analysis is done about specific databases. There are six databases explained in this section, Berlin Database of Emotional Speech (Emo-DB;Burkhardt et al. 2005)is used one as stimuli in our experiment.

First of all, there are three common emotional database types, which are natural, induced and acted. Natural emotions include the purest form of human voice emotion, but as expected, acquiring such data have more difficulty than the others (El Ayadi et al. 2011). In addition to the difficulty of acquiring samples, added noise within any recording reduces the chances of successful voice recognition and analysis.

Induced emotions are acquired by drugs, or presenting emotion inducing videos and imagery. This method allows a high degree of control of emotions. The drawback of this method is the power of emotions, which mostly happens weakly and changes from subject to subject (El Ayadi et al. 2011). Because of these drawbacks and application difficulties, this method was not used in our study.

Acted emotions are the most frequently used, as creating material and content is easier and most importantly the noise level can be controlled (El Ayadi et al. 2011). The primary disadvantage is in itself, using actors for these databases reduces the purity, to be exact; actors overemphasize some cues and may miss subtle ones that might appear in a natural expression of emotion.

The databases are explained in a brief description below. It is informative for the following sections, when databases is mentioned in an research. Emo-DB contains about 500 words elicited by actors. Ten different actors read ten different texts and acted in happy, angry, anxious, fear, bored and disgusted and neutral emotion (Burkhardt

et al. 2005).

Liberman et al. (2002) constructed an emotional database, named as Emotional Prosody Speech and Transcripts (EPST). Data, performed by German actors, were collected from transcripts and audio readings. Data were divided into fourteen emotional categories defined by Banse and Scherer (1996).

In BabyEars database (Slaney and McRoberts 2003), acoustic data include parents (6 mothers and 6 fathers) talking to their infants were collected. Then these records were presented to other parents to classify whether speech included approval, attention or prohibition statement and decide the strength of it.

FERMUSIII is a database containing six emotions of anger, disgust, fear, sadness, surprise and neutral being acted by 13 actors in the languages of English and German (Rigoll et al. 2005). FERMUSIII also includes some records of speech interaction dialogue.

In addition to databases containing the emotional voice recordings, there are many other databases containing emotional dialogue like FERMUSIII. One of the most commonly used was Intelligent Tutoring Spoken Dialogue System (ITSPOKE), which was developed for the students who responded to the physics problem and then students gave feedback about the answers and gave a complete explanation (Litman and Silliman 2004). ITSPOKE was used as a platform for examining whether acoustic-prosodic information to improve the recognition of pedagogically useful information.

In SEMAINE project, a Sensitive Artificial Listener (SAL) is a multi-modal dialogue system (Douglas-Cowie et al. 2008). SAL was designed to have a social skill for human-computer interaction.

2.4.2 Features Used in Emotional Speech Recognition

As a result of the literature searches, according to the categories, features used in emotional speech recognition is given in Table 2.2 and the most used methods are explained in detail after Table 2.2. Combining features that belong to various categories are mostly used to represent the emotional speech (El Ayadi et al. 2011).

Pitch features are relevant with the velocity coming from glottis during the vocal cord vibration. Joy and Surprise hold high velocity whereas anger and disgust hold low velocity (Nogueiras et al. 2001). Using pitch features alone had a weakness, Tolkmitt and Scherer (1986) and Iida et al. (2003) used the autocorrelation method for pitch estimation to increase the accuracy of the emotion classification.

Another useful feature for emotion recognition is Teager energy operator which is based on measuring the airflow through the vocal tract (Teager and Teager 1990). Zhou et al. (2001) classified the neutral and stressed speech using polynomial coefficients according to TEO autocorrelation and they achieved the accuracy of 89.5%, better than MFCC (67%).

Formants linked with the resonances are one of the vocal tract features and these formants give us useful information about the emotional state which subjects are in.

It was found that the first and the second formants give more emotional information than other formants (Tolkmitt and Scherer 1986; France et al. 2000).

Davis and Mermelstein (1980) brought forward a better approach which is Mel-frequency Cepstral coefficients (MFCCs) because fundamental features such as pitch and formant were not enough to achieve a good accuracy of speech recognition. However, Nwe et al. (2003) examined 6 emotions of anger, disgust, fear, joy, sadness and surprise and found that log frequency power coefficients (LFPC) give better results compared to the linear prediction Cepstral coefficients (LPCC) and Mel-frequency Cepstral coefficients (MFCC) feature parameters. Zhou et al. (2001) had also achieved results that prove this superiority.

2.4.3 Machine Learning Algorithms Used in Emotional Speech Recognition

According to a review study on emotional speech recognition conducted by El Ayadi et al. (2011), there are several types of classifiers such as GMM, SVM, ANN, K-NN and there was no certain agreement about which classification method was the most appropriate for emotion classification. Though, El Ayadi et al. (2011) emphasized that HMM was the most widely used classification method for emotion recognition from speech.

HMM is a model in which hidden variables, associated with the Markov operation, control the selection of the components. In the analysis of emotion recognition, the array of features that make up the advanced model brings out the output of the network. Nwe et al. (2003) used HMM with LFPC, MFCC, and LPCC features as the input of neural network to classify 6 different emotions and achieved an increase on the accuracy of human classification rates by % 12.7 and %9.7 for the Burmese and Mandarin database, respectively. Lee et al. (2004) also used HMM on their studies and researched the effect of sound level modeling on emotional speech classification with 4 different emotions: anger, happiness, neutral, and sadness. Overall accuracy results were 76.12% for phoneme-class dependent HMM and 55.68% for prosodic features SVM.

GMM is a model of the probability distribution of features such as spectral features

Table 2.2: Features used mostly in emotional speech recognition

Continuous Features	**Qualitative Features**	**Spectral-based Features**	**Cepstral Based Features**	**TEO Based Features**
Pitch-related	Voice level	LPC	LPCC	TEO- FM-Var
Formants	Voice pitch	LFPC	MFCC	TEO- Auto-Env
Energy-related	Phrase		MDEC	TEO- CB-Auto -Env
Timing / Duration	Temporal Structures			
Articulation				

related to vocal tract. Breazeal and Aryananda (2002) used GMM in their studies to examine the classification of KISMET dataset, which includes approval, attention, prohibition, soothing, and neutral emotional states. 78.77% accuracy was achieved when selecting five accurate features and increased by 3.17 applying hierarchical sequential classification methods. GMM was also applied using other datasets such as BabyEars and FERMUSIII (Slaney and McRoberts 2003 and Schuller 2002). In both datasets, almost the same results were achieved. Using with BabyEars database, best accuracy result was 75% for speaker independent classification while 74.83% classification accuracy was obtained by using FERMUSIII. It had also been observed that the accuracy based on speaker dependent classification increased the accuracy to 89.12% in FERMUSIII datasets (Schuller 2002).

Support vector machine (SVM), based on the use of kernel functions to map the original features in a nonlinear manner, is an important example of a generic classifier. FERMUSIII dataset was also used while applying SVM (Schuller 2002). Best accuracy results were 81.29% and 92.95% for independent and dependent speaker classification, respectively.

Another classifier commonly used in emotional speech recognition applications is the artificial neural network (ANN). Since ANN is more effective in nonlinear transformations modeling than GMM and HMM, ANN has some advantages over them (El Ayadi et al. 2011). In addition, classification accuracy is generally better than HMM and GMM when the training set is small. Nicholson et al. (2000)classified the emotions of joy, mockery, fear, sorrow, disgust, anger, surprise and neutral taken from a local emotional voice database with help of ANN. The best classification accuracy was only 52.87%, was lower than other classifiers. However, Petrushin (2000) achieved 70% average accuracy, while applying boot strap aggregation scheme to the network configuration.

Despite the fact that machine learning had been in used for years, deep learning became state of the art method thanks to the recent developments of GPU (Catanzaro et al. 2008). Deep neural networks began to be trained using much larger training sets in much shorter time. Sánchez-Gutiérrez et al. (2014) used deep learning methods which were Restricted Boltzmann machines (RBM) and deep belief networks (DBN) on the Spanish emotional database in their work. They obtained comparable results for RBM and DBN rather than other classifiers by selecting an appropriate parameter. Spanish emotional speech database is used for their purpose. And they found that with a suitable choice of parameters, RBM and DBN can achieve comparable results to other classifiers, when the parameters were correctly chosen.

In Chernykh et al. (2017), their proposed approach used deep recurrent neural network trained on a sequence of acoustic features calculated over small speech intervals. And special probabilistic-nature Connectionist Temporal Classification (CTC) loss function allowed them to consider long utterances containing both emotional and unemotional parts. They achieved two advantageous reasons: 1) Through the CTC loss function accounts for the fact that emotionality may be contained only in a few frames in the utterance. 2) It can predict the sequence of emotions for one utterance. Chernykh et al. (2017) showed that the results are comparable with the state-of-the art ones in this field. Moreover they analyzed model answers and error distribution along with human performance and came to the conclusion that emotion is a very subjective

notion and even if humans outperform computer the difference is not so significant.

2.5 Emotional Dialogue

There is a vast literature on emotional speech, however, the field that focuses the emotional recognition of the dialogue and the changes of emotional state due to the interaction still needs to be investigated. In most emotional dialogue studies, an emotional speech of each person performing the dialogue is examined separately. There are very few studies that examine the response of listener which triggered by the emotional speech of speaker and how listeners react verbally. In particular, there hasn't been a study to our knowledge that examines the relationship between the verbal response of the listener and the emotional speech created by the speaker. This relationship was investigated in our study and analyzed whether the accuracy rate of classification could be increased by adding the information arising from responses given to emotional statements.

The call-center application dataset and Intelligent Tutoring Spoken dialogue system (ITSPOKE) dataset were used to search classification on the dialogue system. ITSPOKE was a dialogue system that includes the student records while they were taking physics oral exam (Litman and Silliman 2004). Students were asked to provide feedback on the answers to the questions so that they could reflect more emotions.

Litman et al. (2003) used ITSPOKE datasets and hypothesized that their system enhances the spoken dialogue tutorial while predicting emotion itself. As a continuation of previous research mentioned, Forbes-Riley and Litman (2004) automatically extracted features and classified three type of emotions as positive, negative and neutral. They found the prediction accuracy as 84.75%, and it was 44 % higher than error reduction over a baseline.

In one of call-center application research, Lee and Narayanan (2005) aimed to classify negative and non-negative emotions from a dialogue. They combined acoustic, lexical, and discourse features to enhance the accuracy result on emotional dialogue, unlike previous researchers who used only acoustic features. With the method of combining those features, they pointed that they improved classification accuracy by 40.7% for males and 36.4% for females.

In another article, Pittermann et al. (2010) argued that the knowledge about the textual content of an utterance could improve the recognition of the emotional content. However, after their tests and research, confidence measures provided by the combined speech-emotion recognizer feature a high comparability, but no distinction between emotion recognizer confidence.

Lubis et al. (2014) researched the emotional triggers obtained from a colorful database (SEMAINE database) by applying SVM. In dialogue, emotion was triggered depending on the meaning of the word which includes one of four emotions; optimistic, sensible, depressed and angry. The results showed that classification accuracies were random for all categories.

CHAPTER 3

MATERIALS AND METHODS

3.1 Participants

The ethics committee of Middle East Technical University approved the experimental procedure, which included both visual and auditory parts. None of the participants had visual or/and auditory problems. Before the experiment, an informative text about the experiment was read to the participants and a statement that says they were volunteer participants was signed. Participants filled a form of information on demographic, educational and theatrical history (Appendix A). In addition to this, participants answered the Eysenck personality questionnaire (Eysenck and Eysenck 1975). Each of these questionnaires will be discussed throughout the following sections.

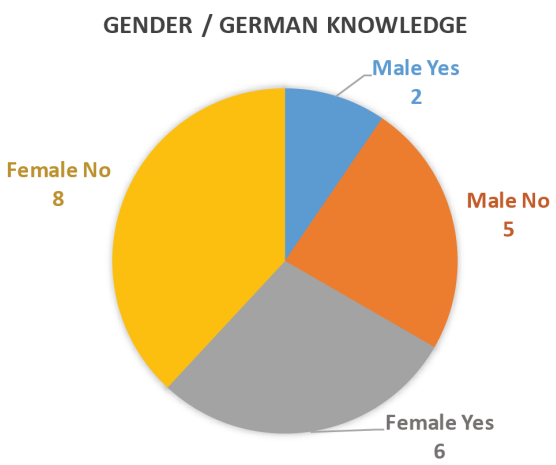
3.1.1 Demographic Information

Twenty-one native Turkish speakers participated in the experiment (Appendix B). All participants had an average age of 26 (range: 19 - 45). From all participants, 14 participants were females and 7 of them were males. Out of 21 participants, eight of them had some German knowledge. The intersection of demographic information with gender is given in Figure 3.1.

3.1.2 Eysenck Personality Questionnaire

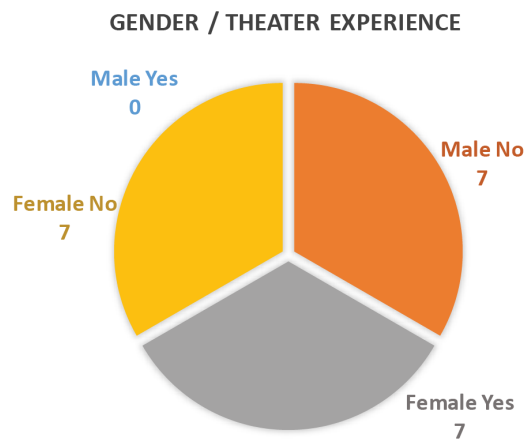
Eysenck personality questionnaire (EPQ) is a group and individual test, which performed to whose age is 16 and above (Eysenck and Eysenck 1975). There is no time limit. It is measured by the behavior of people. EPQ is a 101 items test consisting of 4 parts: Psychoticism (P), Extroversion (E), Neuroticism (N) and Lie (L). The second test is extroversion scale. It is a test scale consisting of 21 items. The subject tends to show extrovert symptoms when the score is equal or bigger than 13 points. For the Turkish version, the questionnaire designated in work of Karancı et al. (2007) was used.

A Typical Extrovert:



(a)

Gender/German Knowledge	Yes	No	Total
Male	2	5	7
Female	6	8	14
Total	8	13	21



(b)

Gender/Theater Experience	Yes	No	Total
Male	0	7	7
Female	7	7	14
Total	7	14	21

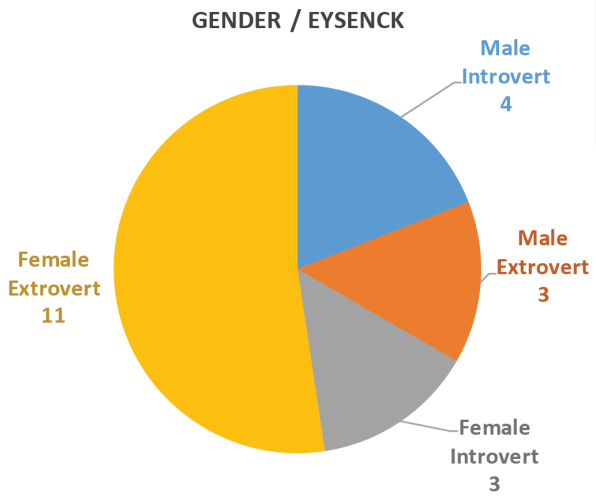
Figure 3.1: Demographic information of participants and distribution of skills according to gender. There were 14 female participants and 7 male participants. Out of 21 participants, 8 participants knew German to some degree (but not advanced), 13 participants didn't. On the other hand, 7 male participants had theater acting experience, while none of the males had.

- is a social person who likes meeting new people.
- has many friends.
- Is looking for a person to speak with
- does not like working and reading on his own.
- searches for exciting things.
- does not avoid taking risks.
- always sticks her/his nose into other people's business.
- enjoys humor.
- is always ready to answer, likes the change.
- is comfortable with new people.
- wants to laugh and be cheerful.
- wants to be on the move and likes it.
- tends to be aggressive and lose his temper quickly.
- does not keep their emotions under control at all times.

A Typical Introvert:

- is calm and is through with many things.
- enjoys books and reading more compared to others.
- stays away from people, except the close friends.
- does pre-plan what to do before he makes any attempt to thoroughly plan.
- does not have any confidence to act on a sudden impulse.
- gives importance to moral values.
- very often keeps his emotions under control and get aggressive rarely
- is defined as a person who does not lose himself.

According to Eysenck questionnaire, 7 of the participants were introverts, 14 of them were extroverts. The intersection of Eysenck result with gender is given in the following Figure 3.2.



Gender/Eysenck	Introvert	Extrovert	Total
Male	4	3	7
Female	3	11	14
Total	7	14	21

Figure 3.2: Eysenck personality of participants and distribution of personality according to gender. Eleven female and three male participants had extrovert personality.

3.2 Experiment

Experiments were carried out in the Neuro Signal Laboratory of the METU Informatics Institute. A cube, which was acoustically insulated and shielded like a Faraday cage was used for experiment procedures. Participants were located inside this cube which had a comfortable chair in front of a LCD monitor on top of a desk. Audio reached participants through headphones which had volume tuner. Voices of participants were recorded using a professional Snowball ICE microphone. Experiment started with a training stage where the subjects were presented with 28 unrelated trials. After the training, participants continued on the main part until they finished. In total, each experiment included 294 trials with a break in every 21 trials. The experimental interface was implemented in MATLAB (R2013a, The Mathworks Inc., Natick, MA) using the Psychophysics Toolbox (Brainard 1997).

3.2.1 Experimental Material

3.2.1.1 Berlin Database of Emotional Speech

Emo-DB (Burkhardt et al. 2005), funded by Technical University of Berlin, was a publicly available emotional German audio database performed by 5 actors and 5 actresses (Mean age: 29.7, range: 21 - 35). The actors were asked to read 10 sentences in 7 different emotions which are anger, boredom, disgust, anxiety/fear, happiness, sadness and neutral. After some of them were eliminated, 535 utterances were built (Table 3.1).

The German database was chosen because:

Table 3.1: The number of emotions, emo-DB contains

Anger	Boredom	Disgust	Fear	Happiness	Sadness	Neutral
128	81	46	68	71	62	79

1. Emotions are more distinct because professional actors are used.
2. It provides good accuracy rate for emotion by using any machine learning algorithms (Yuncu et al. 2014).
3. Also, the correct classification rate of emotions is high while testing by participants. (Yuncu et al. 2014).
4. German is not a widely known language in Turkey. So the emotions of the participants can be measured only with sound without looking at the content integrity of sentences.

Both rates of naturalness and classification for emotions were given in Emo-DB documentation. According to this information, 46 sentences were chosen for each emotion considering the highest rate of correct classification by human ear condition. In total, we used 322 (46 sentences x 7 emotion) voice recordings in our experiment set. From these recordings, 294 were played during the actual test, and 28 played in practice.

3.2.1.2 Establishment of Turkish Sentences

According to Osgood et al. (1957), an emotion emanates from three different axes in various forms, namely pleasure, arousal and dominance. Gökçay and Smith (2012) have realized a normative study, named as TUDADEN, where they researched on a survey of these three main axes instead of emotions themselves. In one part of their study, they used the questionnaire to label the words semantically, whether the words were neutral, negative or positive. Semantically neutral words were adapted to on our own experimental work by constructing "neutral" sentences. The words that were classified as neutral by the highest vote were selected during sentence formation.

All statements were made to include all verbs, general truths, transposition statements and proverbs. Semantically neutral words were chosen so that the participants would not be disturbed by the meaning of the words, when they react vocally. Turkish sentences in our database were comprised of 6-11 syllables (average = 9.28) So that the side effects due to the length of the sentences could be prevented (Rao et al. 2013).Turkish sentences database created is given in Appendix C

3.2.1.3 Establishment of Math Questions

Each trial ends with a mathematical question consisting of four arithmetic operations as one of the following math sequences where *a, b, c* is each positive numbers and smaller than 10.

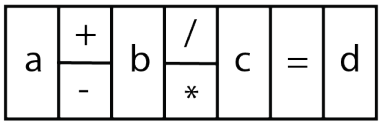


Figure 3.3: Structure of arithmetic equation.

Figure 3.4: Experiment Flow. The flow contains three step; 1) listen German audio, 2) read Turkish sentence and 3) solve math equation

While designing the set of the equation randomly, it was arranged that half of the mathematical equality were correct and the other half were wrong. Participants were asked to respond whether the equality was correct or not. One mathematic question is asked between two trials in order to make the participant less affected from the previously perceived emotion.

3.2.2 Experiment Procedure

The experiment consisted of 14 blocks and each block contains 21 trials. Both actual experiment and practical session contained the same steps. Each session was composed of three intervals as follows ((Figure 3.4):

1. One German sentence was played on each trial where a fixation cross was shown on the screen. The audio record was expected to evoke an emotion on the participants.

2. A randomly chosen and emotionally irrelevant Turkish sentence was presented on the screen for 5 seconds. Participants read it as a reaction of the evoked emotion of what they listen.

3. After completing this process, the participants were tested with an arithmetic equality problem. This was applied to insert some time distance between sequential emotional responses.

The experiment continued by repeating these steps. All paradigms were all randomly selected and each trial got 10 seconds. An example is given in Figure 3.5. Steps are explained in more detail in the following sections.

Example : A sample experiment dialogue

Aduio 1: Die wird auf dem Platz sein, wo wir sie immer hinlegen. *[Speakers16 (Female) reads angrily]*
It will be in the place where we always store it.

Subject 1: Çorbayı yudumladılar.
They sipped the soup

Math 1: (1×4)-2=-2

Audio 2: Das will sie am Mittwoch abgeben. *[Speakers14 (Female) reads with disgust]*
She will hand it in on Wednesday.

Subject 1: Işıkları söndürdü.
He turned off the lights.

Math 2: (4×2)-8=0

Figure 3.5: An example of experiment dialogue. Two different trials of the experiment are shown. 1) Listening 2) Reading 3) Math Part

3.3 General Analysis of Databases

In our study, analysis of trials was carried out through two separate datasets and their hybrid data. One of these two datasets was the Emo-DB database containing the audio recordings that the participants listened in German. The other dataset, which we named Response-DB, contains Turkish responses to the Emo-DB recordings. All processes described in this section were applied to these datasets (Figure 3.6). These operations were pre-processing, feature extraction (MFCC, Formant, Pitch, Moment), feature scaling and classification (MLP), respectively. Since same processes were applied to both datasets, they were described in a general title. The only differences were classification parameters for different datasets. Thus, analyses of those parts were explained separately. General data analysis was performed using in-house build scripts in addition to MATLAB-based toolboxes (R2017a, The MathWorks, Inc., Natick, MA).

3.3.1 Data Preprocessing

Speech signals often contain many silences and noises. Hence, while detecting the pure speech segments in speech analysis, it is necessary to apply the preprocessing methods such as denoising and silence removal.

3.3.1.1 Data Noise Reduction

Berlin emotion database was recorded with high-quality equipment and was noise-free. However, our database contained some background noise; even though it was recorded in an insulated room (Figure 3.7). It was necessary to remove the noise of the speech to prevent the noise from affecting the feature extraction.

Wavelet denoising algorithm (MATLAB Wavelet Toolbox; R2017a, The MathWorks,

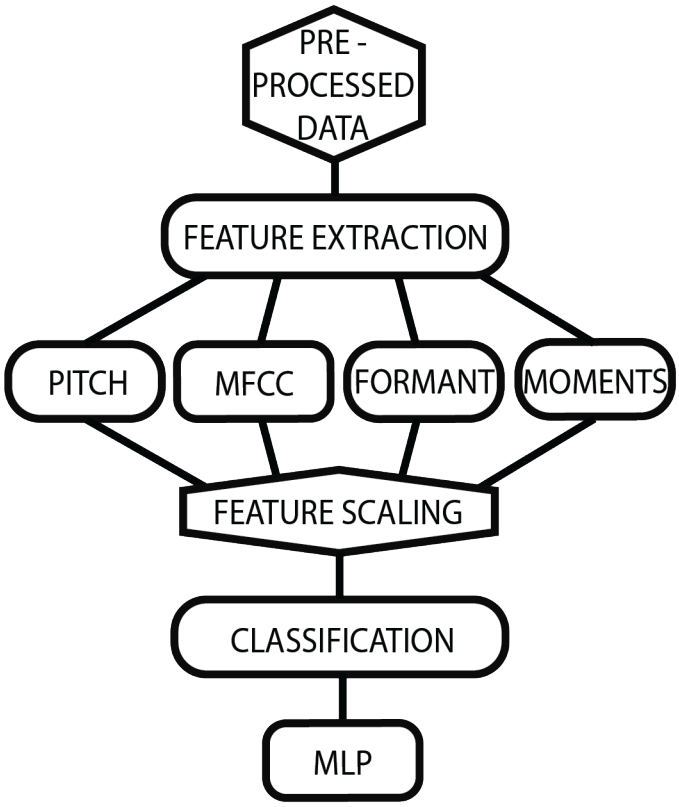


Figure 3.6: Process flow of the speech signal analysis. Processes in order; preprocessing, feature extraction, feature scaling, classification.

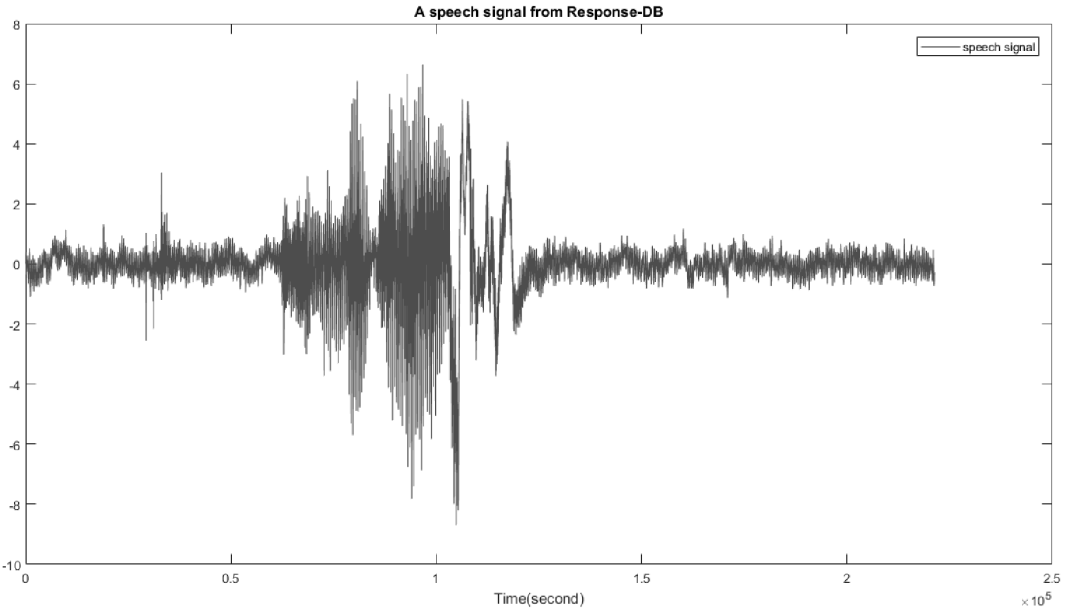


Figure 3.7: Raw data from speech response to Emo-DB.

Inc., Natick, MA), which was based on thresholding estimators, was used to reconstruct signal from the noisy one. The energy of a signal in the Wavelet transform was concentrated in a narrow area and the coefficients in this area are relatively larger than noise. Shrinking those coefficients removed noise or undesired signals. This method helped to obtain the higher quality signal and increase the accuracy of classification.

3.3.1.2 Silence Removal

Silence means that there is not a speech that is produced in a sentence. Since silence can be on any part of a speech such as the beginning, between or the end of it. In the division of speech, silent zones must be well-defined. Generally, in silence removal approach, two feature segments were extracted from the windowed speech signal and the threshold values of those segments were calculated to be applied to their respective regions. After that, consecutive segments were merged. There were several preprocessing steps before applying a threshold and the most popular one was known as Short Time Energy (Jasmine et al. 2016). 50 ms frame size was chosen for short term windows. Since noisy and silent part of the speech energy was lower than voiced speech part, the energy of the signal and spectral centroid were calculated for each frame according to the equations below. Finally, the maxima of the segments, which were obtained from the threshold value mentioned above, were found.

$$E_x = \sum_{n=-\infty}^{\infty} |x(n)|^2 \tag{3.1}$$

$$C_i = \frac{\sum_{k=1}^{N} (k+1) X_i(k)}{\sum_{k=1}^{N} X_i(k)} \tag{3.2}$$

$X_i(k)$, $k = 1 \ldots N$, is the Discrete Fourier Transform (DFT) coefficients of the i^{th} short-term frame, where N is the frame length.

3.3.2 Feature Extraction

Emotional speech recognition contains three basic operations: signal modeling, labeling and matching. Signal modeling is parameterizing the signal to efficient features whereas the labeled features whether they are correctly classified or not. While studying on a speech recognition system, features are constituted the inputs of the neural network. Feature categorizing is important, as it is a pre-selection of efficient features. Prosodic features give a valuable information of what the emotion of the signal is. In general, prosodic and acoustic features are used together (Brown 2006). We extracted MFCC (20), Formant (3), Pitch (1) and Moment (6) features from speech signal for all datasets as seen in Figure. Totally, 30 features were extracted from both datasets.

3.3.2.1 Acoustic features

Acoustic features usually analyze the acoustic speech signal with parameters effectively. Feature extraction is based on traditional methods found in other engineering approaches, which solve a specific problem. Generally, short segment based methodologies are implemented in the extraction of acoustic features. For example, Fast Fourier Transform (FFT), the basis of Cepstral coefficient extraction method, is frequently used in acoustic feature extraction. Diversified acoustic features are tailormade for identifying each information of various phoneme (Moattar and Homayounpour 2012).

Cepstral Coefficients

Most commonly used acoustic features are Mel-Frequency Cepstral Coefficients (MFCC) and variants of this method, which are delta MFCC, and delta-delta MFCC (Bridle and Brown 1974; Mermelstein 1976; Davis and Mermelstein 1980,). There is an efficient use of MFCC features in the emotional speech recognition. Other spectral features such as Linear Prediction Coefficients (LPCs), Linear Prediction Cepstral Coefficients and Perceptual Linear Prediction (PLP) were the main coefficients for emotion classification (Ververidis and Kotropoulos 2003).

Cepstral coefficient extraction methods are based on converting time domain signal to frequency domain signal. Obtaining the power spectra of the signal, the filter bank is used to filter the specific band of the frequency and scale to the logarithmic space. While the Mel-scale filter is used for MFCC, bank-scale for PLP (Hermansky 1990; Stevens et al. 1937). In MFCC, the logarithmically scaled filter bank is converted back to the time domain and the amplitudes of the filter are used as Cepstral coefficients.

It is assumed that a human auditory system processes the speech signal nonlinearly. As is known, the lower frequency components of a speech signal contain more features that are phonetic.

Mel-scale filter bank was used to obtain lower frequency components. The Mel fre-

quency cepstrum was short-term power spectrum, obtained by applying the discrete cosine transform (DCT) to the log spectrum of the signal (Moattar and Homayounpour 2012). It is given by the following equation to convert from normal frequency f to Mel frequency m

$$m = 2595 * log_{10}(\frac{f}{100} + 1) \tag{3.3}$$

MFCC was obtained by performing steps:

1. Preprocessed data was taken.
2. Hamming windows with a 20 ms frame length were shifted on the signal. Each shifting was overlapped with frame length of 5 ms.
3. Discrete Fourier transform (DFT) was applied on each windowed frame to compute magnitude of the signal spectrum.
4. By filtering the magnitude spectrum with a Mel-scale filter bank, Mel spectrum was obtained.
5. The logarithm of Mel spectrum was taken and DCT of log spectrum was computed.
6. DCT coefficients 1-20 were selected, the rest was discarded.

Formants Features

A formant is a condensation of acoustic energy at a certain frequency in a speech signal and provides an articulation of speech. Formant features are generally examined with the segmental feature extraction methods (Väyrynen 2014). The spectrum of the speech signal is analyzed to interpret peak parameter. Frequency peaks are sorted ascending and renumbered e.g. *F1, F2, F3*. LPC is most used extraction method for formant features (Atal and Hanauer 1971; Makhoul 1973).

3.3.2.2 Prosodic features

In phonetics, to understand information about meaning and structure of the utterance, prosodic (supra-segmentally phonology) features are used such as pitch, loudness, tempo, and rhythm. Prosodic features contain linguistic and phonetic information. It is difficult to distinguish between acoustic and prosodic features. Since prosodic features are categorized as segmentally extracted features, some acoustic features can be seen as prosodic. Speakers use prosody to add fluency and depth to expressions and emotions with altering prosodic features in their speech.

Fundamental Frequency

The lowest frequency of the periodic speech signal is called as fundamental frequency (F0). The change of the fundamental frequency has a significant connection with emotions in dialogue. It is also known that changes in fundamental frequency cause changes in other prosodic features such as time and energy. Correlation between emotional states and spectral features is valuable for recognizing not only what the meaning of a text is, also how it is stated (Rao and Koolagudi 2013). Sun (2002) improved pitch determination algorithm by calculating the Subharmonic-to-Harmonic Ratio (SHR), which estimates spectrum of pitch on logarithmic frequency scale. In our work, pitch was extracted by an algorithm based on Sun (2002) work.

Intensity

Intensity is the energy transmitted in seconds across a unit field in a speech sound and is proportional to the square of the pressure change.

$$normalizedIntensity = \frac{\sum_{n=t}^{t+N-1}(X_n)^2}{N} \tag{3.4}$$

where X_n represents the signal x at time sample n and N is the number of time samples.

The logarithmic scale of the normalized intensity is better suited for the human auditory system. The intensity has two disadvantages that other frequency features don't. These are subjective parameters and recording conditions such as background noise and microphone type.

Duration

Prosody period is extracted from the timing and duration of the speech parts during the speech analysis performed to determine the pitch. Duration can be calculated by the ratio of voiced, unvoiced and silent sections. The classification of pauses and voiced segments can be used to calculate correlations between articulation and speech rate. However, the correlation is personality dependent e.g. culture, language or mood (Rao et al. 2013). The intensity and duration features were discarded, due to the uncertainty they created (e.g. subjective parameters and environment conditions).

3.3.2.3 Statistical Features

Statistical modeling of a speech signal is explained in this section. Statistical features are extracted from each speech signal and added to the features dataset. Table 3.2 provides the descriptions and formulas of those features.

3.3.3 Feature Scaling

Statistical normalization has especially been utilized in statistical data processing fields such as data mining. The purpose of the method is reduction of an excessive

Table 3.2: Description and formula of statistical measurement on speech signal

Statistical Modeling		
Features	**Description**	**Formula**
Mean	Avarage value of the speech signal.	$\mu = \frac{1}{N}\sum_{i=1}^{N} A_i$
Std	Standard deviation is a measure of how spread out the speech signal values are.	$S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} \lvert A_i - \mu \rvert^2}$
Var	Variance of the speech signal, is just the square of standard deviation.	$V = \frac{1}{N-1}\sum_{i=1}^{N} \lvert A_i - \mu \rvert^2$
Skew	Skewness is a measure of the asymmetry of the signal around the mean.While skewness of left tailed distribution is negative, skewness of the right tailed distribution is positive. If skewness is zero,distribution is normally distributed.	$s = \frac{E(x-\mu)^3}{\sigma^3}$
Kurt	Kurtosis is a measure of whether the signal is heavily tailed (has more outliers) or light tailed (has less outliers) compared to normal distribution.	$k = \frac{E(x-\mu)^4}{\sigma^4}$
Momentum	k^{th} moment of the speech signal.	$m_k = E(x-\mu)^k$

contrast within the data. Another use is to compare the data in different scaling systems. The aim is to make the data available in different systems to a common system so as to make them comparable.

In our research, *Linear Normalization* method was used. The largest and smallest values in a dataset were considered and all other data were normalized according to these values. The purpose here was to normalize the smallest value to 0 and the largest value to 1 and to spread all the other data to 0-1 range..

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3.5}$$

According to the formula given above, the value of each feature column, obtained

MFCC	..	Formant	Pitch	Moment
-1,72		102,85	178,70	2,43
-0,78		83,70	177,30	2,40
-1,94		259,21	190,61	4,17
:	:	:	:	:
-1,20		101,22	190,98	2,14

Min-Max Normalization

Figure 3.8: Normalization on the vertical value of features

from all the data extracted, were normalized vertically (Figure 3.8).

3.3.4 Feature Classification

Rosenblatt (1960) first used Perceptron in visual perception modeling. Perceptron is one of the oldest neural networks, although it is extremely limited and is based on the principle that a nerve cell produces a binary output by taking more than one binary input (Figure 3.9).

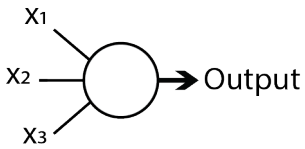


Figure 3.9: A example of Perceptron structure. It has three inputs (X_1, X_2, X_3) with their weights (W_1, W_2, W_3).

where W is weight, X is input value, equation checking whether the weighted sum is greater than a threshold.

$$output = \left\{ \begin{array}{l} 0 \text{ if } \sum_j w_j x_j \leq \text{threshold} \\ 1 \text{ if } \sum_j w_j x_j > \text{threshold} \end{array} \right\}$$

Perceptron can be used in problems that are divided into two parts by a linear function but is not sufficient for non-linearly separable. Backpropagation method is used to solve this problem (Rumelhart et al. 1986). The main purpose of it is to reduce the error between the expected output and the output produced by the network.

Multilayer Perceptron (MLP) is a feedforward neural network, which contains one or more hidden layers between input and output layers. In feedforward neural network, the data flow from input layer to output layer. The hidden layers process the information coming from the input layer and send it to the next layer (Nielsen 2015). MLP used backpropagation learning algorithm to train network (Rumelhart et al. 1986). An illustration of flow is as follows:

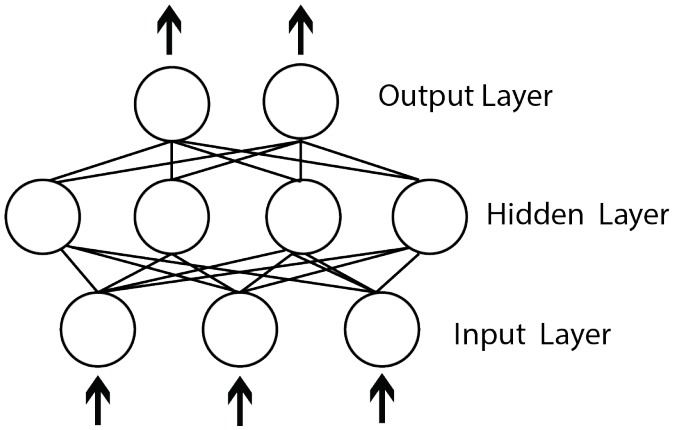


Figure 3.10: A multilayer perceptron structure with one hidden layer. It is built with 3 input, 4 hidden and 2 output nodes.

Problems that are not linearly separable can be solved by the MLP (Cybenko 1992). MLP uses supervised learning strategy (Rosenblatt 1960). The system generates a solution space representing the problem space by generalizing from the training set shown to it. For similar examples shown to test, this solution can produce results and solutions. It is possible to give a more specific result than the other machine learning methods if there is a small size dataset as in our case. Process steps of multilayer network are given below (Swingler 2011).

1. *Determination of dataset:* The part of the dataset divided to learning and test sets
2. *Determination of network topology:* The structure of the network was determined in order for the problem to be learned. It was determined how many input units, how many hidden layers, how many cell elements and how many output elements each hidden layer has.
3. *Determination of learning parameters:* Parameters such as learning coefficient, activation functions, momentum coefficient were determined in this step.
4. *Assigning network initial values:* Weight values of connections and threshold value were assigned.
5. *Demonstration examples from the training set to the network:* The network began to learn; examples of training set were shown randomly to change the weights according to the learning rules.
6. *Calculation of network output value for given input*
7. *Comparing the actual output to the expected output:* Calculated error values generated by the network.
8. *Changing weights:* Applied the backpropagation algorithm to reduce the learning error
9. *Determination of the stopping criterion:* Stopped learning when error falls below a certain value.

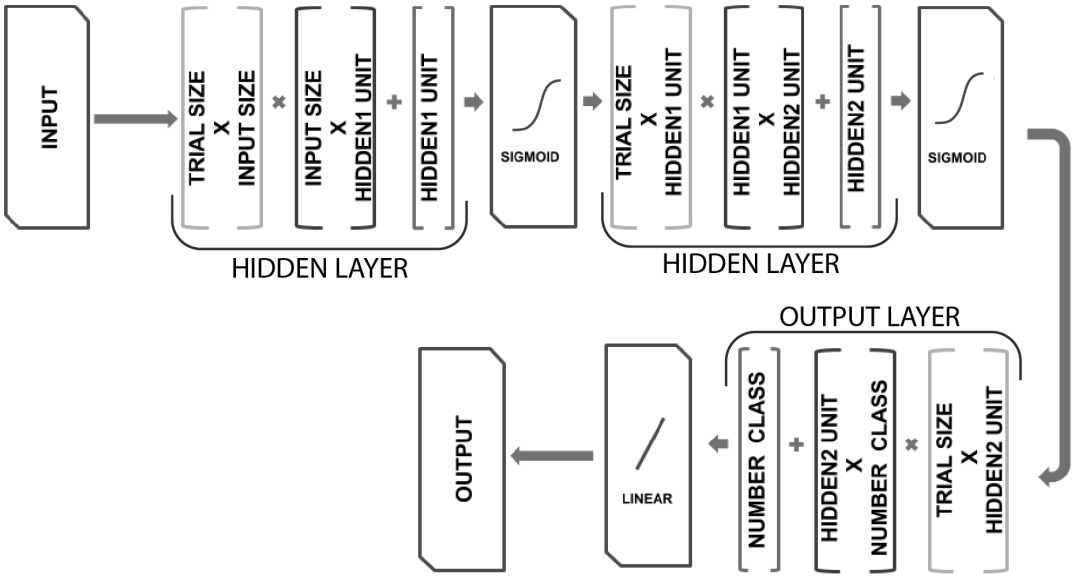


Figure 3.11: Structure of neural network. A structure of MLP with two hidden layers with sigmoid activation function and linear activation function in the output layer.

Network structure with two hidden layers was used for all datasets in our work. The structure is given in (Figure 3.11).

3.4 Analysis of Berlin Audio Data

All the methods described under the general analysis title had been applied to the Berlin audio (emo-DB) dataset and the outputs were obtained and explained in Figure 3.12 - Figure 3.16 in this section.

1. Preprocessed Data

As you can see in Figure 3.12, one sample of voice record from emo-DB dataset is a noise free and clean. For this reason, the distinction between the silent and speech parts is clearly visible.

2. Extracted and Scaled Features

After preprocessing, feature extraction was performed, and each of these features were scaled in itself.

As previously mentioned in the feature extraction section, the *Mel-Frequency Cepstral coefficients*are one of the most important features of emotion recognition research. MFCC feature was extracted from the Emo-DB dataset. MFCC features for all emotions are compared with each other. In order to do this, the standard distribution of each signal categorized in its own emotion was computed and visualized as a line graph. An example of this combination (anger vs. happiness) is given in Figure 3.13. After analysis, it can be said that the coefficients of anger and happiness are

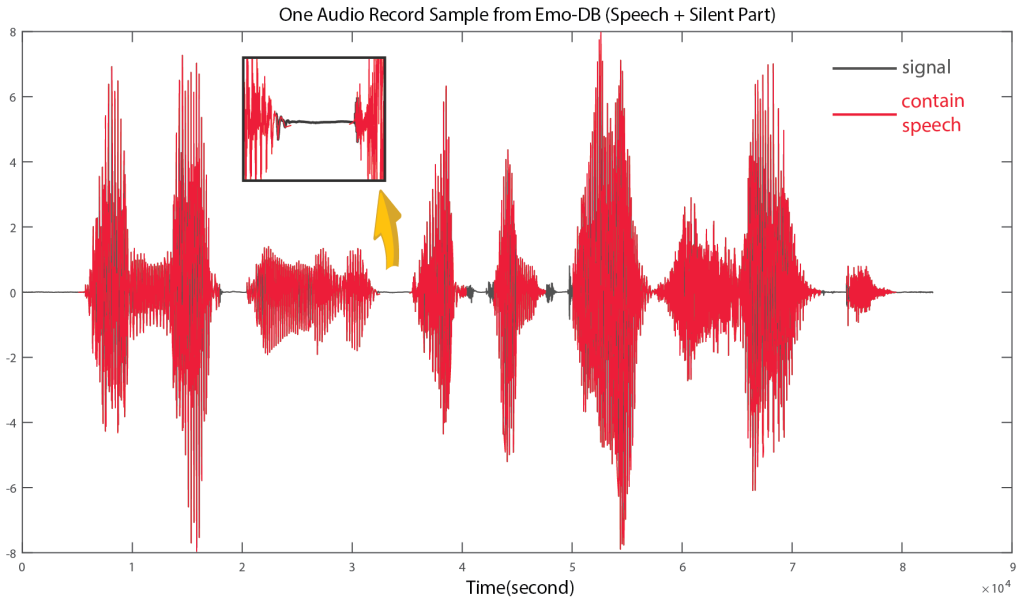


Figure 3.12: Speech signal removed the silent part from Emo-DB audio record. The gray line represents the signal itself, the red line contains speech part. A part of the signal is enlarged in order to better see the distinction between the silent part and the part that contain speech signal.

separated from each other.

Another extracted features were *formant* values of the speech signal, first three formants (*F1, F2, F3*) were taken into consideration in our work. Formant values got increased gradually (Figure 3.14).

Another essential feature of emotion recognition is *pitch* feature. This feature, which is known to reflect the voice characteristics of a person, has been compared according to emotions. The standard deviation and mean error of the categorized emotions were analyzed (Figure 3.15). The distributions of some emotions are narrow, while others are scattered over a wide range.The reason for differentiation on the pitch distributions of emotions is the sentences are acted by both male and female actors in the database. Pitch data varies depending on the gender (Gelfer and Mikos 2005). Figure 3.15 shows that features for some emotion are affected by gender more than other emotions, such as anger.

Statistical calculations were used as the fourth feature extraction method. Six basic statistical measurements were calculated, which were mean, kurtosis, skewness, standard deviation, variance, and the k^{th} moment of data for all emotions. Averages of these measurements are given in Table 3.3.

A comparison of statistical values for all emotions in Emo-DB. Each cell stands for the mean of all statistical features of Emo-DB signal.

3. Feature Classification

First, the dataset was divided into training and testing sets randomly for each run.

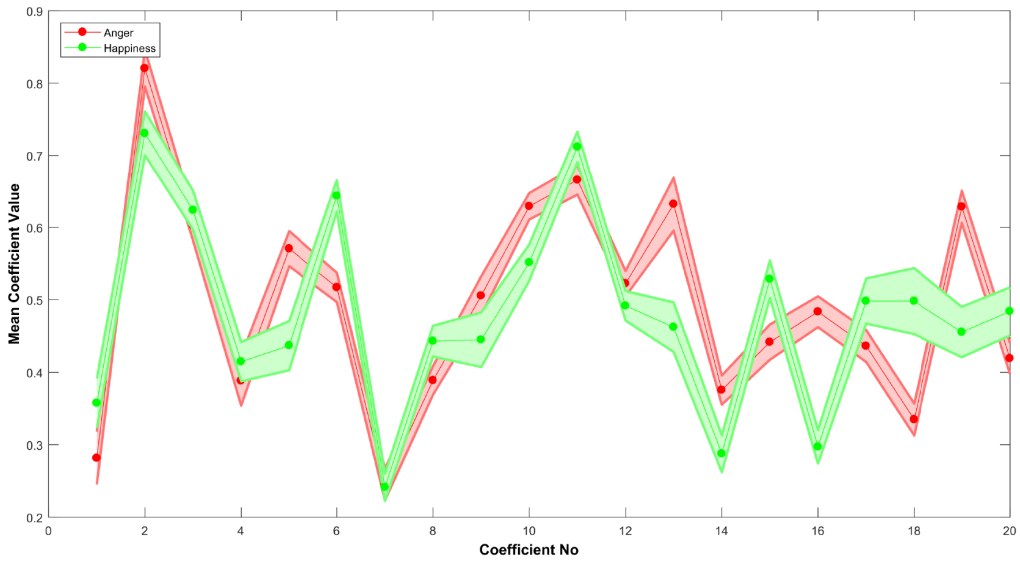


Figure 3.13: A comparison of MFCC coefficients for anger and happiness in Emo-DB. The green line represents mean value of the coefficients of signals categorized as happiness, whereas red line represents mean value of the coefficients of the signals categorized as anger. The shaded parts of lines stand for the distribution of the MFCC coefficients of categorized emotions. The distributions for almost all coefficients are separated from each other.

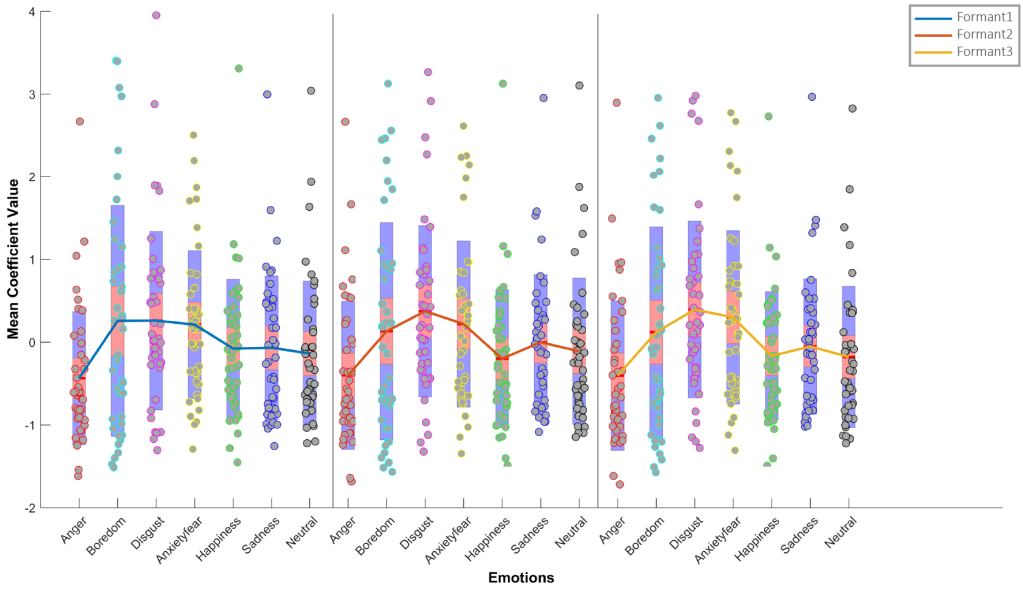


Figure 3.14: A comparison of formants (*F1, F2, F3*) for all emotions in Emo-DB. The red lines behind boxes represent the mean of data, whereas the standard deviation as the blue shaded box. Values for emotions were plotted as colorful circles.

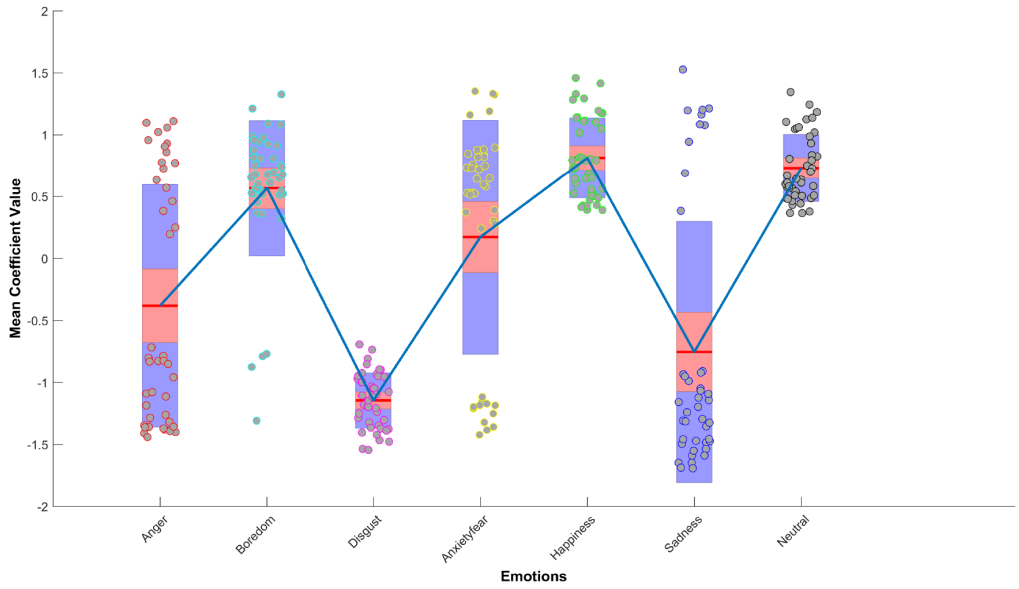


Figure 3.15: A comparison of pitch value for each emotion in Emo-DB. The red lines behind the boxes represent the mean of the data. The standard error of the mean (95% confidence interval) was drawn as the red shaded box, whereas the standard deviation as the blue box. Acquired values were plotted as colorful circles.

Table 3.3: Statistical feature values for all emotions in Emo-DB. Each cell stands for the mean of all statistical features of Emo-DB signal for all emotions where A, B, D, F, H, S, N are the first letters of emotions; anger, boredom, disgust, anxiety-fear, happiness, sadness, and neutral.

	Features	**A**	**B**	**D**	**F**	**H**	**S**	**N**
	Mean	0.579	0.455	0.46	0.229	0.483	0.66	0.383
	Kurt	0.469	0.203	0.266	0.259	0.279	0.232	0.322
Emo-DB	**Skew**	0.427	0.587	0.52	0.559	0.435	0.531	0.523
	Std	0.131	0.082	0.135	0.171	0.262	0.107	0.19
	Var	0.044	0.038	0.057	0.062	0.141	0.044	0.082
	Moment	0.879	0.194	0.43	0.059	0.342	0.411	0.128

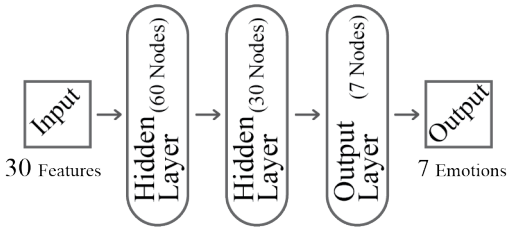


Figure 3.16: Emo-DB neural network structure. A two hidden layer MLP with 30 input nodes and 7 output label. Hidden layers were built with 60 nodes and 30 nodes in turn.

60 % of the dataset was determined as the training set, remaining % 40 was testing set. In other words, out of 294 speech samples, 175 were set as the training set, 119 samples were divided into the test set.

Second, the structure of neural network was built. Multilayer layer perceptron with emotions dataset from Emo-DB was built (Figure 3.16). A gradient descent with momentum and adaptive learning rate was used as the training method with backpropagation learning. Since 30 features were extracted, input layer had 30 nodes. The output layer consisted of 7 nodes representing emotions and outputs were one-hot encoded. Initial weights were distributed sharply peaked as Gaussian.

Finally, good hyper-parameters were set heuristically. A two-hidden-layer was built with first hidden layer had 60 nodes, the second one had 30 nodes (Figure 3.16). Starting learning rate was 0.01 and momentum constant was 0.9. Stopping criteria was controlled by mean square error until it reached the best accuracy goal which is zero or after 1000 iterations.

3.5 Analysis of Turkish Acoustic Response Data

Turkish acoustic response data are a database called Response-DB which contains the audio record of response to the Emo-DB dataset. All the method steps, applied to Emo-DB, were same for Response-DB dataset so feature descriptions are not detailed in this section.

1. Preprocessed Data:

Response data were noisy records because of the experiment environment (background noise) as you can see in Figure 3.17. First, wavelet noise reduction method was applied to remove this noise. Because of the noise, it was less efficient to isolate the parts that contain speech or silence in the Response-DB than in the Emo-DB.

2. Extracted and Scaled Features

In the same way, the feature extraction of Response-DB consists of 4 main groups: MFCC (20), formants (3), pitch (1) and statistical features (6). All the features were explained by figures and tables, it was done using the outputs for a female and a male participant. Minor reminders about the participants are also given the figures.

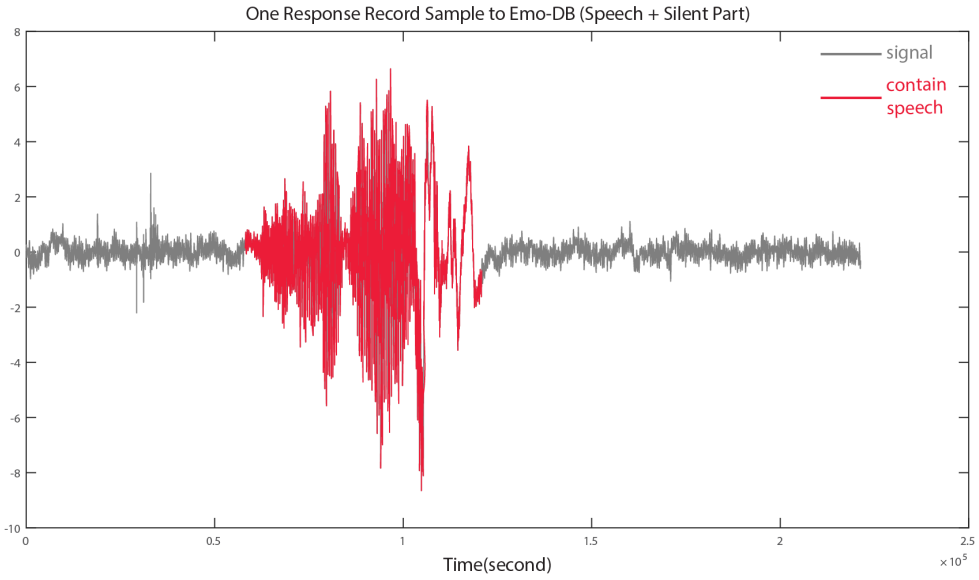


Figure 3.17: Speech signal removed silent part from Response-DB audio record. Grey line represents the response record; red line contains speech part.

MFCC features of subject 2 (female) and subject 7 (male) were given in Figure 16 and Figure 17. In order to compare different emotions, data of 4 different emotions were used to plot MFCC features. The distributions of the coefficients seem to be separated from each other in some values.

The formant and pitch features extracted from the recordings of subject 1 and subject 2 in the Response-DB and the distribution of these features are given in Figure 3.20 and Figure 3.21.

Finally, the statistical features of these subjects (Sub 1 & sub 2) are given in the Table 3.4.

3. Feature Classification For Response-DB dataset, the division of the trial (175 train + 119 test), the structure of the network (30 input + 7 output nodes), and the initial assignments of weights (sharply Gaussian distribution) were all done in the same manner as for the emo-DB, except for the number of layer nodes and the learning rate. After heuristic experiments, it was seen that hidden layers with 30 nodes and learning coefficients of 0.05 provided to obtain better accuracy for response dataset, so the structure of the network was rebuilt accordingly (Figure 3.22).

In order to understand that all emotions were not classified as chance, means that accuracy should be bigger than chance level, the target labels of the test set were randomly mixed and the target result were compared to the output result of network training. The comparison results are explained in the next chapter.

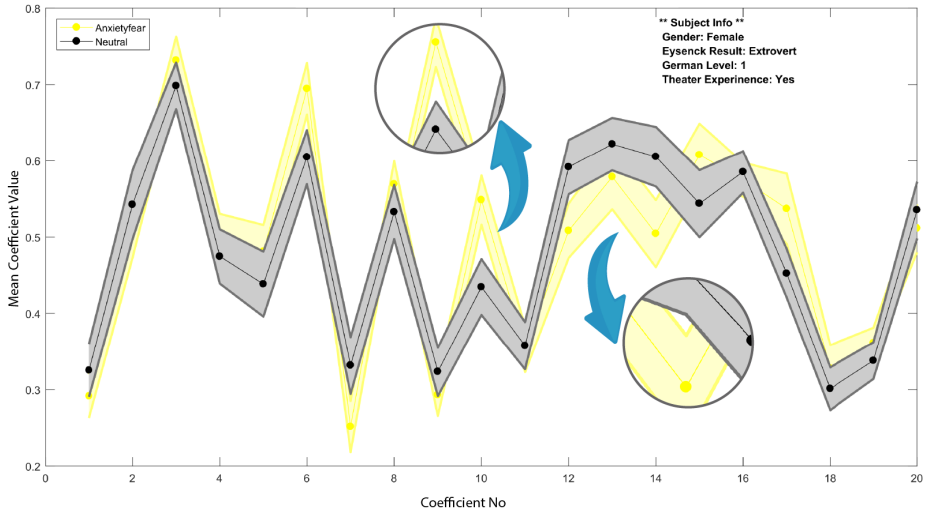


Figure 3.18: A comparison of MFCC coefficients for anxiety-fear and neutral emotions of Sub 2 in Response-DB. Grey and yellow lines represent the mean value of the coefficients of the signals categorized as neutral and anxiety-fear, respectively. The shaded parts of lines stand for the standard deviation of the coefficients.

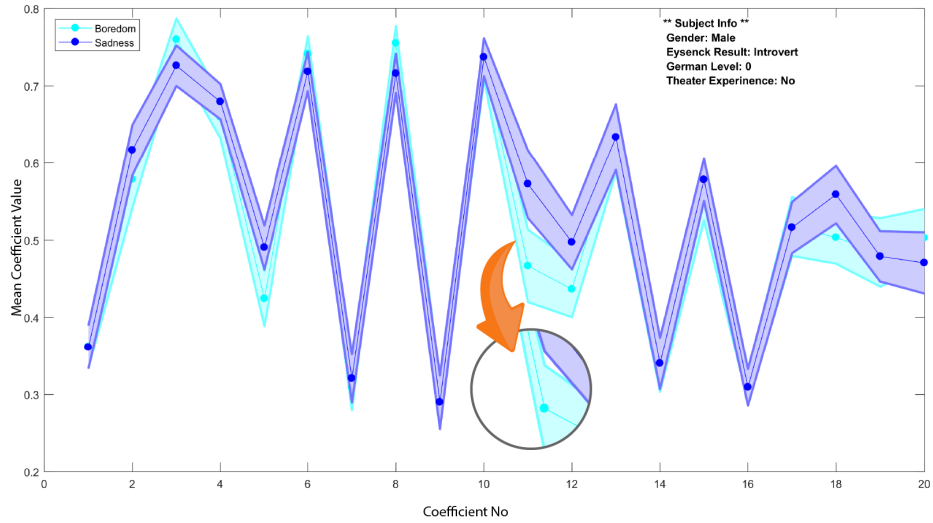


Figure 3.19: A comparison of MFCC coefficients for boredom and sadness emotions of Sub 7 in Response-DB. The dark blue line is coefficients of sadness and shading part represents its distribution whereas the same data are shown in light blue for boredom.

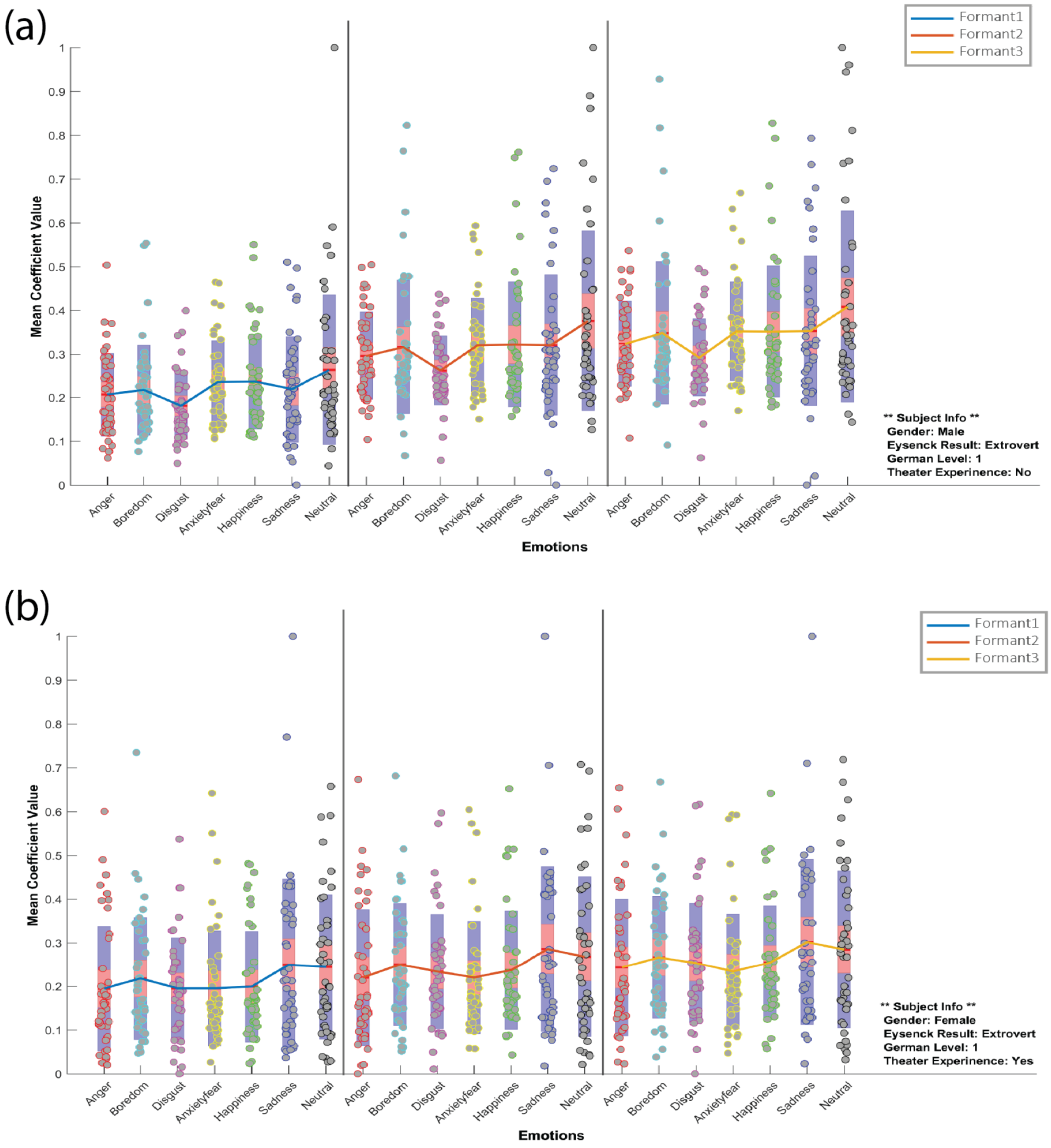


Figure 3.20: A comparison of formants (*F1, F2, F3*) for all emotions in Response-DB. (a) Formants of Subject 1 (b) Formants of Subject 2. Each block represents distribution of the *F1, F2, F2* according to emotion categories, sequentially.

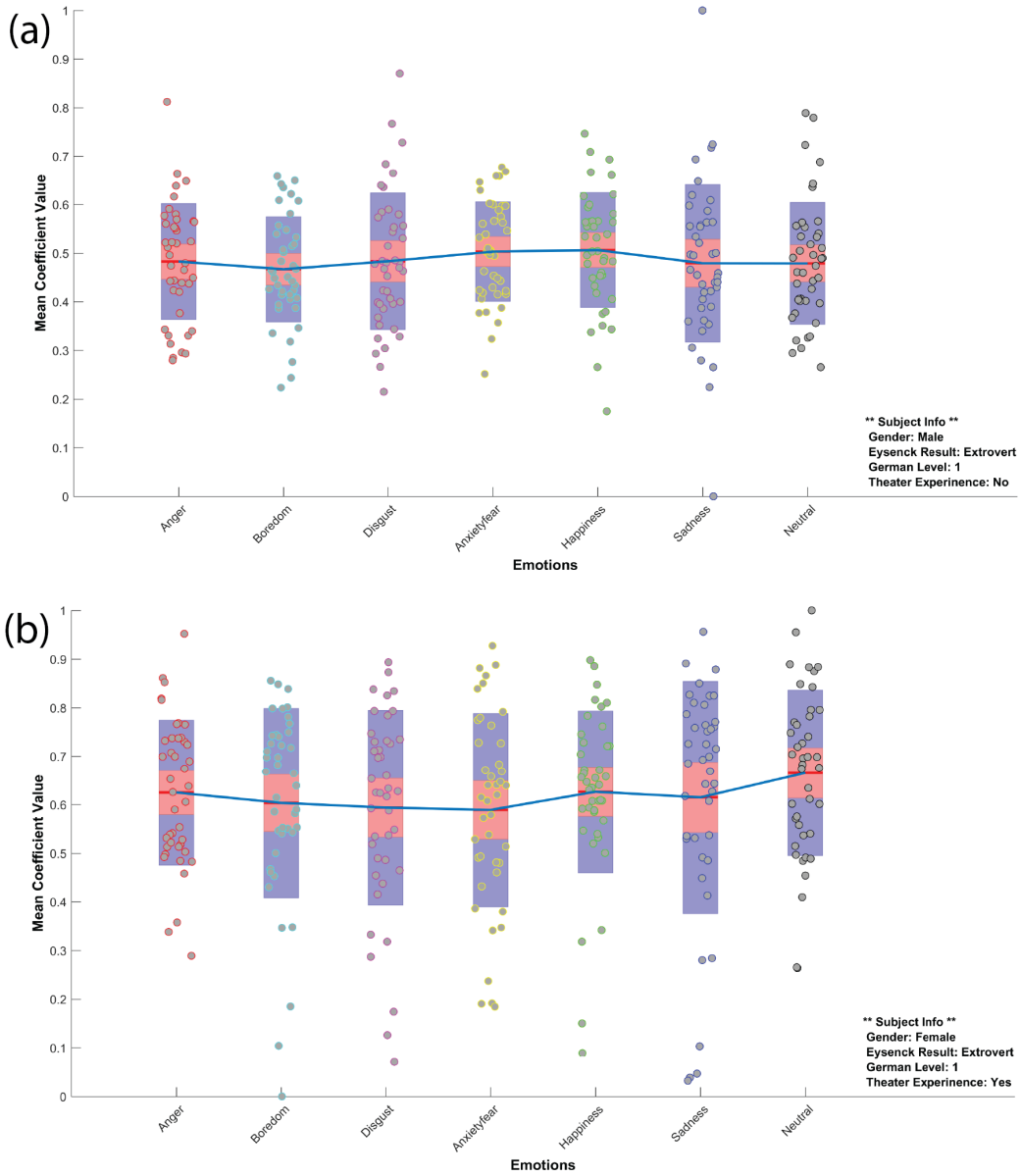


Figure 3.21: A comparison of pitch value for all emotions in Response-DB. (a) Pitch Value of Subject 1 (b) Pitch Value of Subject 2 Red line between the box represents mean value of data. Red box shows standard mean error while blue box shows standard deviation with 95% confidence interval.

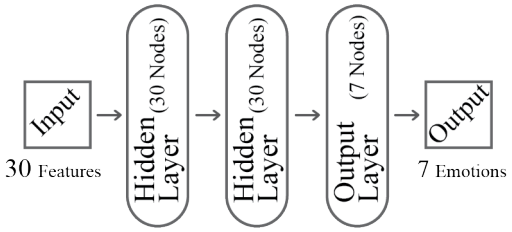


Figure 3.22: Response-DB neural network structure. A two hidden layer MLP with 30 input nodes and 7 output labels. 30 nodes were set for both hidden layers.

Table 3.4: Statistical feature values for all emotions (anger, A; boredom, B; disgust, D; anxiety-fear, F; happiness, H; sadness, S; neutral, N) in Response-DB. Each cell represents the mean of all statistical features; mean, kurtosis, skewness, standard deviation, variance and 3^{rd} moment.

	Features	**A**	**B**	**D**	**F**	**H**	**S**	**N**
	Mean	0.543	0.395	0.794	0.565	0.548	0.564	0.474
	Kurt	0.384	0.308	0.07	0.101	0.185	0.199	0.154
Sub 01	**Skew**	0.62	0.57	0.22	0.661	0.376	0.71	0.562
	Std	0.238	0.261	0.175	0.197	0.142	0.116	0.093
	Var	0.119	0.153	0.075	0.077	0.051	0.044	0.035
	Moment	0.351	0.852	0.062	0.221	0.867	0.918	0.973
	Features	**A**	**B**	**D**	**F**	**H**	**S**	**N**
	Mean	0.736	0.537	0.653	0.407	0.499	0.563	0.67
	Kurt	0.262	0.233	0.337	0.084	0.36	0.091	0.119
Sub 02	**Skew**	0.36	0.487	0.575	0.659	0.455	0.281	0.426
	Std	0.125	0.192	0.113	0.323	0.381	0.166	0.209
	Var	0.046	0.084	0.054	0.166	0.22	0.058	0.082
	Moment	0.854	0.15	0.967	0.906	0.746	0.028	0.036

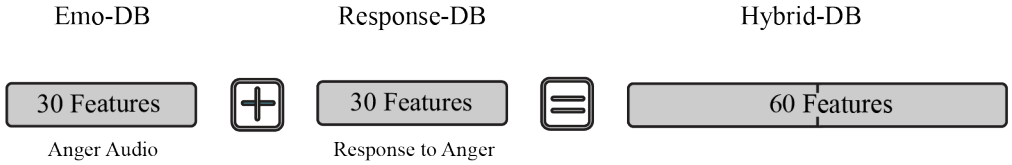


Figure 3.23: An example of merging procedure. Features of Emo-DB audio record that includes anger emotion and features of response to this specific anger audio record 30 features acquired form audio records of anger are merged with Response-DB to create Hybrid-DB.

3.6 Analysis of Hybrid Data

During the experiment, the subjects were asked to give an audible response to the emotion they were hearing. These two records (heard and responded) were matched with each other. The features of the records had been extracted and the new dataset (Hybrid-DB) was created without corrupting these matches. The merging of Response-DB and Emo-DB datasets were done as indicated in Figure 3.23.

In order to verify the correlation between the heard and responded records in Hybrid-DB, a specific emotion's data was merged with a random response data and the resulting database was compared to real results (Figure 3.24).

Learning of Hybrid-DB and *random* Hybrid-DB datasets was performed as well as learning of Response-DB.

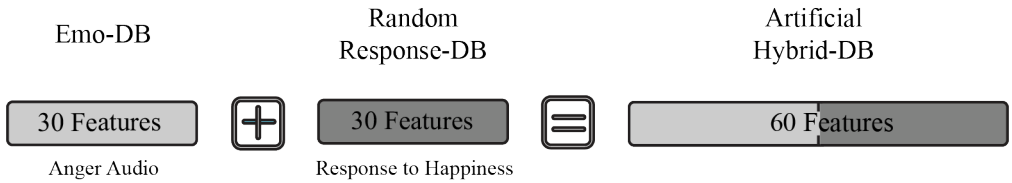


Figure 3.24: An example of *random* Hybrid-DB. Features of Emo-DB audio record that includes anger emotion and features of response to a happiness audio record are merged to create *random* Hybrid-DB.

CHAPTER 4

RESULTS

4.1 Berlin Audio Data Results

The Emo-DB set containing 294 trials were repeatedly trained and tested 10 times. Each time, the dataset was divided into training and testing datasets randomly (*175 train + 119 test sets*). The predicted outputs of the test set were calculated.

There were 7 possible predicted classes: anger, boredom, disgust, anxiety-fear, happiness, sadness and neutral. If the prediction for anger emotion was correct, anger-anger intersection cell was increased by 1 point. If not, cell for misclassified emotion was increased. The classifier made a total of 1190 predictions (each emotion contains 17 test sets which were tested 10 times, i.e. $7x17x10 = 1190$ *predictions*). Hence the best prediction score for an emotion would be 170 with the assumption that all of the data were correctly classified.

Table 4.1: Confusion matrix of emotion classification. Predictions were summed for 10 learning process. Column and row labels represent the first letters of emotions respectively (anger, boredom, disgust, anxiety-fear, happiness, sadness and neutral). Prediction of disgust emotion got the highest score compared to others.

		Truth Label								
		A	**B**	**D**	**F**	**H**	**S**	**N**	**Classification Overall**	**Precision (%)**
Predicted Label	**A**	112	3	2	17	15	1	4	154	72.7
	B	4	95	1	1	5	15	20	141	67.4
	D	8	1	131	16	12	5	2	175	74.9
	F	15	7	13	101	9	12	5	162	62.3
	H	25	6	9	18	117	4	2	181	64.6
	S	4	9	7	9	6	119	8	162	73.5
	N	2	49	7	8	6	14	129	215	60
	Truth Overall	170	170	170	170	170	170	170	1190	
	Recall (%)	65.9	55.9	77.1	59.4	68.8	70	75.9		67.60

In Table 4.1, the first main diagonal showed the number of correct classifications by the trained network. For example, 112 emotions were correctly classified as anger. This corresponded to 5.6% of all 1190 predictions of emotion. Similarly, 95 cases

Table 4.2: Accuracy results of each emotion for each learning trial in Emo-DB. Disgust emotion has the highest accuracy, whereas boredom emotion has the lowest accuracy. A:Anger, B:Boredom, D:Disgust, F:AnxietyFear, H:Happiness, S:Sadness, N:Neutral, All: Mean value of accuracies for each subjects.

# of Loop	A	B	D	F	H	S	N	All
1	64.7	52.9	64.7	71	88	71	71	68.9
2	64.7	64.7	70.6	35	82	82	65	66.4
3	76.5	47.1	94.1	65	77	71	77	72.3
4	58.8	64.7	82.4	65	41	71	71	64.7
5	58.8	58.8	76.5	77	77	59	53	65.5
6	76.5	70.6	88.2	47	65	47	94	69.7
7	70.6	35.3	94.1	65	59	94	88	72.3
8	47.1	52.9	70.6	65	59	77	71	63
9	52.9	52.9	58.8	65	82	77	100	69.7
10	88.2	58.8	70.6	41	59	53	71	63
Mean	**65.88**	**55.87**	**77.06**	**59**	**69**	**70**	**76**	**67.55**

were correctly classified as boredom, corresponded to 4.7% of all emotions.

25 of the anger emotions were incorrectly classified as happiness, which corresponds to 14.7% of all anger data. Similarly, 15 of the anger emotions were incorrectly classified as anxiety-fear, which corresponds to 8.8% of all anger emotions in the data. The other highest misclassified emotion was disgust in all anger data. Eight segments of the anger emotion were incorrectly classified as disgust, which corresponded to 4.7% of the total anger class.

Out of 154 anger predictions, 72.7% were correctly classified whereas out of 141 boredom predictions, 67.4% were correct.

Out of 170 anger cases, 65.9% were correctly predicted as anger and 35.1% were predicted as other emotions. An another example, out of 170 boredom cases, 55.9% were correctly classified as boredom and 44.1% were classified as other emotions.

Overall, 67.6% of the predictions were correct and 32.4% were wrong classifications.

Emo-DB test results showed that total recognition rate was 67.55% (Table 4.2). Disgust had the highest emotion rate with 77.06 %. On the other hand, recognition of boredom had the lowest rate of 55.87%. The highest score of a trial loop acquired was 100% with the class neutral. The measured lowest recognition rate of a trial loop was 35.3% for both emotions disgust and anxiety/fear. Recognition rates for each emotion, total recognition rates and the standard deviation of the accuracy results obtained from all learning loops are given in Figure 4.1.

4.1.1 One-way Analysis of Variance for Emo-DB

In ANOVA test, the null hypothesis was that the mean accuracies were equal within the emotions. *P-value* was close to zero ($6x10^{-33}$), which was much smaller than 0.05 (Figure 4.2a). Hence, the null hypothesis was rejected. The differences between the

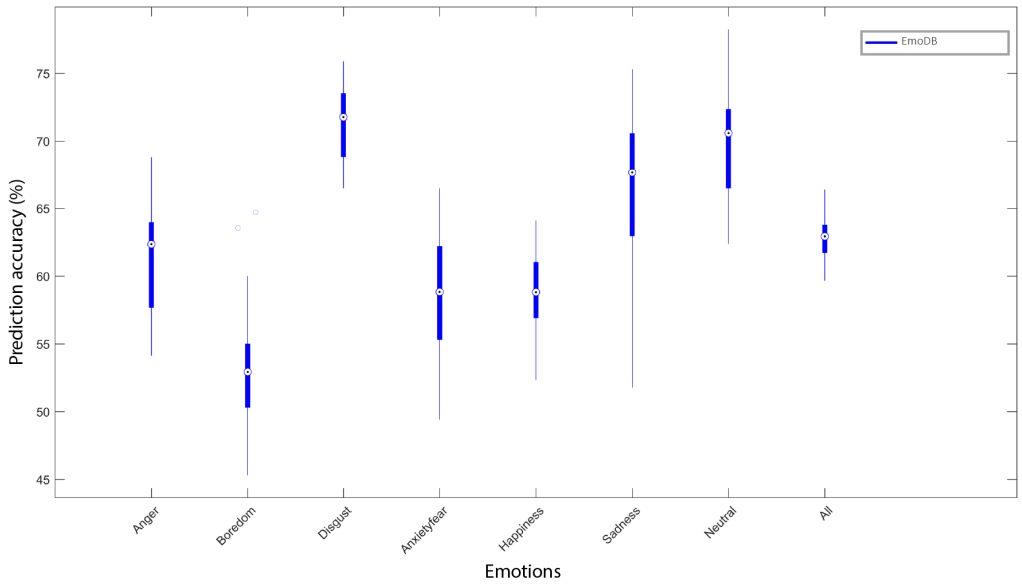


Figure 4.1: Test accuracy results for all emotions of Emo-DB. Disgust has the highest accuracy while boredom has lowest.

means of the emotions were strongly significant. Statistical analysis was performed with Statistics and Machine Learning Toolbox (R2017a, The MathWorks, Inc., Natick, MA)

Interval plot of the emotions was illustrated in Figure 4.2b. Each circle of the bar represented the mean of accuracies, for which 95% confidence level was taken. The controlled mean of accuracies, the intersecting mean of accuracies and the non-intersecting mean of accuracies were represented by three bars; blue, red and gray, respectively. Y-axis of the plot indicates the emotions, while X-axis gave information about the mean of accuracies.

The mean of accuracies for anger was significantly different than the mean of accuracies for all emotions, except for happiness and anxiety-fear. For this reason, the bars indicated by happiness and anxiety-fear were gray, others were red.

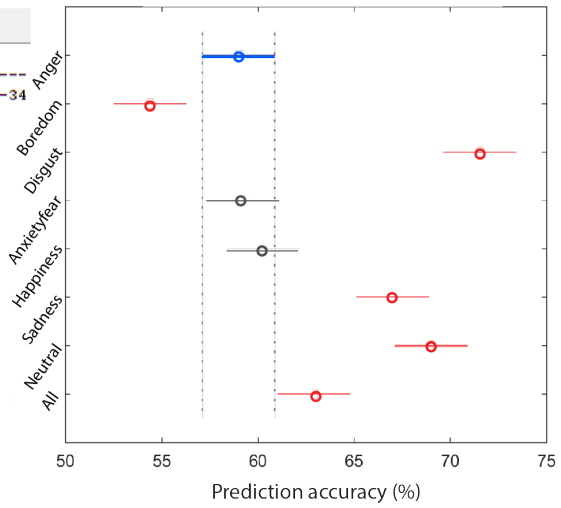
In order to show the differences between the mean accuracy rates, a color map table was depicted (Figure 4.2c). The diagonal of table were colored by gray. If there was a difference between the accuracies of means, the cells were colored as blue. For example, the mean accuracy for anger emotion was bigger than the mean accuracy for boredom emotion, so the cell was colored as light blue. However, the mean accuracy for anger emotion was less than the mean accuracies for disgust, sadness, and neutral emotions, the cell was colored as dark blue. There was a change in the color intensity according to whether it was greater than the controlled the mean accuracy for emotion, or not.

(a)

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Emotion	4973.84	7	710.549	43.48	6.11984e-34
Error	2614.43	160	16.34		
Total	7588.27	167			

(b)



(c)

Y \ X	A	B	D	F	H	S	N
A							
B	<						
D	>	>					
F		>	<				
H		>	<				
S	>	>	<	>	>		
N	>	>		>	>		

Figure 4.2: Two statistical analyses of the mean accuracies for Emo-DB. (a) ANOVA test of the mean accuracies (b) accuracy results of emotions (95% confidence level).The blue bar represents the interval for the mean of anger emotion accuracies. The red bars are the intervals for the mean accuracies for other emotions. Red bars mean the average accuracy for anger is significantly different from the mean accuracy for boredom, disgust, sadness and neutral. Grey bars mean that there is an overlap between intervals for mean accuracy. (c) a color map of the emotions' interaction for 7 emotions where A, B, D, F, H, S, N are the first letters of emotions; anger, boredom, disgust, anxiety-fear, happiness, sadness, and neutral. The light blue cells represent the interval of accuracies for *X* is bigger than the interval of accuracies for *Y*. If equality is opposite, then the cell is represented by dark blue.

4.2 Turkish Speech Response Data Results

There were 294 response trials matched with Emo-DB for each subject in Response-DB. Since 21 subjects participated the experiment, there were $294x21 = 6174$ trials in Response-DB. Hence, the trials for each individual were examined separately and 10 randomly obtained train and test sets were used for classification.

After that, the confusion matrix and test accuracy table were calculated for each trial as it was done for Emo-DB. For each subject, the mean values of these accuracy rates were tabulated in Table 4.4.

In Table 4.8, the subject-dependent average classification accuracy rates (%) range from 16.19% to 31.87%. When data of subjects was examined separately, it can be seen that the recognition rate of anger emotion was higher in most subject. Especially, subject 5 (Female) and subject 6 (Female) show a higher recognition accuracy for anger emotion than other subjects. The lowest recognition rate belongs to disgust emotion. For all emotions, recognition rates were greater than chance level 14.2% (1/7) as seen in Figure 4.3.

Table 4.3: Confusion matrix of emotion classification for Response-DB. Predictions were summed for 10 learning process of twenty-one subjects. Column and row labels represent the first letters of emotions respectively (anger, boredom, disgust, anxiety-fear, happiness, sadness and neutral). Prediction of disgust emotion got the highest score compared to others.

		Truth Label								
		A	**B**	**D**	**F**	**H**	**S**	**N**	**Classification Overall**	**Precision (%)**
Predicted Label	**A**	1131	626	409	410	378	403	368	3725	30.4
	B	550	590	487	435	521	387	405	3375	17.5
	D	407	475	597	513	488	501	517	3498	17.1
	F	385	476	554	740	540	501	465	3661	20.2
	H	342	499	504	525	608	463	484	3425	17.8
	S	356	442	512	470	495	655	640	3570	18.3
	N	399	462	507	477	540	660	691	3736	18.5
	Truth Overall	3570	3570	3570	3570	3570	3570	3570	24990	
	Recall (%)	31.7	16.5	16.7	20.7	17.0	18.3	19.4		20.1

Table 4.4: Prediction accuracies of all emotion and each subject in Response-DB. The accuracies of anger is obviously higher than others. A:Anger, B:Boredom, D:Disgust, F:AnxietyFear, H:Happiness, S:Sadness, N:Neutral, All: Mean value of accuracies for each subjects.

Sub No	**A**	**B**	**D**	**F**	**H**	**S**	**N**	**All**
1	25.87	15.29	5.31	18.81	18.23	18.22	10	15.96
2	49.39	21.75	16.47	9.99	19.4	21.16	17.65	22.25
3	42.36	10.59	21.16	55.89	24.7	21.18	13.53	27.06
4	12.35	15.89	15.28	7.66	10	11.75	14.71	12.51
5	67.06	35.89	24.69	20.56	18.23	18.8	22.92	29.76
6	82.36	12.36	10.58	17.05	27.06	14.71	15.32	25.65
7	20.58	24.11	36.46	22.35	29.39	17.63	19.4	24.27
8	49.99	18.8	9.42	19.41	15.3	24.71	24.7	23.18
9	15.88	17.05	14.12	18.21	16.47	10.01	17.63	15.63
10	24.69	7.65	31.17	10.6	12.93	18.23	17.06	17.49
11	22.92	25.27	9.42	12.94	9.99	10.01	21.17	15.96
12	20.01	14.72	14.69	16.48	22.34	16.45	12.93	16.8
13	50.58	15.87	12.95	11.77	12.95	24.69	57.05	26.55
14	48.82	25.86	11.76	9.42	15.88	19.41	16.46	21.1
15	32.94	12.35	14.12	19.41	14.12	18.82	22.35	19.17
16	20.58	12.93	13.53	16.46	17.06	9.43	17.04	15.29
17	20.59	8.83	8.25	24.09	40.59	11.18	41.19	22.1
18	16.46	23.51	19.98	6.48	9.42	22.34	18.81	16.71
19	6.49	20.56	22.93	11.78	7.06	8.83	8.84	12.34
20	16.46	15.88	20	14.71	13.53	9.43	11.77	14.54
21	22.92	11.19	7.65	12.94	9.41	17.64	22.35	14.86
Mean	**31.87**	**17.45**	**16.19**	**17.00**	**17.34**	**16.41**	**20.14**	**19.48**

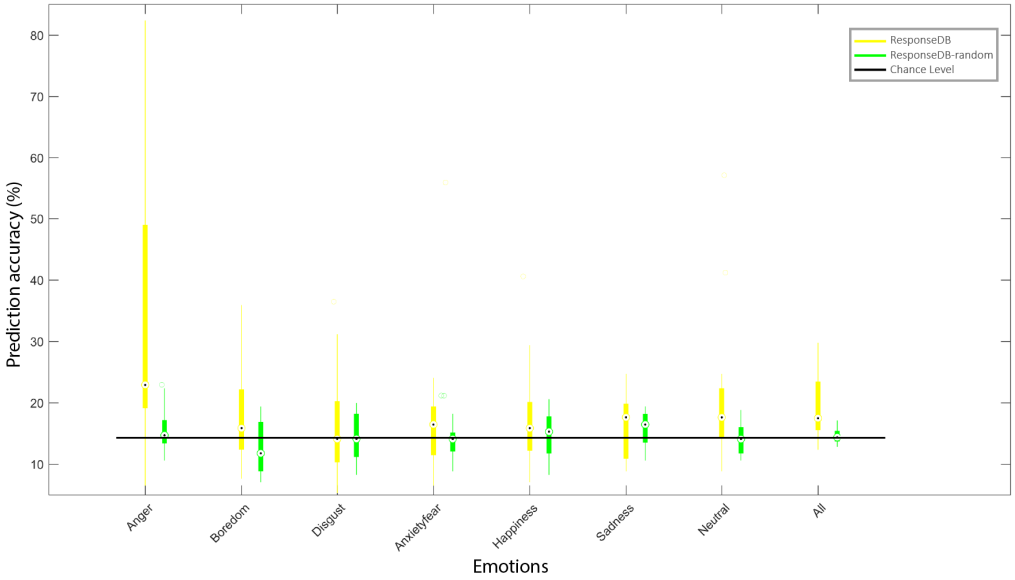


Figure 4.3: Test accuracy results for all emotions of Response-DB and *random* Response-DB. The yellow and green boxes indicates the accuracy and its distribution of Response-DB and *random* Response-DB, respectively. The chance level is represented by the black line. The accuracy for anger is the highest accuracy, while neutral is the second highest one. The accuracy distribution of the randomly generated Response-DB is equal to the chance level.

The accuracy of classification was calculated by looking at whether the estimated output and the expected output belong to the same class. For this reason, the expected outputs were randomly mixed to see whether the inputs were predicted by chance. As expected, the accuracies for emotions were about at the chance level. The accuracy results of *random* Response-DB are given in Table 4.5.

The accuracy distribution of *random* Response-DB is indicated by the green boxes while the actual Response-DB is given by the yellow boxes and chance level indicated by black line in Figure 4.3.

4.2.1 N-way Analysis of Variance for Response-DB

In the ANOVA table (4.4a), first three rows corresponded to the factors gender, Eysenck personality and emotion, respectively. The *p-value* 0.214 indicated that the mean accuracies for male and female were not significantly different. Similarly, the *p-value* 0.6668 indicated that the mean responses for introvert and extrovert, of the factor Eysenck personality test, were not significantly different. However, the *p-value* 0.0031 was small enough to conclude that the mean accuracies were significantly different for emotions.

From 4^{th} row to 6^{th} row represented the intersections of the factors; gender and Eysenck, gender and emotion, Eysenck and emotion. All three of them was greater than 0.05, so they failed to reject the null hypothesis. There was no significant effect

Table 4.5: Prediction accuracies for all emotions and each subject in *random* Response-DB. The accuracies for all emotions are at chance level. A:Anger, B:Boredom, D:Disgust, F:AnxietyFear, H:Happiness, S:Sadness, N:Neutral, All: Mean value of accuracies for each subjects.

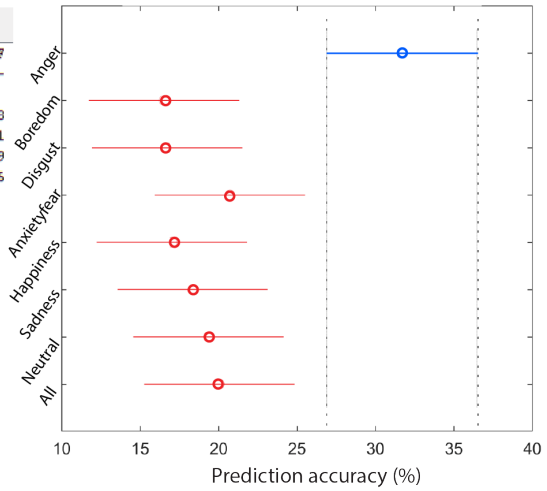
Sub No	**A**	**B**	**D**	**F**	**H**	**S**	**N**	**All**
1	15.87	8.82	8.25	10.58	12.95	18.8	18.24	13.35
2	17.05	18.82	17.63	14.12	18.23	12.36	11.18	15.62
3)	17.65	7.07	9.42	18.23	11.17	12.94	13.55	12.86
4	16.47	11.19	11.18	14.7	16.48	14.11	14.11	14.03
5	15.88	15.28	12.95	14.7	13.54	16.46	11.77	14.38
6	17.64	8.25	18.23	8.84	15.89	11.76	10.6	13
7	14.13	19.4	19.4	18.22	17.63	13.53	11.76	16.3
8	10.59	11.19	12.94	14.12	11.76	17.05	15.86	13.36
9	11.77	18.83	13.52	11.18	20.57	18.22	16.48	15.81
10	14.71	11.19	18.23	8.83	10	18.22	16.46	13.93
11	10.58	15.88	18.83	12.34	15.3	18.84	10.59	14.6
12	14.69	8.84	10.01	16.47	15.88	17.06	12.94	13.69
13	21.77	8.83	16.47	13.52	11.77	19.42	15.89	15.37
14	22.34	11.77	15.87	12.35	10.59	17.64	11.77	14.62
15	13.52	11.77	15.3	14.71	15.29	14.1	15.88	14.38
16	13.52	15.3	19.99	14.1	8.24	18.83	12.36	14.59
17	12.94	7.06	12.95	14.7	18.83	10.58	18.82	13.7
18	15.28	18.81	14.13	10.59	15.3	15.28	12.96	14.63
19	14.13	16.46	18.8	21.16	19.99	15.29	14.11	17.14
20	22.94	18.22	10.01	21.18	15.89	13.52	17.63	17.06
21	12.94	10.58	11.17	14.71	18.82	17.64	14.13	14.28
Mean	**15.54**	**13.03**	**14.54**	**14.25**	**14.96**	**15.79**	**14.15**	**14.60**

(a)

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Gender	163.3	1	163.283	1.56	0.214
Eysenck	19.5	1	19.505	0.19	0.6668
Emotion	2388.8	7	341.254	3.26	0.0031
Gender*Eysenck	326	1	325.972	3.11	0.0799
Gender*Emotion	238.6	7	34.093	0.33	0.9415
Eysenck*Emotion	746.9	7	106.695	1.02	0.421
Error	14984.9	143	104.789		
Total	20359	167			

(b)



(c)

Y \ X	A	B	D	F	H	S	N
A							
B	<						
D	<						
F	<						
H	<						
S	<						
N	<						

Figure 4.4: Two statistical analyses of the mean accuracies for Response-DB. (a) n-way ANOVA test of the mean accuracies (b) accuracy results of emotions. The blue bar shows the interval of the mean accuracies for anger emotion. The red bars are the intervals for other emotions and all are significantly from the interval of anger. (c) a color map of the emotions' interaction. The provided table represents information on the combination of 7 emotions; anger, boredom, disgust, anxiety-fear, happiness, sadness and neutral initialized by A, B, D, F, H, S, N, respectively. If the interval of accuracies for *X* is bigger than the interval of accuracies for *Y*, the cell is colored by light yellow. If not, the cell is colored by dark yellow.

on accuracies of classification due to these interactions.

Accuracies for emotions can be observed in (Figure 4.4b).The blue bar showed the interval for the mean accuracies for anger emotion. The red bars were the intervals for the mean accuracies for other emotions. None of the red bars overlap with the blue bar, which means the mean accuracies for anger was significantly different from the mean accuracies for other emotions (colored by red).

Besides, a different visualization of the combinations for all emotions is given in the Figure 4.4c. Yellow cells represented the difference in the mean accuracies of the emotions made in combination. The first column showed the mean accuracies for anger emotion was different and bigger than other emotions, that's why these intersecting cells were colored with light yellow. On the other hand, the intersection between boredom and anger emotion (at second column and first row) was colored with dark yellow, because mean accuracy for boredom was smaller than mean accuracy for anger. The mean accuracies for other emotions were not different than boredom, so the cells were left as white.

Table 4.6: Prediction accuracies for all emotions and each subject in Hybrid-DB. The accuracy for disgust is higher than others and the second highest accuracy is anger whereas the accuracy for anxiety-fear is the lowest one. A:Anger, B:Boredom, D:Disgust, F:AnxietyFear, H:Happiness, S:Sadness, N:Neutral, All: Mean value of accuracies for each subjects.

Sub No	A	B	D	F	H	S	N	All
1	57.06	38.81	60.59	49.42	44.7	49.99	52.94	50.49
2	65.29	44.12	58.83	36.48	44.13	62.94	59.99	53.12
3	68.24	46.47	68.25	65.88	54.68	68.81	57.06	61.34
4	44.13	37.65	57.64	32.34	38.82	40.59	45.29	42.36
5	83.52	67.66	67.65	48.25	57.06	55.29	62.36	63.11
6	88.24	42.92	62.37	48.24	50.58	64.13	54.12	58.63
7	71.19	59.99	73.54	59.43	56.47	68.24	54.11	63.27
8	64.11	53.53	60.58	45.89	48.23	52.36	64.13	55.54
9	52.94	48.83	56.48	42.35	51.76	37.65	44.7	47.8
10	53.52	37.66	64.71	34.13	47.65	54.12	61.77	50.49
11	57.64	53.52	55.89	38.22	49.41	57.65	57.65	52.85
12	51.17	44.7	57.64	33.53	50.01	60	47.65	49.26
13	64.11	63.53	72.37	42.94	58.82	65.3	73.54	62.92
14	51.76	45.28	71.19	42.35	45.29	55.89	54.09	52.27
15	54.72	47.65	60.02	42.95	52.94	58.82	50	52.44
16	46.47	38.24	64.71	34.11	48.23	58.24	52.93	48.98
17	54.69	47.06	59.42	44.72	55.27	53.54	67.06	54.54
18	45.89	58.23	64.71	37.65	44.71	54.69	60.59	52.35
19	48.83	38.24	59.41	42.94	44.71	47.07	49.4	47.23
20	60.6	42.95	58.81	38.83	51.16	51.17	42.97	49.5
21	50.59	40.02	59.42	38.82	42.96	58.23	55.88	49.4
Mean	**58.80**	**47.48**	**62.58**	**42.83**	**49.41**	**55.94**	**55.63**	**53.23**

4.3 Hybrid Data Results

The creation of the Hybrid-DB dataset had been described in Section 3.6.

For each subject, classification accuracy rates, are given in Table 4.6 and range from 32,34% to 88.24%. The recognition rates of anger, disgust and neutral emotion were higher than other emotions in all accuracy rates of each subject. When the averages of the participants' results were taken, this interval changed from 42.83% to 62.58%. According to this information, disgust had the highest accuracy, while the second highest accuracy value belonged to the anger emotion. On the other hand, anxiety-fear had the lowest accuracy rate. Overall accuracy rate was 53.23%.

When combined with the subject response features, another test was conducted to check whether response data actually contributed a new quality and helped for a better classification or not. To test the genuineness of actual responses, another *random* Hybrid-DB was created with a random emotion responses and compared with original Hybrid-DB. Section 3.6 describes how the *random* Hybrid-DB was obtained.

The result was indicated by the red boxes in Figure 4.5 and is given in Table 4.7.

Table 4.7: Prediction accuracies for all emotions and each subject in *random* Hybrid-DB. The accuracy for disgust is the highest accuracy while prediction for anxiety-fear has the lowest accuracy rate. A:Anger, B:Boredom, D:Disgust, F:AnxietyFear, H:Happiness, S:Sadness, N:Neutral, All: Mean value of accuracies for each subjects.

Sub No	A	B	D	F	H	S	N	All
1	50	43.53	58.83	37.06	52.94	55.9	54.11	50.33
2	38.83	43.52	56.47	37.66	44.69	45.89	43.54	44.37
3	48.24	40	62.36	29.41	38.24	57.65	47.06	46.14
4	42.35	37.05	60	42.35	42.35	57.06	48.22	47.06
5	18.22	35.29	54.69	30.59	45.89	48.23	43.54	39.5
6	27.04	44.71	52.35	48.24	50.01	50.01	46.47	45.55
7	51.76	38.23	54.11	34.13	41.78	45.88	48.23	44.9
8	38.81	45.31	61.75	36.47	48.22	42.95	41.78	45.03
9	48.81	42.35	53.53	37.65	47.63	52.34	44.12	46.63
10	41.18	37.65	55.89	45.3	44.7	47.04	52.95	46.37
11	51.75	48.82	62.94	41.76	42.95	54.11	54.11	50.91
12	50	44.11	58.23	40.58	45.31	55.28	43.53	48.14
13	41.18	48.81	67.05	41.77	55.88	52.38	45.29	50.31
14	34.7	39.4	62.95	36.48	50.59	48.83	57.07	47.13
15	34.71	47.07	62.34	37.65	48.22	48.81	50.59	47.07
16	43.52	40.59	54.71	31.75	38.83	54.71	45.29	44.2
17	45.29	44.71	54.12	41.18	48.24	50.58	42.96	46.72
18	52.36	35.29	57.07	28.82	51.77	45.87	45.89	45.29
19	50	48.84	61.78	35.89	46.47	44.1	50.58	48.22
20	48.24	44.7	57.06	39.41	44.13	57.06	49.41	48.57
21	44.71	33.54	57.66	37.05	47.65	54.7	44.12	45.63
Mean	**42.94**	**42.07**	**58.38**	**37.68**	**46.50**	**50.92**	**47.56**	**46.57**

As it can be seen in Figure 4.5, all accuracies decreased for all emotions when Hybrid-DB was merged randomly. In particular, this decline was conspicuous for the emotions of anger and neutral.

4.3.1 N-way Analysis of Variance for Hybrid-DB

In the ANOVA table (Figure 4.5a), the *p-value* 0.1339 and the *p-value* 0.2901 for gender and Eysenck personality indicated that the mean accuracies were not significantly different. On the other hand, the *p-value zero* was smaller than (< 0.05) so there was strong evidence that the mean accuracies for emotions were different.

When emotions were examined within themselves, the interval of anger emotion was indicated by the blue bar (Figure 4.5b). The mean accuracy interval of boredom and anxiety-fear emotions were different than the mean accuracy for anger emotion, so they were shown with red bars. However, there were no significant differences between the mean accuracy for anger emotion and the mean accuracies for disgust, happiness and neutral emotions.

Looking at the color map (Figure 4.5c), a table of all combinations for emotions

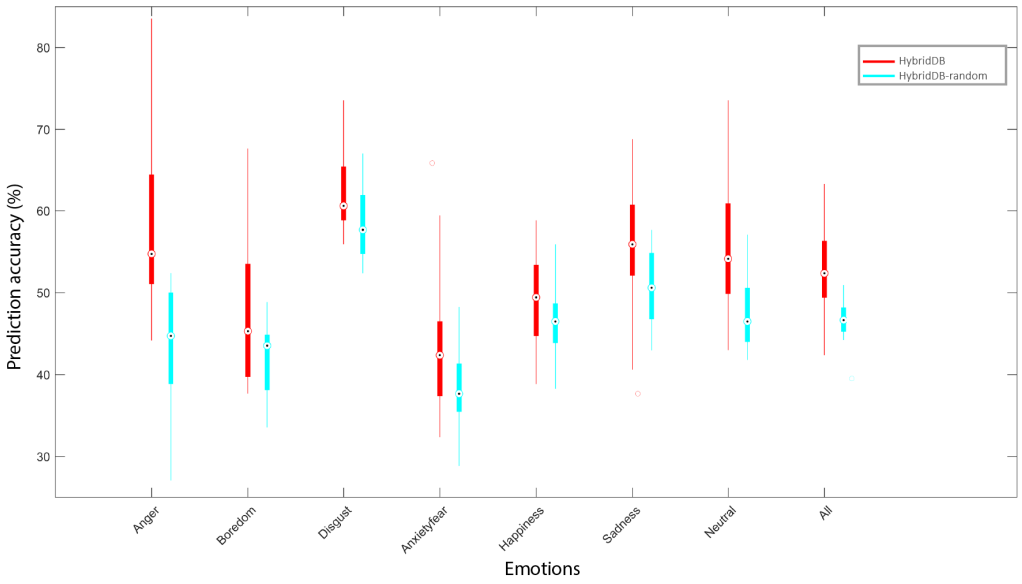


Figure 4.5: Test accuracy results for all emotions of Hybrid-DB and *random* Hybrid-DB. The red and light blue boxes represents the distribution of accuracy rate of Hybrid-DB and *random* Hybrid-DB, respectively. It is seen that the accuracy values of the randomized data decrease for all emotions.

represented in the Figure 4.5b is shown. If there was a difference between the mean accuracy intervals of X and Y of emotion and Y was smaller than X, the intersecting cell was colored with light orange, but Y was greater than X, the intersecting cell was colored with dark orange. When the 5^{th} column was examined, the mean accuracy for happiness emotion was significantly different than the mean accuracies for disgust and anxiety-fear emotions, but not different from others. The mean accuracy for happiness emotion was greater than the mean accuracy for anxiety-fear but smaller than the mean accuracy for disgust emotion

Finally, the intersections of the factors were examined. Gender & emotion and Eysenck & emotion interaction was greater than 0.05, so they failed to reject null hypothesizes. Neither of them had significant effect on the accuracies of classification. However, gender & Eysenck was less than the significance level, it could reject the null hypothesis (Figure 4.7a). The interaction of gender & Eysenck had a statically significant effect on accuracies. Mean and standard deviation table of accuracies for gender and Eysenck personality is given in Figure 4.7a-b. There was only a difference between the mean accuracies for male and introvert subjects and the mean accuracies for female and introvert subjects (Figure 4.7c). When compared to the other interactions, there was no significant difference.

4.4 Overall Analysis of Results

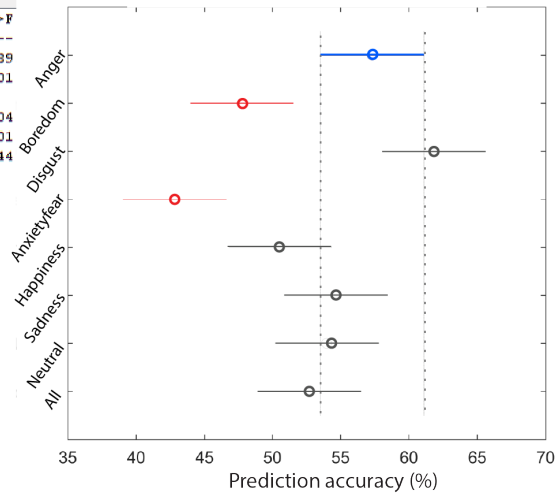
The recognition rate results of Emo-DB, Response-DB, Hybrid-DB, *random* Response-DB, *random* Hybrid-DB were compared with each other (Figure 4.8). These five dif-

(a)

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Gender	139.1	1	139.082	2.27	0.1339
Eysenck	69	1	69.034	1.13	0.2901
Emotion	3665.4	7	523.624	8.55	0
Gender*Eysenck	806.5	1	806.488	13.17	0.0004
Gender*Emotion	192	7	27.433	0.45	0.8701
Eysenck*Emotion	382.3	7	54.62	0.89	0.5144
Error	8754.5	143	61.22		
Total	15426.3	167			

(b)



(c)

Y \ X	A	B	D	F	H	S	N
A							
B	<						
D		>					
F	<		<				
H			<	>			
S				>			
N			<	>			

Figure 4.6: Two statistical analyses of the mean accuracies for Hybrid-DB. (a) n-way ANOVA test of the mean accuracies (b) accuracy results of emotions. The blue bar indicates the interval of the mean accuracies for anger emotion. If there is no intersection between the interval of the mean accuracies for anger emotion and the interval of the mean accuracies for other emotions, the bars are shown by a red color. Otherwise, the bars are shown by a gray color. (c) a color map of the emotions' intersection. The table gives data on the intersection from 7 different emotions (anger, A; boredom, B; disgust, D; anxiety-fear, F; happiness, H; sadness, S; neutral, N). When the interval of accuracies for X is bigger than the interval of accuracies for Y, the cell is colored by light orange. In the other direction of the equation, the cell is colored by dark orange.

ferent test sets were compared and the average results of datasets were given in Table 4.8. In Emo-DB, disgust and neutral emotion had higher recognition rate, while boredom had lowest. From these three datasets, the Response-DB set had the lowest classification rate. The two highest rates were anger and neutral emotion where disgust was lowest in Response-DB. Anger and disgust emotion had the highest recognition rate in merged dataset

In addition, ANOVA test was performed depending on the emotions for accuracy comparison of 5 datasets (Response-DB, Hybrid-DB, Emo-DB, *random* Response-DB, *random* Hybrid-DB). As a result of this test, Response-DB had the lowest accuracy rate from the set of actual data as seen in Figure 4.9, Emo-DB had the highest accuracy rate. *Random* Response-DB was located at the chance level, and there was a decrease in rate for Hybrid-DB when the features were merged randomly.

The rank of accuracies had already been described in this chapter with figures and tables. On the other hand, the main consideration in Figure 4.9 was the intersection for Hybrid-DB (red) and Emo-DB (dark blue). In Emo-DB, the accuracy for anger emotion was dramatically less than the accuracy for disgust emotion. However, for Hybrid-DB, the accuracy rate for anger was close to the accuracy rate of disgust.

(a)

	Mean Table of Accuracy				Std Table of Accuracy		
	Int	Ext	Avg		Int	Ext	Avg
Male	57,00	51,68	54,34		9,80	8,17	8,99
Female	50,25	53,11	51,68		7,70	10,39	9,04
Avg	53,63	52,39	53,01		8,75	9,28	9,02

(b)

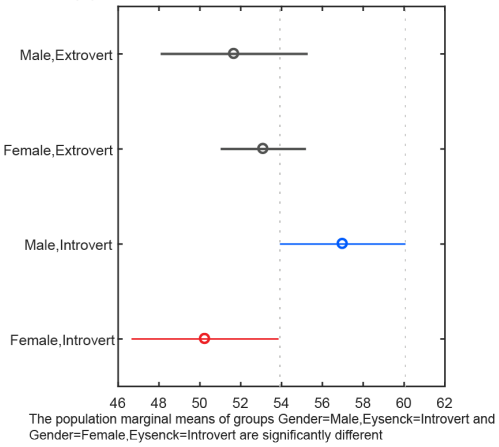


Figure 4.7: Statistical analysis of the mean accuracies which is resultant of the interaction between gender and Eysenck personality factors. (a) mean table and standard deviation table of accuracies. (b) A multi comparison result of the interaction of gender and personality factors.

Table 4.8: Table visualization of the accuracy rates for all datasets. Row titles indicates as datasets (Response-DB, Response-DB) whereas column titles represent emotions; A:Anger, B:Boredom, D:Disgust, F:AnxietyFear, H:Happiness, S:Sadness, N:Neutral, All: Mean value of accuracies for each datasets.

	A	**B**	**D**	**F**	**H**	**S**	**N**	**All**
Emo-DB	65.88	55.87	77.06	59	69	70	76	67.55
Response-DB	31.87	17.45	16.19	17.00	17.34	16.41	20.14	19.48
random Response-DB	15.54	13.03	14.54	14.25	14.96	15.79	14.15	14.60
HybridDB	58.80	47.48	62.58	42.83	49.41	55.94	55.63	53.23
random Hybrid-DB	42.94	42.07	58.38	37.68	46.50	50.92	47.56	46.57

Besides, the accuracy for anger was higher than the accuracies for sadness and neutral, in contrast to Emo-DB.

Analysis of accuracies (Figure 4.4 & 4.6) showed that gender and personality factors were not related to emotion factor, general view of the interaction is represented in Figure 4.10 and Figure 4.11.

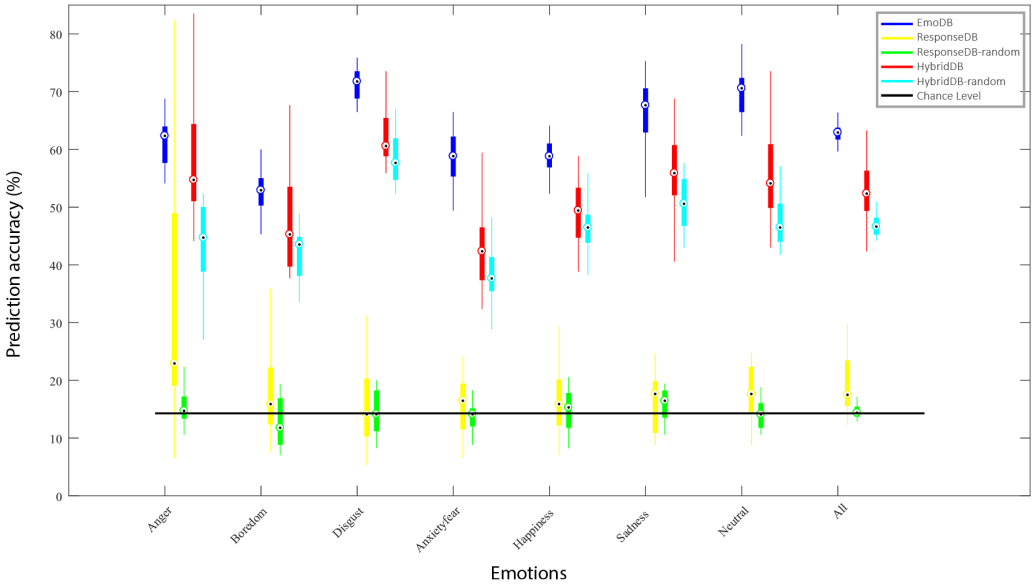


Figure 4.8: Graphical visualization of test accuracy values for all datasets (Emo-DB, dark blue; Response-DB, yellow; Hybrid-DB, red; *random* Response-DB,green ;*random* Hybrid-DB, light blue). Y-axis indicates prediction rate of datasets whereas clustered bars are represented by each emotion in X-axis.

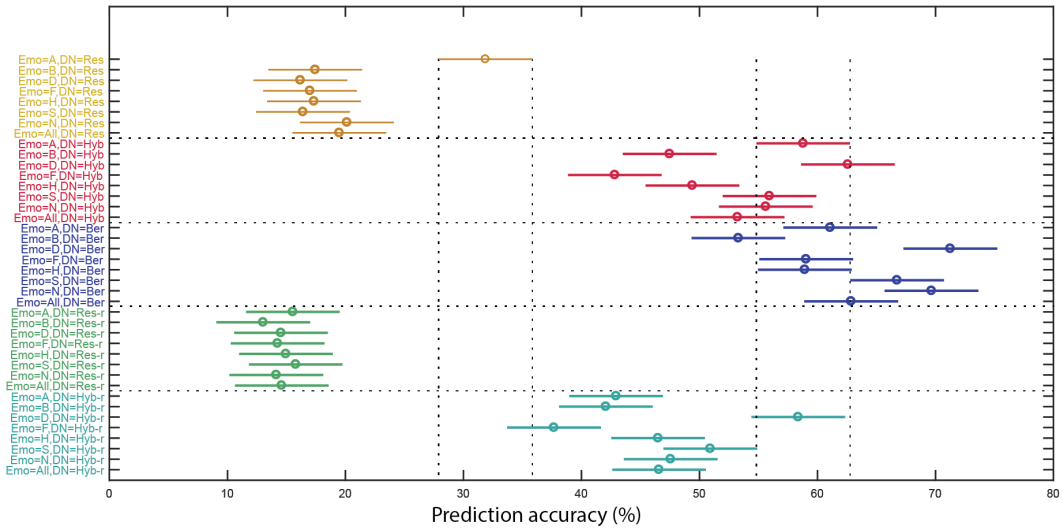


Figure 4.9: A two-way ANOVA test of all datasets (Response-DB, Hybrid-DB, Emo-DB, *random* Response-DB, *random* Hybrid-DB), are represented by yellow, red, dark blue, green, light blue bars, respectively.Each clustered column represents 7 emotions; anger, boredom, disgust, anxiety-fear, happiness, sadness and neutral from the top on down.

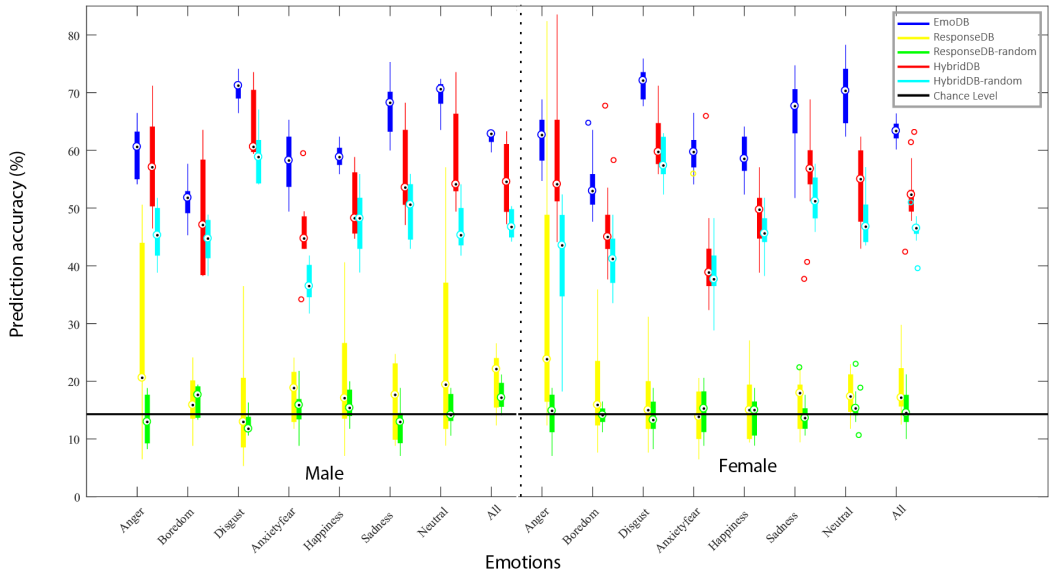


Figure 4.10: Test accuracies of all datasets according to gender.In the figure, is divided into equal parts with each part having accuracy results for gender categories (Male / Female). Datasets: Response-DB, Hybrid-DB, Emo-DB, *random* Response-DB, *random* Hybrid-DB. The black line shows the chance level of classification prediction.

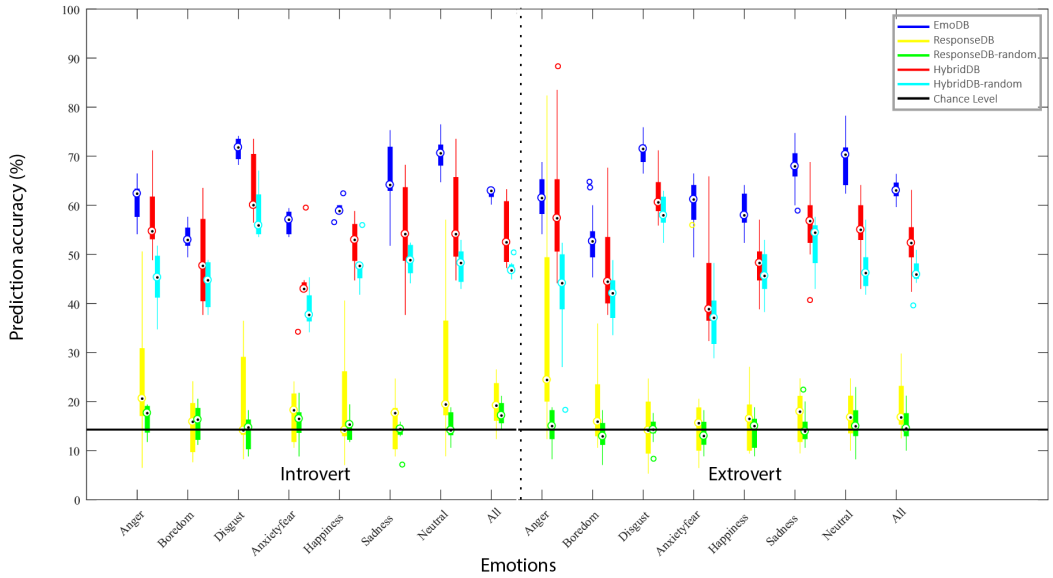


Figure 4.11: Test accuracies of all datasets according to Eysenck personality (Introvert / Extrovert). Datasets: Response-DB, Hybrid-DB, Emo-DB, *random* Response-DB, *random* Hybrid-DB. The chance level of recognition rate is indicated by black line

CHAPTER 5

CONCLUSION

5.1 Discussion

In our thesis, two hypotheses were proposed. First of them was that participants who hear an emotional voice, reacts to these recordings in an emotional way. In addition, it was possible to classify these reactions given based on their own speech responses. Therewithal, how well the classification accuracies for emotions were distinguished from each other was also the field of our thesis. Results of analysis was verified for both statements. The other hypothesis tries to prove that the information obtained from the emotional response may positively contribute to the classification of the emotion of the recording. The contribution to the classification had resulted accuracies of Emo-DB to fall contrary to expectations. Cause of this fact is explained in detail in this section.

The results of the analysis of three datasets are interpreted in order. Our results showed that the highest classification accuracy was for anger, whereas the lowest accuracy rate was disgust for Response-DB (Table 4.4). A study by conducted by Yuncu et al. (2014) used the Berlin database, showed that participants, who heard to voice recordings, classified anger emotion better. It may be said that, emotion classification even differ by also human ear, so emotion perceptibly differs when compared to other emotions.

In some studies, it has been found that the perception of anger is more successful than the perception of other emotions. Hansen and Hansen (1988) found that in emotional perception from face, it was easier to select anger emotion among the crowd. Pell et al. (2009) researched on recognition of the emotions in three foreign languages (English, German, Arabic) and native language (Spanish) with the help of human listeners. They found that the most perceived emotion was different on each foreign languages. The best perceptions for emotions in German language were anger and neutral. In Pell and Kotz (2011) work shows that the time of perception also differs for each emotion. We may say that time of perception can affect the expression of emotions in vocal responses.

In order to examine whether the sound recordings in Response-DB correspond to a certain emotion, the target output was mixed randomly. Moreover, the accuracy rates for all emotions were found to be at the chance level in this randomly mixed Response-DB (Table 4.5). This suggests that the actual Response-DB has a knowl-

edge of emotional response and confirms the first hypothesis we are advocating.

On the other hand, Table 4.8 showed that Response-DB had the lower classification accuracy when compared to Emo-DB. The reasons for this may be that the sound recordings in the dataset have noise and that the participants are not professional actors. It had been proven in many studies that the database of professional actors like Berlin Emo-DB was getting much better results (El Ayadi et al. 2011). In our study, the results were obtained in the same direction as this information (Table 4.8).

The interpretation of Hybrid-DB showed that accuracies of this dataset are also less than Emo-DB. Adding new features to the feature vector causes an increase for accuracy rate unless they are extracted from a noisy data. The features of Emo-DB added to the Hybrid-DB had a great contribution to the emotion classification. It is possible to understand this contribution from the similar fluctuation in the accuracy curve of Emo-DB and Hybrid-DB (Figure 4.8).

The accuracy rates obtained for each emotion except anger show the same skewed distribution. Anger was different and had increased accuracy rate. The features for anger in Response-DB had the highest accuracy rate enhancement, which is reflected in the accuracy of Hybrid-DB. We couldn't prove that our second hypothesis that the information obtained from the response dataset is a positive contribution to the classification of an emotion of the voice heard. However, it may be positive contribution after overcome the limitations.

When we examine the Emo-DB, it can be seen that two emotions predicted with the highest accuracies are disgust and neutral, respectively. However, in the Hybrid-DB obtained by adding new features, it can be seen that these rankings change for disgust and anger (Table 4.8). This demonstrates the importance of the response dataset in terms of accuracy rate.

The Eysenck personality test and gender interaction were examined to find whether there was any effect on the emotion classification (Figure 4.7). It seemed that the accuracy for introvert males was significantly better than the accuracy of introvert females. Probably, the unbalanced distribution of the subjects caused this effect. It requires further examination with more subjects and more advanced statistical analysis. Because there were few and unbalanced subjects in our dataset, even a single subject could affect the outcome in an undesirable way.

5.2 Limitations

In this thesis, there are a few limitations on the accuracy improvement. One of them is the noise of the raw data. Noisy data cause a decrease in both feature extraction and classification. Another reason, why obtained accuracy rates for Response-DB dataset were limited, was that the participants were not professional actors. Since a simple approach was used in the merging process of two datasets (Emo-DB and Response-DB), accuracy rates for Hybrid-DB is lower than that for Emo-DB. A result can be enhanced by applying a better hierarchical fusion methods. A small size of dataset also creates a limitation in the application of learning methods. It is possible that more efficient and precise results can be achieved by extending dataset size. Enriching

the dataset in all aspect will be useful for obtaining a good classification accuracy. Gender and personality data can be made balanced and more subject can be provided. Furthermore, feature types can be selected in this large feature vector, increasing the number and type of features extracted from the data and improving accuracy. Trying to classify seven emotions is another influence that limits the accuracy. The more labels (emotions) to classify, the more difficult to handle the accuracy rate.

5.3 Future Work

To sum it up, this study suggested two hypotheses in the fields of the thesis's purpose and these first hypothesis was confirmed. It was shown that people who listen to emotional voice records react also emotionally. Especially for anger, the accuracy of the classification was verified. The hybridization approach applied to datasets caused the accuracy of Emo-DB decrease. After reducing the effect of the limitation, mentioned in the previous section, an increase in the accuracy can be achieved by using different classification methods.

In future work, response audios to an emotional speech will be recorded in a noiseless environment. The length of datasets will be increased by working with more subjects. In addition, new feature extraction methods will be added to increase the size of feature vector, such as speed of speech and articulations (Khanna and Sasikumar 2011). The features that trigger a specific emotion can be analyzed.

Thus, we might try to improve accuracy by applying advanced classification methods such as DNN which shows better accuracy with a wider dataset. According to Fayek et al. (2017) review, Convolutional Neural Networks (ConvNets) performed better prediction accuracy compared to other architectures. In future work, we will enhance the accuracy of Response-DB by using ConvNets.

According to some theories, emotions are assumed to be influenced by cultures.How the person feels, expresses and perceives the emotions depends on the cultural structure of the community surrounding the person (Richerson and Boyd 2005). Since culture and personality of person play a role in the expression of emotions, it may have been a cultural influence in the correct classification of anger emotion (Markus and Kitayama 1991; Gunkel et al. 2014). Other emotions may be easier to express for other societies, so it may lead an emotion to have higher accuracy. This can be considered as an another future study.

Based on Pell et al. (2009) and Pell and Kotz (2011) researches, we may make subjects listen to the voice recordings in other languages and we may investigate whether we can also obtain any useful information in those languages. In our work, we researched the acoustic information of the sound recordings, more than the semantic integrity of them. For this reason, voice recordings containing German language were selected. However, Turkish voice recordings will be played to the subjects in order to create a control group for purposes of the comparison of new foreign language.

REFERENCES

Afshar, H. S. & Rahimi, M. (2014). The Relationship among Critical Thinking, Emotional Intelligence, and Speaking Abilities of Iranian EFL Learners. *Procedia - Social and Behavioral Sciences*, *136*, 75–79. doi:10.1016/j.sbspro.2014.05.291

Alloway, T. P., Copello, E., Loesch, M., Soares, C., Watkins, J., Miller, D., ... Ray, S. (2016). Investigating the reliability and validity of the Multidimensional Emotional Empathy Scale. *Measurement*, *90*, 438–442. doi:10.1016/j.measurement.2016.05.014

Atal, B. S. & Hanauer, S. L. (1971). Speech Analysis and Synthesis by Linear Prediction of the Speech Wave. *Journal of the Acoustical Society of America*, *50*, 637–655. doi:10.1121/1.1912679

Banse, R. & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*. doi:10.1037/0022-3514.70.3.614. arXiv: 0022-3514.70.3.614 [10.1037]

Bezooijen, R. (1984). Characteristics and Recognizability of Vocal Expressions of Emotion. Berlin, Boston: De Gruyter Mouton. doi:10.1515/9783110850390

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436. doi:10.1163/156856897X00357. arXiv: 1011.1669v3

Breazeal, C. & Aryananda, L. (2002). Recognition of affective communicative intent in robot-directed speech. *Autonomous Robots*, *12*(1), 83–104. doi:10.1023/A:1013215010749

Bridle, J. S. & Brown, M. D. (1974). *An experimental automatic word recognition system*. Joint Speech Research Unit, Ruislip, England.

Brown, K. (2006). *Encyclopedia of Language & Linguistics*. doi:10.1016/B0-08-044854-2/09040-4

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A database of german emotional speech. In *In proceedings of interspeech, lissabon* (pp. 1517–1520).

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, *42*(4), 335–359. doi:10.1007/s10579-008-9076-6

Cabanac, M. (1992). Pleasure: the common currency. *Journal of Theoretical Biology*, *155*(2), 173–200. doi:10.1016/S0022-5193(05)80594-6

Cairns, D. A. & Hansen, J. H. L. (1994). Nonlinear analysis and classification of speech under stressed conditions. *Journal of the Acoustical Society of America*, *96*(6), 3392–3400. doi:10.1121/1.410601

Catanzaro, B., Sundaram, N., & Keutzer, K. (2008). *Fast Support Vector Machine Training and Classication on Graphics Processors*. doi:10 . 1145 / 1390156 . 1390170

Chernykh, V., Sterling, G., & Prihodko, P. (2017). Emotion recognition from speech with recurrent neural networks. *CoRR*, *abs/1701.08071*.

Coffey, S., Vanderlip, E., & Sarvet, B. (2017). The Use of Health Information Technology Within Collaborative and Integrated Models of Child Psychiatry Practice. *Child and Adolescent Psychiatric Clinics of North America*, *26*(1), 105–115. doi:10.1016/j.chc.2016.07.012

Cybenko, G. (1992). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *5*(4), 455–455. doi:10.1007/BF02134016

Davis, S. B. & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics,Speech, and Signal Processing*, *28*(4). doi:10.1109

Demircan, S. & Kahramanlı, H. (2014). Feature Extraction from Speech Data for Emotion Recognition. *Journal of Advances in Computer Networks*, *2*(1), 28–30. doi:10.7763/JACN.2014.V2.76

Douglas-Cowie, E., Cowie, R., Cox, C., Amier, N., & Heylen, D. K. J. (2008). The Sensitive Artificial Listner: an induction technique for generating emotionally coloured conversation. In L. Devillers, J.-C. Martin, R. Cowie, E. Douglas-Cowie, & A. Batliner (Eds.), *Lrec workshop on corpora for research on emotion and affect* (WP 08-02, pp. 1–4). ELRA.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases. *Pattern Recogn. 44*(3), 572–587. doi:10.1016/j.patcog.2010.09.020

Eysenck, H. J. & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire*. doi:10.1177/014662168000400106

Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, *92*(Supplement C), 60–68. Advances in Cognitive Engineering Using Neural Networks. doi:https://doi.org/10.1016/j.neunet.2017.02.013

Forbes-Riley, K. & Litman, D. J. (2004). Predicting emotion in spoken dialogue from multiple knowledge sources. *Proceedings of the 4th Meeting of the North American Chapter of the Association for Computational Linguistics: : Human Language Technologies,* 201–208.

France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., & Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk.

IEEE Transactions on Biomedical Engineering, *47*(7), 829–837. doi:10.1109/10.846676

Gelfer, M. P. & Mikos, V. A. (2005). The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels. *Journal of Voice*, *19*(4), 544–554. doi:10.1016/j.jvoice.2004.10.006

Gökçay, D. & Smith, M. (2012). TÜDADEN: Türkçe'de Duygusal ve Anlamsal Değerlendirmeli Norm Veri Tabanı. *Bilgisayar ve Beyin, Pan Yayıncılık,Bingöl H, editor.* 2012.

Gunkel, M., Schlägel, C., & Engle, R. L. (2014). Culture's Influence on Emotional Intelligence: An Empirical Study of Nine Countries. *Journal of International Management*, *20*(2), 256–274. doi:10.1016/j.intman.2013.10.002

Hansen, C. H. & Hansen, R. D. (1988). Finding the face in the crowd an anger superiority effect. *Journal of personality and social psychology*. doi:10.1037/0022-3514.54.6.917

Hansen, J. H. & Cairns, D. A. (1995). ICARUS: Source generator based real-time recognition of speech in noisy stressful and Lombard effect environments {star, open}. *Speech Communication*, *16*(4), 391–422. doi:10.1016/0167-6393(95)00007-B

Hayley, A. C., de Ridder, B., Stough, C., Ford, T. C., & Downey, L. A. (2017). Emotional intelligence and risky driving behaviour in adults. *Transportation Research Part F: Traffic Psychology and Behaviour*, *49*, 124–131. doi:10.1016/j.trf.2017.06.009

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, *87*(4), 1738–52. doi:10.1121/1.399423

Honda, M. (2003). Human Speech Production Mechanisms. *NTT Technical Review*, *1*(2), 24–29.

Iida, A., Campbell, N., Higuchi, F., & Yasumura, M. (2003). A corpus-based speech synthesis system with emotion. *Speech Communication*, *40*(1-2), 161–187. doi:10.1016/S0167-6393(02)00081-X

Ingale, A. & Chaudhari, D. (2012). Speech Emotion Recognition. *International Journal of Soft Computing and Engineering*, *2*(1), 235–238.

Jasmine, J. M., Sandhya, S., Ravichandran, D. K., & Balasubramaniam, D. D. (2016). Silence Removal from Audio Signal Using Framing and Windowing Method and Analyze Various Parameter. *International Journal of Innovative Research in Computer and Communication Engineering*, *4*(3), 3103–3108. doi:10.15680/IJIRCCE.2016.0404046. arXiv: 1105.3232v1

Karancı, N., Dirik, G., & Yorulmaz, O. (2007). Eysenck Kişilik Anketi - Gözden Geçirilmiş Kısaltılmış Formunun (EKA-GGK) Türkiye'de geçerlik ve güvenilirlik çalışması.

Khanna, P. & Sasikumar, M. (2011). Recognizing emotions from human speech. In S. J. Pise (Ed.), *Thinkquest2010: Proceedings of the first international conference on contours of computing technology* (pp. 219–223). New Delhi: Springer India. doi:10.1007/978-81-8489-989-4_40

Lee, C. M., Yildirim, S., Bulut, M., Kazemzadeh, a., Busso, C., Deng, Z., ... Narayanan, S. (2004). Emotion Recognition based on Phoneme Classes. *Database*, (1), 889–892.

Lee, C. M. & Narayanan, S. S. (2005). Toward Detecting Emotions in Spoken Dialogs. *13*(2), 293–303.

Liberman, M., Davis, K., Grossman, M., Martey, N., & Bell, J. (2002). Emotional Prosody Speech and Transcripts.

Litman, D. J. & Silliman, S. (2004). ITSPOKE: An Intelligent Tutoring Spoken Dialogue System. *Proc. of the Human Language Technology Conference: 4th Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, 233–236.

Litman, D., Forbes, K., & Silliman, S. (2003). Towards emotion prediction in spoken tutoring dialogues. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology companion volume of the Proceedings of HLT-NAACL 2003–short papers - NAACL '03*, *2*, 52–54. doi:10.3115/1073483.1073501

Lubis, N., Sakti, S., Neubig, G., & Toda, T. (2014). Emotion and Its Triggers in Human Spoken Dialogue : Recognition and Analysis. *Situated dialog in speech-based human-computer interaction*, 224–229.

Lugović, S., Dunđer, I., & Horvat, M. (2016). Techniques and applications of emotion recognition in speech. In *2016 39th international convention on information and communication technology, electronics and microelectronics (mipro)* (pp. 1278–1283).

Makhoul, J. (1973). Spectral Analysis of Speech by Linear Prediction. *IEEE Transactions on Audio and Electroacoustics*, *21*(3), 140–148. doi:10.1109/TAU.1973.1162470

Markus, H. & Kitayama, S. (1991). Culture and the self: Implications for cognition,emotion, and motivation. *Psychological Review*. doi:10.1037/0033-295X.98.2.224

McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M., & Stroeve, S. (2000). Approaching automatic recognition of emotion from Voice: A rough benchmark. *Proceedings of the ISCA Workshop on Speech and Emotion*.

Mermelstein, P. (1976). Distance Measures for Speech Recognition–Psychological and Instrumental. In *Joint workshop on pattern recognition and artificial intelligence*.

Moattar, M. & Homayounpour, M. (2012). A review on speaker diarization systems and approaches. *Speech Communication*, *54*(10), 1065–1103. doi:http://dx.doi.org/10.1016/j.specom.2012.05.002

Morales-Perez, M., Echeverry-Correa, J., Orozco-Gutierrez, A., & Castellanos-Dominguez, G. (2008). Feature extraction of speech signals in emotion identification. *Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society.* 2590–2593. doi:10.1109/IEMBS.2008.4649730

Murray, I. R. & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*. doi:10.1121/1.405558

Nicholson, J., Takahashi, K., & Nakatsu, R. (2000). Emotion Recognition in Speech Using Neural Networks. *Neural Computing & Applications*, *9*, 290–296. doi:10.1007/s005210070006

Nielsen, M. (2015). Neural networks and deep learning.

Nogueiras, A., Moreno, A., Bonafonte, A., & B. Mariño, J. (2001). *Speech emotion recognition using hidden Markov models.*

Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication*, *41*(4), 603–623. doi:10.1016/S0167-6393(03)00099-2

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). The measurement of meaning. *University of Illinois Press*, 1–352.

Pan, Y., Shen, P., & Shen, L. (2012). Feature Extraction and selection in speech emotion recognition. *Proceeding of the onlinepresent. org*, *2*(1050), 64–69.

Pell, M. D. & Kotz, S. A. (2011). On the time course of vocal emotion recognition. *PLoS ONE*, *6*(11). doi:10.1371/journal.pone.0027256

Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. A. (2009). Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior*, *33*(2), 107–120. doi:10.1007/s10919-008-0065-7

Petrushin, V. (1999). Emotion in speech: Recognition and application to call centers. *Proceedings of Artificial Neural Networks in Engineering*.

Petrushin, V. A. (2000). Emotion recognition in speech signal: Experimental study, development, and application. *Proceedings of the Sixth International Conference on Spoken Language Processing*, (Icslp), 5.

Pittermann, J., Pittermann, A., & Minker, W. (2010). Emotion recognition and adaptation in spoken dialogue systems. *International Journal of Speech Technology*, *13*(1), 49–60. doi:10.1007/s10772-010-9068-y

Rao, K. S. & Koolagudi, S. G. (2013). Robust emotion recognition using pitch synchronous and sub-syllabic spectral features. In *Robust emotion recognition us-*

ing spectral and prosodic features (pp. 17–46). New York, NY: Springer New York. doi:10.1007/978-1-4614-6360-3_2

Rao, K. S., Koolagudi, S. G., & Vempada, R. R. (2013). Emotion recognition from speech using global and local prosodic features. *International Journal of Speech Technology*, *16*(2), 143–160. doi:10.1007/s10772-012-9172-2

Richerson, P. J. & Boyd, R. (2005). *Not by Genes Alone: How Culture Transformed Human Evolution*.

Rigoll, G., Müller, R., & Schuller, B. (2005). Speech Emotion Recognition Exploiting Acoustic and Linguistic Information Sources. *Specom*, 61–67.

Robinson, P., Baltruaitis, T., Davies, I., & Pfister, T. (2011). The emotional computer. *Papers.Laurelriek.Org*.

Rosenblatt, F. (1960). Perceptron Simulation Experiments. *Proceedings of the IRE*, *48*(3), 301–309. doi:10.1109/JRPROC.1960.287598

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learining Internal Representations by Error Propagation. doi:10.1016/B978-1-4832-1446-7.50035-2. arXiv: 1011.1669v3

Russell, J. A. (1980). A Circumplex Model of Affect.

Sánchez-Gutiérrez, M., Albornoz, E., Martiney-Licona, F., Rufiner, H., & Goddard, J. (2014). Deep Learning for Emotional Speech Recognition. *6th Mexican Conference on Pattern Recognition (MCPR2014)*, 311–320. doi:10.1007/978-3-319-07491-7_32

Schafer, E. C., Wright, S., Anderson, C., Jones, J., Pitts, K., Bryant, D., ... Reed, M. P. (2016). Assistive technology evaluations: Remote-microphone technology for children with Autism Spectrum Disorder. *Journal of Communication Disorders*, *64*, 1–17. doi:10.1016/j.jcomdis.2016.08.003

Schuller, B. (2002). Towards intuitive speech interaction by the integration of emotional aspects. In *Ieee international conference on systems, man and cybernetics*. doi:10.1109/ICSMC.2002.1175635

Schuller, B., Steidl, S., & Batliner, A. (2009). A.: The interspeech 2009 emotion challenge. In *In isca, ed.: Proceedings of interspeech* (pp. 312–315).

Slaney, M. & McRoberts, G. (2003). BabyEars: A recognition system for affective vocalizations. *Speech Communication*, *39*(3-4), 367–384. doi:10.1016/S0167-6393(02)00049-3

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A Scale for Measurement of the Psychological Magnitude Pitch. *Journal of the Acoustical Society of America*, *8*(jan 1937), 185–190. doi:10.1121/1.1915893

Stuckless, R. (1994). Developments in real-time speech-to-text communication for people with impaired hearing. *Communication access for people with hearing loss*, 197–226.

Sun, X. (2002). Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. In *2002 ieee international conference on acoustics, speech, and signal processing* (Vol. 1, pp. I-333-I-336). doi:10.1109/ICASSP.2002.5743722

Swingler, K. (2011). Multi-Layer Perceptrons.

Taylor, J. (2012). *Raising generation tech: Preparing your children for a media-fueled world*. Sourcebooks.

Teager, H. M. & Teager, S. M. (1990). Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract. *Speech Production and Speech Modelling*, *55*, 241–261. doi:10.1007/978-94-009-2037-8

Terzis, V., Moridis, C. N., & Economides, A. A. (2012). The Effect of Emotional Feedback on Behavioral Intention to Use Computer Based Assessment. *Computers & Education*. doi:10.1016/j.compedu.2012.03.003

Tolkmitt, F. J. & Scherer, K. R. (1986). Effect of Experimentally Induced Stress on Vocal Parameters. *Journal of Experimental Psychology: Human Perception and Performance*. doi:10.1037/0096-1523.12.3.302

Van den Eijnden, R. J., Lemmens, J. S., & Valkenburg, P. M. (2016). The Social Media Disorder Scale. *Computers in Human Behavior*, *61*, 478–487. doi:10.1016/j.chb.2016.03.038

Väyrynen, E. (2014). *Emotion recognition from speech using prosodic features*.

Ververidis, D. & Kotropoulos, C. (2003). A Review of Emotional Speech Databases. *Proceedings of the 9th Panhellenic Conference on Informatics (PCI)*, 560–574.

Vošner, H. B., Kokol, P., Bobek, S., Železnik, D., & Završnik, J. (2016). A bibliometric retrospective of the journal computers in human behavior (1991â2015). *Computers in Human Behavior*, *65*(Supplement C), 46–58. doi:https://doi.org/10.1016/j.chb.2016.08.026

Womack, B. D. & Hansen, J. H. L. (1996). Classification of speech under stress using target driven features. *Speech Communication*, *20*(1-2), 131–150. doi:10.1016/S0167-6393(96)00049-0

Wu, C.-H., Yeh, J.-F., & Chuang, Z.-J. (2009). Emotion perception and recognition from speech. In J. Tao & T. Tan (Eds.), *Affective information processing* (pp. 93–110). London: Springer London. doi:10.1007/978-1-84800-306-4_6

Yang, N., Muraleedharan, R., Kohl, J., Demirkol, I., Heinzelman, W., & Sturge-Apple, M. (2012). Speech-based emotion classification using multiclass svm with hybrid kernel and thresholding fusion. In *2012 ieee spoken language technology workshop (slt)* (pp. 455–460). doi:10.1109/SLT.2012.6424267

Yuncu, E., Hacihabiboglu, H., & Bozsahin, C. (2014). Automatic speech emotion recognition using auditory models with binary decision tree and svm. In *2014 22nd international conference on pattern recognition* (pp. 773–778). doi:10.1109/ICPR.2014.143

Zhou, G., Hansen, J. H. L., & Kaiser, J. F. (2001). Nonlinear feature based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing*, *9*(3), 201–216. doi:10.1109/89.905995

APPENDIX A

DEMOGRAPHIC INFORMATION QUESTIONNAIRE

Experiment No	Experiment Date
General Information	
Name Surname	
E-Mail	Birth Date (dd/mm/yy)
Vision Problem ☐ Yes ☐ No	Gender ☐ F ☐ M
Hearing Problem ☐ Yes ☐ No	Marital Status
Education Information	
University	Mother Tongue
Deparment	German Knowledge ☐ Yes ☐ No
Level ☐ UG ☐ BS ☐ MS ☐ PhD	Level (1-5)
	Theather Experience ☐ Yes ☐ No

APPENDIX B

TURKISH SENTENCE DATABASE

Actual Experiment Section				Actual Experiment Section		
No	**Turkish Sentences**	**NS**		**No**	**Turkish Sentences**	**NS**
1	Mağazaları dolaştım	8		31	Hem çalışıyorum hem okuyorum	11
2	Masa tahtadan yapılmış	8		32	Ayakta bekliyordu	7
3	Kitaplar yatağın altında	9		33	Makineye deterjan koydu	9
4	Kamera çantanın içinde	9		34	Veli toplantısı başlıyor	9
5	Gözlük masanın üstünde	8		35	Kırmızı bir yelek aldım	8
6	Kaşık masanın üstünde	8		36	Her sıcaklıkta buharlaşma olur	11
7	Havlular katlanmış	6		37	Merdiveni duvara yasla	9
8	Masa örtüleri sepette	9		38	Dantel oyası öğreniyorum	10
9	Dünya yuvarlaktır	6		39	Kovboy filmi izliyorlar	8
10	Masanın rengi kahverengi	9		40	Ovalar az eğimli yerlerdir	10
11	Araba yıkandı	6		41	Kuruma uygun giyinmişti	9
12	Adam güvertede duruyor	9		42	Etik kuralları gözden geçirdik	11
13	Çanta odanın içinde	8		43	Maliye kısmını o halledecek	10
14	Portakal turuncu renklidir	9		44	Bankadan numara almak gerek	10
15	Masanın dört ayağı var	8		45	Bodrum'un halıcıları meşhur	10
16	Dolabın üç çekmecesi var	9		46	Baca temizliği bugün yapıldı	11
17	Şişenin mavi kapağı var	9		47	Köy halkı meydanda toplanmıştı	10
18	Bu borular metalden yapılmış	10		48	Ders listesi duvarda asılı	10
19	Odada bir sürü bidon var	9		49	Yarın ormana gidilecek	9
20	Yere, desenli bir halı serdim	10		50	Şirket ithalatla ilgileniyor	11
21	Bu bir çöp kutusu	6		51	Enstitü yirmi yıl önce kuruldu	11
22	Meyveler buzdolabında	8		52	Firma kuruluş aşamasında	10
23	Dünya bir gezegendir	7		53	Binanın önünde buluşalım	10
24	Kız, trenle seyahat ediyor	9		54	Kilise ayini sabah başlıyor	11
25	Köpek oyuncak topu getirdi	10		55	Bozkırlarda yağış miktarı azdır	11
26	Tablo duvarda asılı	8		56	Emlak borsası çok değişti	9
27	Futbol bir spordur	6		57	Kürek sporları suda yapılır	11
28	Çocuk yarın eve varıyor	9		58	Bu hafta dersler başladı	8
29	Kıyafetler çok renkli	7		59	Kibritsiz kandil yaktı	7
30	Gençler müzik dinliyorlar	8		60	Yanımıza konserve almıştık	10

61	Kız radyoyu açtı	6		100	Kendime bere ördüm	7
62	Bu bir belgesel filmi	7		101	Geleneklerine bağlı bir toplum	11
63	Bu bir tarih kitabı	7		102	Anayasa, kurallar bütünüdür	11
64	Bahçe çiçeklerle dolu	8		103	Yarışma kurallarını okuduk	11
65	İnsanlar tiyatroya gidiyor	10		104	Ve perde açıldı	6
66	Bugün konsere gidecekler	9		105	İhtiyacına yönelik seçti	10
67	Ormanda ağaçlar var	7		106	Tencere takımını da aldık	10
68	Apartman dairesinde kalıyor	11		107	Tampon bölgede oturuyorlardı	10
69	Okulun yurdunda kalıyor	9		108	Garson tepside çay taşıyordu	10
70	Penguenler kutuplarda yaşar	10		109	Hayvanlar ahırda kalıyor	9
71	Biri pencereyi kapatmış	9		110	Arapça soldan yazılır	8
72	Evin duvarlarını boyadık	10		111	Yumurtaları sepete koy	9
73	Her gün sekiz saat uyurum	9		112	Yoldaki adam adres soruyor	10
74	Kahvaltı önemli bir öğündür	10		113	Çadırlarını buraya kurarmış	11
75	Banka sokağın karşısında	9		114	Ürünler rafta duruyor	8
76	Ablam seneye geliyor	8		115	Kapının malzemesi ahşap	9
77	Bu yüzey sert maddeden yapılmış	10		116	Spor aboneliğim sona erdi	11
78	Çatallar çekmecenin içinde	10		117	Grafikleri inceliyorum	10
79	Belgeler kurula ulaştı	9		118	Duvara çivi çaktım	7
80	Ev ilanlarına bakıyordu	10		119	Gazete yazılı basındır	9
81	Kasaba meydanında toplandık	10		120	Kardeşinin saçlarını taradı	11
82	Kahvaltıda mutlaka elma yer	10		121	Arabanın sahibi geldi	9
83	Yol genişliği 4 metre kadar	9		122	Kayınpeder koltukta oturuyor	11
84	Tenekenin içinde yağ var	9		123	Ütü iyice ısınmış	8
85	Başlık büyük harflerle yazılır	10		124	Noterlik bir kamu hizmetidir	10
86	Hoca tahtaya bir şeyler yazdı	10		125	Vinçle yükleri kaldırdık	8
87	Ceketin yakasını ütüledi	11		126	Nehrin akıntısı hızlandı	9
88	Bugün kalın montunu giymişti	10		127	Pastayı sekiz parçaya böldük	10
89	Fırçayı tabloya dokundurdu	10		128	Vitesi dörde attım	7
90	Demir en yaygın dördüncü metaldir	11		129	Madde atomlardan oluşur	9
91	Televizyon açık kalmış	8		130	Oy vermek vatani bir vazifedir	11
92	Türkiye'de çok maden çeşidi var	11		131	Çark dönmeye başladı	7
93	Üyeler toplantıya çağırıldı	11		132	Puro bir tütün mamulüdür	9
94	Radyonun ayarı düzeltilmeli	11		133	Petrol mineral yağıdır	8
95	Bir bayisi de buraya açılmış	11		134	Ses mağarada yankılandı	9
96	Endüstri gelişmekteydi	9		135	Hep salonda oturulur	8
97	Hakan Bey notları sisteme girdi	11		136	Köpeği aşıya götürdük	9
98	On adet rakam vardır	7		137	Kendi markasını almıştı	9
99	Bu resme montaj yapılmış	8		138	Direksiyona o geçti	8

139	Üç işçi bir haftada boyadı	10		178	Tibet tapınağına çıktık	9
140	Belgeleri imzalattık	8		179	Akım devre üzerinde dolaşır	11
141	Karar kamuoyuna sunuldu	10		180	Günlük söylemine başlamıştı	10
142	Kursiyerler bugün derse başladı	11		181	Eşyalarını kutuya yerleştir	11
143	Ağacın gölgesinde oturuyor	11		182	Tarihçi derse girdi	7
144	Sınıf başkanı seçildi	8		183	Şehir merkezini geziyorlar	10
145	Futbolcular sahaya çıktı	9		184	Dönem geçişleri yaşanıyor	10
146	Dizel içten yanmalı bir motordur	11		185	Otobüsün kapıları açıldı	11
147	Karga düz gagalı siyah bir kuştur	11		186	Teyzem kırmızı montunu giymiş	10
148	Eve yeni sayaç takıldı	9		187	Gökyüzü mavi renklidir	8
149	Vezir, satrançta en güçlü taştır	10		188	Türkiye yarımadadır	8
150	Kumandayı uzatır mısın?	9		189	Dünya yedi kıtadan oluşur	10
151	Çocuklar ip atlıyorlar	8		190	Çocuk gözlük kullanıyor	8
152	Ankara Türkiye'nin başkentidir	11		191	Telefon çalmaya başladı	9
153	Yemekleri buzluğa koydum	9		192	Öğrenciler yarın sunum yapacak	11
154	Bu gece dolunay var	7		193	Yazın memlekete geri dönecek	11
155	Buzullar yaz kış erimezler	9		194	Sabaha kadar çamaşırlar kurur	11
156	Deney yaparken maske takılmalı	11		195	Günde 3 litre su içmeliyiz	10
157	Bir saat altmış dakikadır	9		196	Evini kiraya verdi	8
158	Kurbağalar yüzebilirler	9		197	Her gün okula gidip geliyor	10
159	Hipofiz bir salgı bezidir	9		198	Günlük hafif bir yürüyüş yapmalı	11
160	İnanç bir şeye bağlılık duymaktır	11		199	Kuşlar gökyüzünde uçuyorlar	10
161	Alet çantasını getir	8		200	Masaya tabakları koydu	9
162	Kemik, dokuların en sertidir	10		201	Kaşıklar çekmecede	7
163	Saçıma sprey sıktım	7		202	Çorbayı yudumladılar	8
164	On kiloluk bagaj hakkımız var	10		203	Vardığında kahvaltı başlamıştı	11
165	Eşyaları koliledik	8		204	Bekçi görev yerinde bekliyor	10
166	Değnek tahtadan yapılmış	8		205	Soru cevap kısmına geçelim	10
167	Banka saat dokuzda açılıyor	11		206	Körfez boğazda su çıkıntısıdır	11
168	Yol açık görünüyor	7		207	Anahtarlığı çantasına attı	11
169	Futbol turnuvası düzenlediler	11		208	Çantasını koluna taktı	9
170	Boşnak böreği yapabiliyor	11		209	Sokakta kitap bakıyor	8
171	Nesne cansız bir varlıktır	8		210	Ellerini yıkadı	7
172	Belgeyi tercüme ettirdi	9		211	Bugün misafir gelecek	8
173	Arılar bal peteği yaparlar	10		212	Pazardaki tüm tezgahları gezdi	11
174	Dört kelimelik bir cümle oluştur	11		213	Manavdan meyve alıyor	8
175	Proje hibeleri açıklanmış	11		214	Işıkları söndürdü	7
176	Şehrin planlaması çok güzel	10		215	Toplantı salı günü yapılacak	11
177	Deyimler sözlüğüne bakmalısın	11		216	Her gün mahallede yürüyüş yapar	11

217	Kırtasiyeye girdi	7
218	Sokakta ıslık çalıyordu	9
219	Damat tıraşı oluyor	8
220	İki katlı köşkte oturuyorlar	11
221	Plastik sandalye almışlar	9
222	Radyoda müzik çalmaya başladı	11
223	Günde bir şişe soda içilmeli	11
224	Molada oturmuş konuşuyorlar	11
225	Teneffüs zili çaldı	7
226	Öğle saati yaklaşmış	8
227	Okullar ara döneme girdi	10
228	Projede bayağı yol kat etmişler	12
229	Bisiklet turu yarın yapılacak	11
230	Ormandaki ağaçlar yeşil renkli	11
231	Kalorifer odayı ısıtıyor	11
232	Çiçekler pencerenin önünde	10
233	Lisansı sekiz dönemde bitecek	11
234	Kalemin ucu kırılmış	8
235	Aynada kendimi gördüm	8
236	Arabanın kapısını kapattım	11
237	Üstüme temiz bir şeyler giyeyim	11
238	On yedi asal bir sayıdır	10
239	Yarın on sekiz yaşına girecek	11
240	Dışarda serin bir hava var	9
241	Piksel görüntünün yapı taşıdır	11
242	Ateş, taş devrinde bulundu	9
243	Resmin gölge yoğunluğu derindi	11
244	İki, tek çift asal sayıdır	9
245	Bu kurumun belli kuralları var	11
246	Şemsiyesi kırmızı renkti	9
247	Masalar paralel dizilmişti	10
248	Yemeğini bitiren kalkıyordu	11
249	Bugün pembe kazak giymiş	8
250	Annem perdeleri açmış	8
251	Camları bugün sildirdim	8
252	Akşam trenine yer ayırttım	10
253	Tiyatro biletlerini aldım	10
254	Nöbetçi eczaneye girdi	9
255	Yeşil ayakkabılarını giydi	11
256	Günübirlik gezi düzenlenecek	11
257	Asfalt yolda gidiyorlar	8
258	Eve girince montunu çıkardı	11
259	Yeni kitabı yayınlanacak	10
260	Dizi izlemeye başladı	9
261	Emin adımlarla yürüyordu	10
262	Hoca birazdan ara verecek	10
263	Bir gün kitap yazabilirim	9
264	Kısa film festivali başladı	10
265	Telefon şarj oluyor	8
266	Arkadaşlarını bekliyor	9
267	Bu gece yeni dizi başlayacak	11
268	Çiçekleri suluyor	7
269	Dört raflı bir dolap aldım	8
270	Topu karşıya doğru savurdu	10
271	Kağıtları zımbaladı	8
272	Fotoğraf albümünü düzenliyor	11
273	Tabloları duvara astı	9
274	Dereler tepeler aştı	8
275	Televizyonun karşısında uyur	11
276	Koşarak otobüse bindi	8
277	Yemekler servis edilecek	9
278	Gözlerini dinlendiriyordu	10
279	Şu an gazetesini okuyor	10
280	Sokakları adımlıyordu	9
281	Motosiklet sürüyor	7
282	Yeni yeni yürümeye başladı	11
283	Doktora kontrole gittim	9
284	Hemen her konuda anlaşırız	10
285	Bir köy kasabasına gelmişlerdi	11
286	Handan içeri girdiler	8
287	Ata binmeyi öğrenecek	9
288	Fırına yemeği koydu	8
289	Saçlarını taradı	7
290	Duvarlar boyandı	6
291	Alışverişe gidelim	8
292	Taburede oturuyordu	9
293	Köşeyi dönünce onu gördü	10
294	Durağa gelince inecek	9
AVARAGE =		9,28

Practical Section		
No	**Turkish Sentences**	**NS**
1	Fazla oturmadılar	7
2	Konuşmacıyı dinledi	8
3	Bugün yapabilirim	7
4	Bugün gitmekten vazgeçti	6
5	Onun birçok kitabı var	8
6	Öğrenciler okulun bahçesinde	11
7	İki grupta incelenecek	10
8	Bu kitabı ben de okudum	9
9	Çayınızı hazırlarım şimdi	10
10	Benden kazak aldı	6
11	Her gün kitap okumalısın	9
12	Öğrencilere ödev vereceğim	11
13	Gökyüzüne öylece bakakaldım	11
14	İşe erkenden giderim	8
15	Evleri bizimkinden büyüktür	10
16	Gelişmesi için çalışıyoruz	11
17	Otobüse yetişebilir	9
18	Dışarıda birkaç kişi vardı	10
19	Çocuklara dağıttılar	8
20	İçeri girdi bize selam verdi	11
21	Akşam maç yapacaklar	7
22	Sende ders notları varmış	8
23	Dumandan çizgiye baktı	8
24	Yarın bir tanıdığa gideceğiz	10
25	Ev alma komşu al	6
26	Öğrencilerimizden olan Muhsin	11
27	O da ziyarete gelecekmiş	10
28	Çanakkale'yi de gezerdik	9
AVARAGE =		8,89

APPENDIX C

PARTICIPANTS INFORMATION

ID	Age	Gender	Eysenck Result	German Level	Acting Experience	Education Level
1	26	m	ext	1	No	Phd
2	28	f	ext	1	Yes	Phd
3	26	f	ext	2	Yes	Ms
4	26	f	ext	1	Yes	Ms
5	24	f	ext	0	No	Ms
6	26	f	ext	0	No	Ms
7	26	m	int	0	No	Phd
8	24	m	ext	0	No	Ms
9	31	f	int	2	Yes	Phd
10	31	f	int	0	No	Ms
11	21	f	ext	1	Yes	Bs
12	25	f	ext	1	No	Ms
13	22	m	int	0	No	Ms
14	27	f	ext	0	No	Phd
15	20	f	int	0	No	Bs
16	18	m	ext	0	No	Bs
17	24	m	int	0	No	Ms
18	24	f	ext	0	No	Ms
19	28	m	int	1	No	Ms
20	41	f	ext	0	Yes	Ms
21	45	f	ext	0	Yes	Bs