A NOVEL APPROACH TO EMOTION RECOGNITION IN VOICE: A CONVOLUTIONAL NEURAL NETWORK APPROACH AND GRAD-CAM GENERATION

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

SALİH FIRAT CANPOLAT

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

JUNE 2019

A NOVEL APPROACH TO EMOTION RECOGNITION IN VOICE:

A CONVOLUTIONAL NEURAL NETWORK APPROACH AND GRAD-CAM GENERATION

Submitted by SALİH FIRAT CANPOLAT in partial fulfillment of the requirements for the degree of **Master of Science in Cognitive Science Department, Middle East Technical University** by,

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

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

Prof. Dr. Deniz Zeyrek Bozşahin
Supervisor, **Cognitive Science Dept., METU** _____

Examining Committee Members:

Prof. Dr. İcâl Ergenç
Linguistics Dept., Ankara University _____

Prof. Dr. Deniz Zeyrek Bozşahin
Cognitive Science Dept., METU _____

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

Date: **06/27/2019**

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: **Salih Fırat, Canpolat**

Signature **:** _____

ABSTRACT

A NOVEL APPROACH TO EMOTION RECOGNITION IN VOICE:
A CONVOLUTIONAL NEURAL NETWORK APPROACH AND GRAD-CAM GENERATION

Canpolat, Salih Fırat
MSc., Department of Cognitive Sciences
Supervisor: Prof. Dr. Deniz Zeyrek Bozşahin

June 2019, 75 pages

Emotion is one of the essential components in human and human-machine interaction. One of the most common communication channels is the sound. Understanding the underlying mechanisms of emotion recognition in the sound signal is an essential step in improving both types of interaction. For this purpose, we developed an emotion recognition model, and a Turkish-specific database, referred to as the Turkish Emotion-Voice (TurEV) database. The database contains one-word-vocalizations of four emotion types; angry, calm, happy, and sad in three different frequency bands. The model was trained using TurEV, and human validation studies were conducted. The results indicate that the model is feasible for emotion recognition tasks. The comparison of the humans with the computational model indicate that the model achieves better results using feature-rich frequency bands, the humans use all other aspects of the sound signal.

Keywords: cnn, emotion, voice, corpus, Turkish

ÖZ

SESTE DUYGU TANIMLAMASI ÜSTÜNE YENİ BİR YAKLAŞIM:
KONVOLUSYONEL SİNİR AĞLARI VE GRAD-CAM OLUŞTURULMASI

Salih Fırat Canpolat

Yüksek Lisans, Bilişsel Bilimler Bölümü

Tez Yöneticisi: Prof. Dr. Deniz Zeyrek Bozşahin

Haziran 2019, 75 sayfa

Duygu, insan ve insan-makine etkileşiminin temel bileşenlerinden biridir. İnsan ve insan-makine etkileşiminde sık kullanılan iletişim kanallarından biri de sestir. Ses sinyalinde duygu tanımayı sağlayan temel yapıları anlamak iki tip etkileşimi de geliştirmek için önemli bir basamaktır. Bu amaçla, bu çalışma kapsamında, yeni bir duygu tanıma modeli ve Türkçeye özgü olan, Türk Ses-Duygu (TurEV) veritabanı geliştirildi. Veritabanı, dört duygu tipinin (kızgın, sakin, mutlu ve üzgün) üç farklı frekans bandında bir kelimelik seslendirmelerinden oluşmaktadır. Model, TurEV kullanılarak eğitildi ve insan doğrulama çalışmaları yapıldı. Sonuçlar, modelin duygu tanımada kullanılabilir bir yapıya sahip olduğuna işaret etmektedir. Karşılaştırmalı analizler, bilgisayarlı modellerin özellik bakımından zengin frekans bantlarını kullanarak daha iyi sonuç almasına karşın, insan zihninin ses sinyalinin diğer tüm özelliklerini kullandığını göstermektedir.

Anahtar Sözcükler: cnn, duygu, ses, Türkçe

DEDICATION

I dedicate this study to the most important people in my life, my mother and my father; Sezin and Şahin Canpolat. When everyone lost hope, you were there. Words can not describe my gratitude.

ACKNOWLEDGMENTS

Foremost, I would like to express my deepest gratitude for my advisor Deniz Zeyrek Bozşahin. Under her guidance I have gained more than mere knowledge, I have gained discipline and grown as a scientist.

Besides my advisor, I would like to thank my thesis committee: İcâl Ergenç and Cengiz Acartürk for their insightful comments, hard questions, and encouragements.

Moreover, I would like to offer my gratitude to my undergraduate professors Timuçin Aktan and Kahraman Kıral, for their encouragement and wisdom.

In every study, there are heroes that make it possible. These heroes are unnamed, unknown, and forgotten. These heroes are assistants who collected data, actors, and participants. I'm deeply indebted to them. I would like to honor them by thanking each of them by name. First of all, Şiyar Morsünbül, thank you for being the best team leader and your infinite patience. Moreover, I'd like to thank Kübra Yılmaz, Oğuzcan Ülgen, Cansu Aşıcı, Ayşe Nur Balli, Dilşah Suretli, Gamze Gücenmez Üngörmüş, Mustafa Özaydın, Enis Dönmez, Emre Erçin, Baki Çağdaş Aydın, Reyyan Baş, Ümit Murat, Dilara Uslu, Müge Çelikel, Büşra Gizem Sönmez, and İrem Yıldız.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANOVA	Analysis of Variance
CNN	Convolutional Neural Network
CSV	Comma Separated Value
DNN	Deep Neural Network
F_0	Fundamental Frequency
FFT	Fast Fourier Transform
FN	False Negative
FP	False Positive
GMM	Gaussian Mixture Model
GPU	Graphical Processing Unit
Grad-CAM	Gradient Category Activation Map
HMM	Hidden Markov Model
Max	Maximum
MFCC	Mel Frequency Cepstral Coefficients
Min	Minimum
NN	Neural Network
PCM	Pulse Code Modulation
ReLU	Rectified Linear Unit
STD	Standard Deviation
STFT	Short Time Fourier Transform
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TurEV	Turkish Emotion Voice Database
VoIP	Voice to IP
WAV	Windows Audio Waveform

Chapter I

1. Introduction

Human emotion is one of the components that represent human cognition interculturally. According to Ekman, emotions are stable through different cultures (Ekman, 1972; Ekman et al., 1987). Emotions leave markers in vocalizations, prosody, facial expressions, and biological processes such as hormone levels and blood pressure (Cowie & Cornelius, 2003; Liscombe, Venditti, & Hirschberg, 2003; Vogt & André, 2005). To analyze emotion, multimodal models were also developed and they attained a high level of success. Among the models, those based on vocalization were the most common ones. In fact, vocalization is one of the least intrusive and the most accessible form of marker for the emotions.

Emotions are one of the primary requisites of a healthy interaction between individuals. Lacking the ability to convey or understand the emotions is part of many pathologies (A. S. Cohen, Najolia, Kim, & Dinzeo, 2012). Studies have been conducted on the subject matter of special education and rehabilitation of such individuals (Konstantareas, 2006). Understanding the underlying mechanisms in emotion recognition, at least in vocalizations is essential to understand human interaction. Emotion recognition does not only affect communication between humans, but it is also one of the critical factors in human-machine interaction. In the last two decades, the amount of human-machine interaction has increased drastically. Human is generally the unpredictable part in human-machine interaction; therefore, understanding human emotion is one of the critical aspects of improving this interaction.

In the field of emotion recognition, a great deal of improvement has been made through vocal markers. The models which performed barely above chance now have over 90% accuracy rating (Nwe, Foo, & De Silva, 2003; Wang, 2014). As computational power increased, older models were replaced with support vector machines (SVM) and different kinds of neural networks. However, most of these studies have tackled the emotion recognition problem through a performance perspective. To the best of our knowledge, studies concerned with the performance of both machines and human judges are rare.

Corpus studies are an essential part of emotion recognition. Although computational approaches can be a part of the corpus studies, their main focus is creating a validated corpus of emotion. To date, emotion corpora with different properties have been compiled. For example, Berlin Emotional Speech Database (Burkhardt et al., 2005) consists of a single language German, has 800 vocalizations, and is free. INTERFACE (Hozjan, Zdravko, Asuncion, Antonio, & Albino, 2002) on the other hand consists of English, Slovenian, Spanish, and French has 175 to 190 vocalizations and is commercially available.

The Turkish language hardly shares the rich literature in the emotion recognition field other languages enjoy. The link with emotion and voice is even less studied. Among

the studies that exist, we can name works on fundamental frequency (Fidan, 2007), and on machine learning models in emotion recognition (Erdem, 2014). However, to the best of our knowledge, there are no voice-emotion corpora studies in Turkish. Thus, this study aims to address these issues and by providing a machine learning model, it intends to set up the first step to reveal the possible link between voice and emotion in Turkish.

1.1 Goals

In this thesis, we aimed to reach four goals. Our first goal was to develop a novel model type for emotion recognition that is trained using a set of Turkish words, which is robust to noise and manipulation, and whose decision-making process is visualized with various techniques. Our second goal was to compile the Turkish Emotion Voice Database (TurEV Database) that is representative of the acoustic changes in Turkish one-word vocalizations and is validated by expert judges. We hope that the TurEV Database will be a precursor for the future studies in emotion-voice corpora studies. Our third goal was to do a comparative analysis between the predictions of the model and the judges.

1.2 Scope

The machine learning model used in this thesis is a variant of the Convolutional Neural Network Model. The model is designed to allow Grad-CAM model construction and heat map generation. Unlike other emotion recognition models, the model used in this thesis accepts images. The Turkish Emotion Voice Database (TurEV) Database consists of vocalizations of four emotional categories; angry, calm (neutral), happy, and sad voiced by six amateur actors (3 male, three female). TurEV consists of words in three different frequency bands.

TurEV is based on 82 words, each of was vocalized in each emotional category by each amateur actor. Each vocalization in TurEV is accompanied by a spectrogram of that vocalization.

In the thesis, the performance of the model and the judges are compared using a series of analyses, the results of which are finally presented in contingency tables. The judges' assessment of emotions is considered as the golden key on the basis of which machine learning is assessed.

This thesis is primarily concerned with producing an emotion-voice database, and a neural network model that predicts four emotions. The focus of the thesis is on these productions and their analysis.

1.3 The Research Gap in Contemporary Computational Voice-Emotion Studies

The review of literature in Chapter II will show that there are certain gaps in the contemporary computational voice-emotion studies. For example, the models produced are accurate and fast, yet they are either not fully transparent in their decision-making process, or they merely provide the underlying feature maps. The human voice and its interaction with emotion create non-linear relationships in the features. It is possible to derive information about the features, yet the existing models only provide linear features such as the fundamental frequency (F0) contour. A novel type of model that tackles the problem of emotion recognition is entirely missing in Turkish.

Secondly, the number of corpora used in emotion recognition studies has significantly increased. These corpora have many different properties, such as the number of languages, several vocalizations, accessibility, validation techniques, so on and so forth (El Ayadi, Kamel, & Karray, 2011). However, an established voice-emotion corpus for the Turkish language that is open to the public does not exist. The studies conducted in this area have either built their own dataset or used corpora of other languages (Erdem, 2014).

1.4 Contributions of the Thesis

In this thesis, we aimed to contribute to the voice-emotion studies literature in three significant aspects, namely, methodologically, empirically, and cognitively.

1.4.1 The Methodological Contribution

We used a novel method in order to create a feature map that is representative of the Turkish sound signal in time and frequency domain providing different frequencies. The method we chose was short-time Fourier transformation, and feeding this information into a convolutional neural network (CNN) (Owens & Murphy, 1988). We have chosen the CNN because it allows the object detection paradigm to be applied to the data (Cai, Fan, Feris, & Vasconcelos, 2016). Moreover, the CNN architecture allowed us to use Grad-CAM (Class Activation Map), which enabled us to learn more about the model's decision process and extract the heat maps that show essential features.

1.4.2 The Empirical Contribution

Corpora play a critical role in voice-emotion studies. In this thesis, we compiled a comprehensive database called the Turkish Emotion Voice Database (TurEV). TurEV consists of a corpus component and an analysis component. The corpus component includes vocalizations of 82 words, three different versions of these vocalizations in different frequency bands, spectrograms derived from these vocalizations, and heat maps. The analysis component includes F_0 values, validation statistics, model

decision statistics, and heat maps for the validation set. Moreover, the corpus is validated by expert judges for each frequency band, and the data from the comparative analysis is included in the database.

1.4.3 The Cognitive Contribution

The present study investigates the voice-emotion corpus through a computational model as well as an ablation study.[1] In the context of our study, ablation involves removing the voice properties from certain frequency bands of the model. We then test the predictions of the model in each frequency band as well as asking the human judges to assess their perception of emotion category in each frequency band. The discrepancies and the parallelisms between the model and the judges are revealed and analyzed. In this way, we were able to make inferences about the nature of phonological information provided in different frequency bands, because we could understand how the human ear perceives emotion even when information in certain frequency bands is missing.

1.5 Structure of the Thesis

Chapter I - Introduction

In this chapter we presented a brief view of the thesis in terms of its goals, contributions, the gaps in research in this emotion-voice studies, and what our contribution will be.

Chapter II - Background

In this chapter, we present the theoretical background of emotion-voice studies. Theoretical background includes computational approaches to emotion-voice studies, the phonology of emotion, the structure of emotion, and corpora studies conducted on emotion-voice.

Chapter III - Method

In this chapter, we present the methodological approach used in this thesis. The chapter starts with how the data is handled including its collection and validation, it continues with the presentation of the neural network architecture, and ends with the information on statistical analysis conducted on the fundamental frequency (F_0).

Chapter IV – Analysis and Results

In this chapter, we present the results of the analysis conducted on the model, the validation study, and statistical analysis explained in Chapter III. The chapter starts with the analysis of the model, continues with the analysis of the human judges' decisions, and ends with the results of an analysis comparing machine learning with human learning.

[1] Ablation is a type of study in which models are tested with modified, removed, or damaged features. This allows models to be tested in the absence of manipulation of these features.

Chapter V – Turkish Emotion-Voice Database (TurEV Database)

In this chapter, we describe the TurEv database. We present the results of the Turkish Emotion-Voice Database study. We start with information on the corpus and the coverage of the database. Lastly, we present the evaluation of the corpus component of the TurEV Database.

Chapter VI – Conclusion

In this chapter, we summarize the results obtained in Chapter IV and conclude in light of the information presented in Chapter II. We discuss the implications and limitations of the thesis, and lastly, we conclude this chapter with suggestions for further research.

Chapter II

2. Background

2.1 Introduction

In this chapter we present the theoretical background of emotion, the emotion-voice studies, and give brief information on phonological aspects of emotion-voice. We present an overview of what emotion is, the relevant contemporary research on emotion, the features, and model types in emotion-voice studies, then we briefly introduce the role of phonology in emotion-voice studies. We conclude this chapter with information on the existing emotion-voice databases and their properties.

2.2 Emotion

Emotion is an essential part of not only interpersonal communication but also human-machine interaction. Various emotions are conveyed through voice as well facial expressions, posture, gestures, and such (Busso, Bulut, Lee, & Narayanan, 2008). However, the majority of the human-machine interaction happens through the voice channel with the assistants such as Microsoft's Cortana and Apple's Siri (Hoy, 2018). On the interpersonal communication side, phones allow humans to communicate through voice-only channels. This communication channel has been enhanced by the introduction of mobile phones and voice to IP (VoIP) communication systems. These changes have made emotion-voice studies inevitable. In order to proceed, one question requires to be answered: what is emotion?

2.2.1 What is emotion?

In order to work on a concept, it is imperative first to operationalize it. According to Scherer, emotion can be operationalized with a component process model, which divides the concept into three significant parts, namely, the function, subsystems, and components (Scherer, 1982). In this regard, emotion is an interconnected change that effects the subsystems in the component process model and that change must be in synch and be caused by an event that has importance to the organism (Scherer, 1987, 2001). Emotion also has its valence, namely positive and negative emotions. Within the boundaries of the thesis and the component process model, emotion can be measured using action functions that use the somatic nervous system manifested as a motor expression. This could be a facial expression or changes in intonation.

2.2.2 Affect, Emotion, and Mood

Emotion has a type-token relationship with affect and mood with both affect and mood being types of emotion; they have significant differences that affect their functions. Mood is a type of emotion that has a slow rate of change and relative longevity that lasts from hours to days. In pathological conditions such as depression, mood might not change for six months or longer. Affect, on the other hand, is short lasting, it can last from milliseconds up to an hour in some cases. During a conversation, affect can change in various ways.

The present thesis aims to study the affect type of emotion. During the study, affect and emotion will be used interchangeably.

2.3 Contemporary Research on Emotion Recognition

Contemporary research on emotion recognition through computational approaches is focused on audio, visual and audiovisual methods (Y. Kim, Lee, & Provost, 2013). Audiovisual methods provide ample amount of data and offer high precision which is a requirement in pathological cases such as depression (Cohn et al., 2009). In exchange for the increased accuracy, audiovisual models suffer from increased data size and problems in parallelization of the sound and image. On the other hand, using sound as the sole source of information in machine learning models lower data density and the total amount of information. Lowering the signal to a single type, however, removes the need to parallelize two different data types and allows the models to use the sound in various ways.

The sound signal can be utilized in different ways to create estimators that categorizes and recognizes different emotions. A widely used approach is to extract handcrafted features from the signal then use these features to classify the emotions. This is a robust approach in a given dataset with strong internal validity. Handcrafting the features allows researchers to pinpoint the specific properties of the data and extract them as needed. This approach is robust in terms of the classification made for the dataset they are developed for; however, they it lacks external validity. Therefore, such approaches are not accurate on the novel data. One exception to this phenomenon is the depression studies. According to the Diagnostic and Statistical Manual of Mental Challenges, V (DSM V), the major depressive disorder, is a mood disorder characterized by an extensive and unrealistic feeling of sadness for a prolonged time (American Psychiatric Association, 2013). The research performed on the categorization of the depression through vocal prosody can successfully be generalized to the cross-corpus applications (Alghowinem, Goecke, Epps, Wagner, & Cohn, 2016; Mitra, Shriberg, Vergyri, Knoth, & Salomon, 2015). This approach has high external validity in terms of the stability of depressive expressions across cultures. As depression prolongs sadness, it lowers the energy in the sound signal, widens the gaps in speech, and lowers the variation of pitch regardless of the culture

and language. Depression categorization through vocal biomarkers also focuses on the binary categorization of depressed. Therefore, it is possible to use handcrafted features.

Feature extraction can also be done with the brute force approach and is shown to be a successful way of extracting features (Pachet & Roy, 2009; Vogt & André, 2005). The brute force approach generates thousands of features from the present data then eliminates those using different procedures. Culling of the features that do not improve the model can be done in various ways. Genetic algorithms can be used in this manner, change in explained variance or cross-correlation can be used as an indicator. The brute force approach is slow and computationally expensive. Advantages of a brute force approach are using a computational method to extract most robust features instead of handcrafted features. This eliminates the need of human expertise in feature extraction, it also eliminates human error. Therefore, this approach adds to the overall strength of the model.

2.4 Common Feature Types Used in Emotion Recognition

Depending on the type of research, the model, the estimator, and the corpus, different set of feature types can be used. Except for rare occasions, features are singular estimations of the window functions, and their vectorized representations are used in the models. These features can be categorized into three different groups being; continuous, spectral, and TEO-Based (El Ayadi et al., 2011).

2.4.1 Continuous Features

Pitch, energy, and formants are the most common continuous features. These features are not mutually exclusive, and they can have a part-whole relationship. They are all time dependent and generally used within a window. A formant range has its energy value, and different pitches have different densities at different formats. Different measures can be used for these continuous features. Most commonly used measures are mean, standard deviation, range, skewness, and kurtosis.

Pitch is the frequency of the sound -- a high pitched sound has high amplitude at higher frequencies, whereas a low-pitched sound has high amplitude at lower frequencies. Unlike artificially created sounds that can have a single pitch value such as 123.47 for the musical note A#, the human voice is a mix of different frequencies at different amplitudes.

Formants are spectral shifts in the vocal tract that result from acoustic resonance, and they are represented by intervals in hertz. In most phonology studies, five formants are used (up to 5000 Hz). Formant 1 (F_1) resides between 500 and 1000 hertz. The values of the formants are presented in Table 1. This thesis, however, uses frequencies that reach up to 8000 Hertz. In the context of the thesis, frequency band between 5000 and 8000 will be called the non-formant band.

Table 1: Formants and their frequency bands

Name	Lower Limit	Higher Limit
Formant 1 (F_1)	500	1000
Formant 2 (F_2)	1000	2000
Formant 3 (F_3)	2000	3000
Formant 4 (F_4)	3000	4000
Formant 5 (F_5)	4000	5000

Energy is the total signal strength. Studies have shown that one of the dependencies of energy is the emotional state (Cowie & Cornelius, 2003). The energy of a signal with finite length can be explained by its power. The power is the sum of energy in a given time. The energy of a signal varies for a given timeframe; the energy may be high or low depending on the signal. The emotion categories sad and calm produce low energy sound signals whereas the emotion categories angry and happy produce high energy sound signals. Moreover, the energy may vary depending on the phonological properties.

2.4.2 Spectral Features

The continuous features are dependent on the time domain, whereas the spectral features are dependent on the frequency domain. They represent the spectral distribution of the sound in a given window or the whole signal itself. Autocorrelation, Mel frequency cepstral coefficients (MFCC), and fast Fourier transformations (FFT) are several examples of spectral features (El Ayadi et al., 2011). The thesis uses a hybrid feature. The hybrid-feature is spectrogram derived from short time Fourier transformation (STFT). The result of STFT represents the change of frequencies and their amplitude in time. Because the STFT is both time-dependent and frequency-based, it is an hybrid type of feature. The STFT represents the change in frequencies across time. The STFT and spectrograms are rarely used in speech recognition. A recent study has shown that extracting and processing features as images using the STFT method is a viable option (Wang, 2014). However, this study uses the STFT method only as a feature extraction and processing tool rather than using it as the main input data.

2.5 Commonly Used Models in Emotion-Voice Studies

Various types of models have been used in emotion-voice studies depending on the available computational power and the selection of the features. Moreover, these models can be used in conjunction with other models or as hybrid models such as the Subspace Mixed Gaussian Hidden Markov Model which is a hybrid of neural networks and hidden Markov models. Each of these models is briefly explained below.

2.5.1 Hidden Markov Models

The Hidden Markov Models (HMM) are commonly used in speech signals for various purposes ranging from emotion recognition to word classification (Nwe et al., 2003). The Hidden Markov Models use the outcomes of the first-order Markov chains, and from there estimates what the chains consist of (Rabiner & Juang, 1986). Commonly used features in the HMMs are spectral features such as MFCCs. Although the HMMs no longer have their old popularity, their hybridizations with other models still produces robust results (Mao, Tao, Zhang, Ching, & Lee, 2019).

2.5.2 Gaussian Mixture Models

The Gaussian Mixture Models (GMM) are simple yet robust models; like the HMMs they use spectral features. The GMMs create a map of probabilities and categorize the sound signal according to its position in the probability space. Contemporary research shows that the GMMs have high performance when used in conjunction with optimized algorithms (Tang, Chu, Hasegawa-Johnson, & Huang, 2009).

2.5.3 Neural Networks

The Neural networks (NN) are prevalent in many different fields. However, they do not offer robustness or yield good performance when they are used on their own in emotion recognition through sound. There are several types of neural networks, such as the deep neural networks (DNN), recurrent neural networks (RNN), and convolutional neural networks (CNN). The DNNs and RNNs are generally used in conjunction with other models such as the HMMs. However, the CNN architecture is very rarely used. To the best of our knowledge, there are few published studies that use the CNN architecture for emotion recognition. For example, one study uses spectral estimates (Weißkirchen, Böck, & Wendemuth, 2018) and the other one uses one dimensional convolutions with long short term memory (Trigeorgis et al., 2016). Moreover, the study which had used spectral estimates (Weißkirchen et al., 2018) produced an average accuracy score of 52%, which is much lower than the average accuracy score achieved in this thesis.

The thesis utilizes CNN as its model framework. CNNs generally accept images as inputs, and they conduct convolution operation on the images. They are robust to the changes in orientation, distortion, and partial exposure of the subject. In the thesis, we aim to utilize the CNNs' ability to recognize objects in emotion recognition.

2.5.4 Support Vector Machines

The Support Vector Machines (SVM) are the most commonly used model types in emotion recognition. They accept both the spectral and the continuous features, and they have high performance. The SVMs work by increasing dimensionality and

making categorization at higher dimensions. The SVMs discriminate two classes with one versus all paradigm. However, the new algorithms allow the SVMs to make multiclass discrimination (Wu, Lin, & Weng, 2004).

In this thesis, we used the CNN architecture instead of the SVM architecture because the CNN architecture is robust to the noise in the data and it can produce heat maps. The heat maps can be used to probe into inner workings of a CNN model.

2.6 Phonology of Emotion

Studies on the phonology of emotion have mostly focused on fundamental frequency (F_0), and its properties such as value, pitch, jitter, contour, and tilt (Busso et al., 2008; Johnstone & Scherer, 1999; McGilloway et al., 2000; Paeschke, Kienast, & Sendlmeier, 1999). In the thesis, we analyzed the F_0 values as part of the corpus study, but our model was designed to learn the features in all five formants, as well as the features that reside in the non-formant frequencies. The frequencies over 5000 hertz are not studied within phonetics and generally have no formant labels. We refer to these frequencies as non-formant frequencies.

2.7 Emotion-Voice Databases

The validated corpora [2] is the essential component of any emotion-voice study. The construction of an emotion-voice database is a complex process. Some of the important components of the emotion-voice corpora and databases are presented below.

The essential components of emotion-voice corpora are;

- Emotion types
- Source of emotions
- Number of vocalizations and utterances
- Language(s)

The essential components of emotion-voice databases are;

- The corpus
- The validation study
- Peripherals such as spectrograms and statistics

[2] A In this thesis, the term corpus will be used for the linguistic component of the database such as vocalizations and utterances, and the term database will be used for the data as a whole such as the validation results, extracted statistics, models, etc.

We will compare the Turkish Emotion-Voice (TurEV) Database with the exemplars while explaining the components above to emphasize the differences and parallelisms.

2.7.1 Emotion-Voice Corpus Components

2.7.1.1 Emotion Types

The Emotion types are the number of emotion categories. This number can vary significantly between corpora, it can be four as in the KES Database (E. H. Kim, Hyun, Kim, & Kwak, 2007) and in the TurEV database. It can be as high as 20 as in The EU Emotion-Voice Database (Lassalle et al., 2018). Most studies stay within the confinement of the six base emotions, i.e., anger, disgust, joy, fear, sadness, surprise, as stated by Ekman (Ekman, 1972). A neutral emotional state is sometimes added. TurEV uses angry, calm (neutral), happy, and sad as its emotion categories.

2.7.1.2 Source of Emotions

The source of emotions can be professional or amateur actors, participants, or in-vivo samples.

Nearly all of the studies conducted in the emotion-voice area is conducted on the recordings generated by either professional or amateur actors. TurEV uses amateur actors (i.e. individual who have no training in acting), who simply act out a type of emotion and generate the required voice as if they were experiencing that particular emotion. Another way of using actors is by collecting the data from already recorded media such as TV series, movies, or other similar mediums.

Participants, on the other hand, participate in an experimental condition to induce emotion. In this kind of methodology, emotion induction through imagination is used (Johnstone & Scherer, 1999). This methodology is rarely used, but it produces the data with the most potent external validity.

In-vivo sampling is another method to obtain emotion-voice samples. It consists of the lengthy procedure of obtaining voice samples from call centers, auto dealers, etc., then processing them in chunks. Due to ethical implications, this method is rarely used. An example of this kind of sampling is Natural Database containing emotion-voice samples obtained from call centers (Morrison, Wang, & De Silva, 2007). This sampling method is used commonly in clinical studies for depression within ethical boundaries (Mitra et al., 2015).

2.7.1.3 Number of Vocalizations and Utterances

The number of utterances and vocalizations changes drastically among corpora. It can be as low as 80 as in the Pereira Database (Pereira, 2000) or can be as high as 16000

in the SUSAS Database (Hansen & Bou-Ghazale, 1997). The TurEV has over 1700 vocalizations produced by 6 actors vocalizing 82 words in 4 emotional categories.

2.7.1.4 Language(s)

The language aspect of a corpus determines the area(s) it can be used. Most corpora are developed for a single language. Most commonly used language is English. A multilingual corpus is one of the most valuable sources for emotion-voice studies because it allows cross-cultural studies. The EU Emotion-Voice Database with English, Swedish, and Hebrew languages is one of the recent and prime examples for this kind of practice (Lassalle et al., 2018). The TurEV supports only Turkish language.

2.7.2 Emotion-Voice Database Components

2.7.2.1 The Corpus

The corpus is the most critical component of an emotion-voice database. It is the only component that is necessary for the existence of an emotion-voice database. The accessibility component of a corpus is one of the main bottlenecks in emotion-voice studies. Nearly all of the emotion-voice corpora that exist are either private or have license fees. There are only a few corpora on a public domain such as The Berlin Emotional Speech Database (Burkhardt et al., 2005). The TurEV is also planned be on the public domai.

2.7.2.2 The Validation Study

The validation study ensures the validity of the study performed on the database. A validation study is generally performed by rating the items of the corpus by the human judges. The EU Emotion-Voice Database is one of the prime examples. In The EU-Emotion-Voice Database study, expert judges eliminated the recordings that do not reflect the emotional state, then non-expert judges validated the corpus on a grand-scale (Lassalle et al., 2018). In the validation study of the TurEV, three expert judges were used to rate the finished corpus with respect to which category each word fits. The detailed process of validation study is presented at 3.2.7.

2.7.2.3 The Peripherals

The peripherals are not an essential part of the database. However, they enrich the study. Peripherals can be frequency and formant information, statistical information, machine learning models, and similar information. TurEV offers all these peripherals with the addition of heat maps and spectrograms that allows researchers to conduct both a visual analysis and exploit differentmachine learning models.

2.8 Summary

In this chapter, we presented general information and theoretical background on the contemporary emotion-voice studies. We defined emotion in the context of our study, tapped on the contemporary research on the subject matter. We presented information on the standard features and model types used in contemporary research. We ended the chapter with information on the databases used in emotion-voice studies.

Chapter III

3. Method

3.1 Introduction

In the previous chapter, we presented information on the theoretical background of emotion-voice studies. In this chapter, we present the methodology used in the thesis. This chapter is divided into three main topics; the data, the machine learning experiment, and the frequency based statistical analysis.

3.2 Data

In this section, we introduce our data processing pipeline. We first explain the process from a bird's eye view, then move on to how we selected the words. We explain our procedure which involves amateur actors, data collection, data cleaning, and segmentation procedures. Finally, we explain our data validation procedure.

3.2.1 Overview

Single-word vocalizations were collected from amateur actors who joined the study through a convenience sampling method. Collected vocalizations were analyzed for their noise profile using the silent parts of the recordings. Each recording produced a unique noise profile; this noise profile was used for removing the noise from the end-to-end section of each recording. The de-noised recordings were trimmed after this operation. The trimming was done by cutting the excess silence from the sound file leaving 150 milliseconds of silence before the vocalization starts and after the vocalization ends. The result of this data cleaning process was used in two different analyses. The neural network model is the heart of the thesis as well as the frequency-based analysis as a part of the two-stage validity verification procedure.

3.2.2 Selection of One-Word Vocalizations

Eighty-two words were selected from *Türkçenin Ses Dizgesi* (Ergenç & Bekar Uzun, 2017). These words reflect different phonological properties of Turkish language. For instance; the initial sound ç in *çilek* is voiceless. However, when the same sound appears word-finally as in *kulaç, it* is voiced (Ergenç & Bekar Uzun, 2017) Such phonological variation allowed our model to be tested for different conditions and increased its robustness. Moreover, these words had already been investigated in their neutral (calm) emotional state, and their F0 values, density graphs, and their spectrograms up to 5000 hertz are provided (Ergenç & Bekar Uzun, 2017). In short, this set of words provided a perfect data and the baseline of our experiments.

3.2.3 The Amateur Actors

The amateur actors were recruited using convenience sampling. The actors were named amateur because they did not have professional acting experience. Six amateur actors were recruited, 3 of them being female and 3 of them being male. Age of the actors ranged from 23 to 35. Mean age of the actors was 26.83 with a standard deviation of 4.35. Amateur actors were presented with digital copies of the amateur actor manual, the amateur actor number, the amateur actor consent form, and a list of words. The information package presented to the actors are given in APPENDIX A.

3.2.4 Data Collection

Voice recordings were done by amateur actors without supervision; i.e. the amateur actors were not supervised by an assistant and they conducted the recording procedure by themselves base on the guidelines (cf. Appendix A). Collected emotional states were angry, calm (neutral), happy, and sad. These emotions were chosen because they fit in the different axes of the valence-arousal axis (Figure 18) (Barrett, 1998). The amateur actors were requested to vocalize each word in the list as if they were in that emotional state and were feeling the emotion in high intensity. The amateur actors uttered the words as many times as needed, and they picked the best result that expressed the aimed emotion. The amateur actors were given the freedom to present their single-word vocalizations to others and ask their opinion to informally validate the correctness of the emotional states. According to the feedback given by the amateur actors, they used outside feedback in their decision process. Moreover also according to the feedback given by amateur actors, the process of recording the single-word vocalizations took between 4 to 8 hours for four emotional states for each actor.

The amateur actors were told to vocalize all the words in the normal tone of voice but as if they are feeling the emotion in high intensity. Thus, all emotions were vocalized in high intensity except for calm (neutral) which is obviously the default state.

The single-word vocalizations were collected in two different environments depending on the hardware available to the amateur actors. The amateur actors with access to computers recorded their vocalizations using the freely available application Audacity 2.3.0. Audacity 2.3.0 is available for all of the commonly used operating systems. The amateur actors who had access to smartphones but not to a computer that can record voice adequately used the Sound Recorder application from Sony Mobile Productions. Regardless of the platform and the operating system, the vocalizations were recorded in the mono channel, had the WAV format, and were in 44100-hertz sampling rate.

3.2.5 Data Cleaning

The collected data were ported into Audacity 2.3.0 (Audacity, 2018). Each recording was first analyzed for its noise profile using parts of the recording that is supposed to contain no sound. Those parts of the recording were supposed to be silent; however,

they had noise. The procedure sampled this noise. The profile was then used for removing the noise from the file which it was generated from. Steps of noise removal are presented in Figure 1. The process involved the following steps:

1- The recording was loaded into Audacity 2.3.0.
2- The signal view was changed from the waveform into the spectrogram.
3- The area that contained no vocalization, therefore, supposed to be silent was selected.
4- The noise profile that was generated before was used in conjunction with the noise remover effect for the entirety of the recording.
5- The recording was trimmed leaving 150 milliseconds of silence at the beginning and end of it, and the rest of the recording was removed.
6- The recording was exported in 32-bit Float Pulse Code Modulation (PCM) WAV format (Microsoft, n.d.).

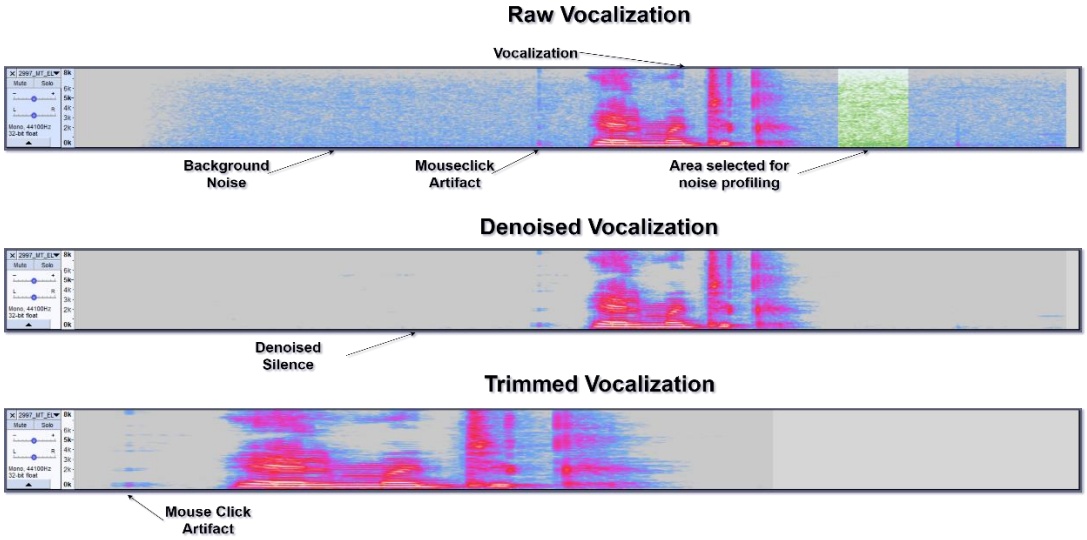

Figure 1: Steps of noise removal procedure in Audacity.

In this procedure, the spectrogram view was explicitly used in order to be able to view the low amplitude high-frequency parts of the vocalization as well as the noise recorded. The spectrogram view also allowed the file to be inspected more clearly. The noise profile generation used an internal statistical procedure within Audacity. The selected area for the profile generation should be at least 2048 samples long, which translates into 0.05 seconds when the 44100-hertz sampling rate is used (Audacity Team, n.d.). Uniform areas of noise longer than 2048 samples did not make any improvements in noise reduction, however more significant areas that contain different noise characteristics improve the noise reduction performance. In some cases, the noise reduction procedure had failed to remove a band of the noise that settled at 0-500 hertz band; in these cases, a second profile generation and the reduction was applied. Depending on the environment, the recording device, and the amateur actor, the denoising procedure managed to remove noise with varying amount of success.

The de-noised data was then inspected in the spectrogram view as well as the waveform view in order to determine the beginning and end of the single-word vocalization. After the inspection, it was inspected again by listening to the recording between 1x and 0.05x speeds. Once the beginning and the end of the recording was determined, 150 milliseconds of silence was kept at the beginning, and at the end of the recording, the rest of the silence was trimmed out. The main reason of adding this extra step was standardization of the recordings to a particular format, and the second reason was to eliminate excess silence. Elimination of excess silence without changing the structure of the word allowed the model to be able to work in a more data-dense space. The silence that was kept at the beginning and the end of the vocalization was used for compensation of possible human errors done in recording and trimming, and allow for future human inspection.

The denoising process removed the majority of the noise.This process was done with extreme care. Denoised recordings were inspected in different playback speeds. The process is depicted at Figure 1.
The data cleaning process was finished by exporting the processed recording in 32-bit Float PCM WAV format. During the exporting procedure, the file was scanned for possible tags embedded in the original WAV file and found tags were removed. 32-bit float PCM allowed for a broader representation of data with the 32-bit precision. The 16-bit signed integers were not preferred because they did not have the resolution the 32-bit float format can offer.

3.2.6 Segmentation of Data into Different Frequency Bands

The recordings were subjected to three different frequency manipulation. Each manipulation was conducted by the Butterworth filter.

A. The frequencies over 8000 hertz were trimmed out, and only the frequencies between 0 and 8000 hertz were kept. This band represented the frequencies expressed by human speaking voice.
B. The frequencies over 5000 hertz were trimmed out, and only the frequencies between 0 and 5000 hertz were kept. This band only contained the frequencies represented by the formants.
C. The frequencies over 8000 hertz and under 500 hertz were trimmed out, and only the frequencies between 500 and 8000 hertz were kept. This band represented the frequencies expressed by human speech voice but lacked fundamental frequency (F0).

Band A was used for training the model (i.e. frequencies between 0 and 8000).

Band A represented human speaking voice in almost all of the cases even when the speaker has extremely high-pitched voice or when certain vocalizations contained high pitched sound signals. Band A, then, has the most representative frequency range for human speech sound.

All three frequencies (A, B, C) were used for the validation of the model, and they were also presented to the human judges for human validation.

3.2.7 Data Validation

10% of the all recorded vocalizations were selected as the validation sample. The selected validation sample was used in two different validation processes. The neural network validation process and the human validation process used the same validation data set. Because both the neural network and the human raters used the same validation data set, it became possible to compare the performances of the neural network and the human raters.

Three judges were recruited to evaluate the collected vocalizations. The judges had a BA degree in psychology. Two of them also had a MSc degree in psychology, and one was a graduate student in a psychology MSc program. Each judge rated one spectrogram band. Each judge was presented with the vocalizations with the name of the vocalization changed from ID_EMOTION_WORD to YXXXXX where X and Y are figures, and Y is greater than zero. Each judge listened to the validation data set independent of other judges. The judges were allowed to listen to the vocalizations as many times as they would like. Then they rated the vocalizations into one of the emotional categories angry, calm, happy, and sad.

3.3 The Machine Learning Experiment

3.3.1 Overview

The preprocessed recordings were fed to a spectrogram producing a script written in Python 3.6.8 using SciPy libraries Numpy, Pandas, and Matplotlib (Hunter, 2007; McKinney, 2010; Oliphant, 2007; Van Der Walt, Colbert, & Varoquaux, 2011). The script produced a specifically tailored spectrogram that has optimum spatial and temporal resolution using short time Fourier transformation (Owens & Murphy, 1988) with highly overlapping windows.

The spectrogram files were split into training and validation sets by using a random sampling method through a pseudo-random algorithm. 80% of the data was used for the training set and 20% of the data was used for the validation set. For every ten spectrogram files, 8 training and 2 validation samples were selected. Then, a separate six-fold cross-validation was conducted over the whole dataset in order to validate the main model.

The neural network model belongs to the convolution neural network architecture. This architecture is mainly used for computer vision tasks. The general architecture of our model consisted of 4 convolutional layers, one dense layer, and one softmax layer. The model was trained for 9 epochs using the training samples for fitting and the validation samples for validation. The resulting model was saved by the h5 specifications.

The convolutional neural network (CNN) model was loaded back into the memory, and the last convolutional layer was copied as well as the output of the softmax layer. Gradients of the last convolutional layer were calculated according to the output of the softmax layer, and a heat map was generated. This heat map was applied to the original spectrogram image to make a comparative analysis of activations.

3.3.1.1 What is a spectrogram?

A spectrogram is a representation of the signal on time domain. In a spectrogram, the signal is represented as the change of amplitude in different frequencies over time (Pacific Northwest Seismic Network, n.d.). In the thesis, the X-axis represents time, whereas the Y-axis represents frequency. Change in the luminosity represents the change in the amplitude, the colour closer to white means the sound is higher in amplitude. The spectrogram image presented at Figure 2 demonstrates the change in amplitude and frequencies in time.

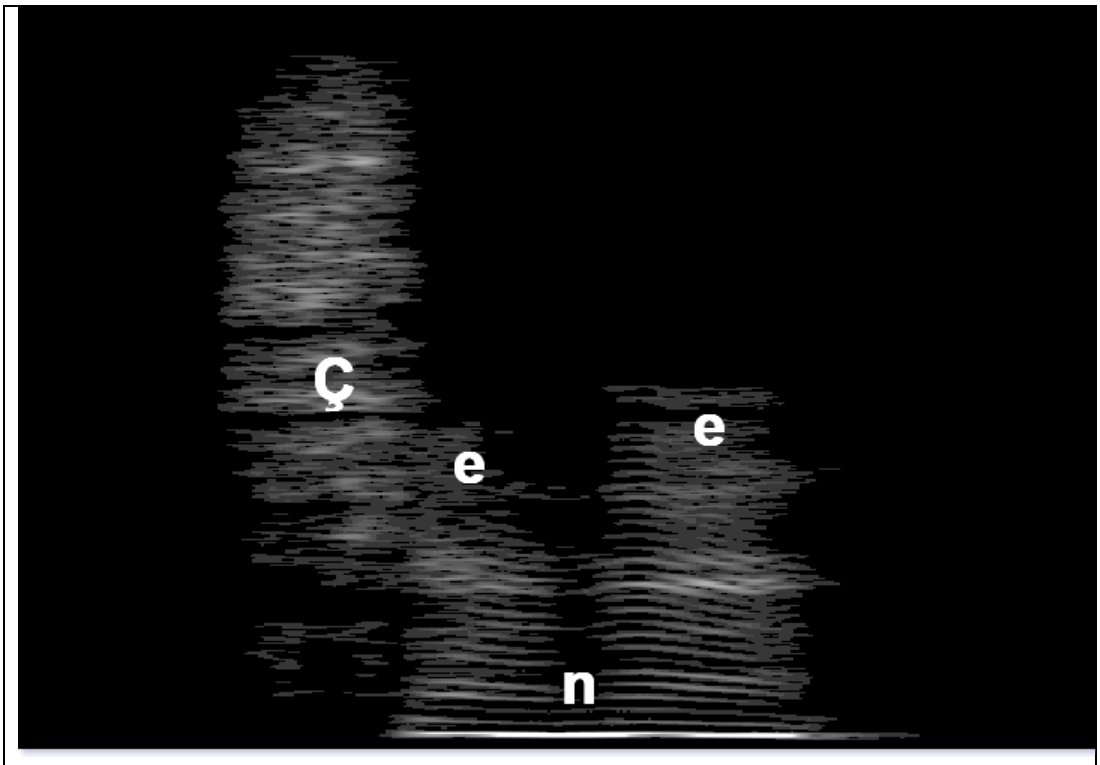

Figure 2: Spectrogram of the vocalization of the word *çene* in emotional state angry

3.3.1.2 What is a heat map?

A heat map is a specialized type of a shaded matrix; however, it can have more than two dimensions. In a heat map, the data is represented by color or luminosity

(Wilkinson & Friendly, 2009). Spectrograms are also heat maps in this context; however, within the boundaries of the thesis, heat maps represent the gradient activations in the spectrograms. In Figure 3 a demonstration of a heat map is presented for the vocalization of the word *çene* in the emotional state angry. The redder areas in the heat map represents the strong activations the model created.

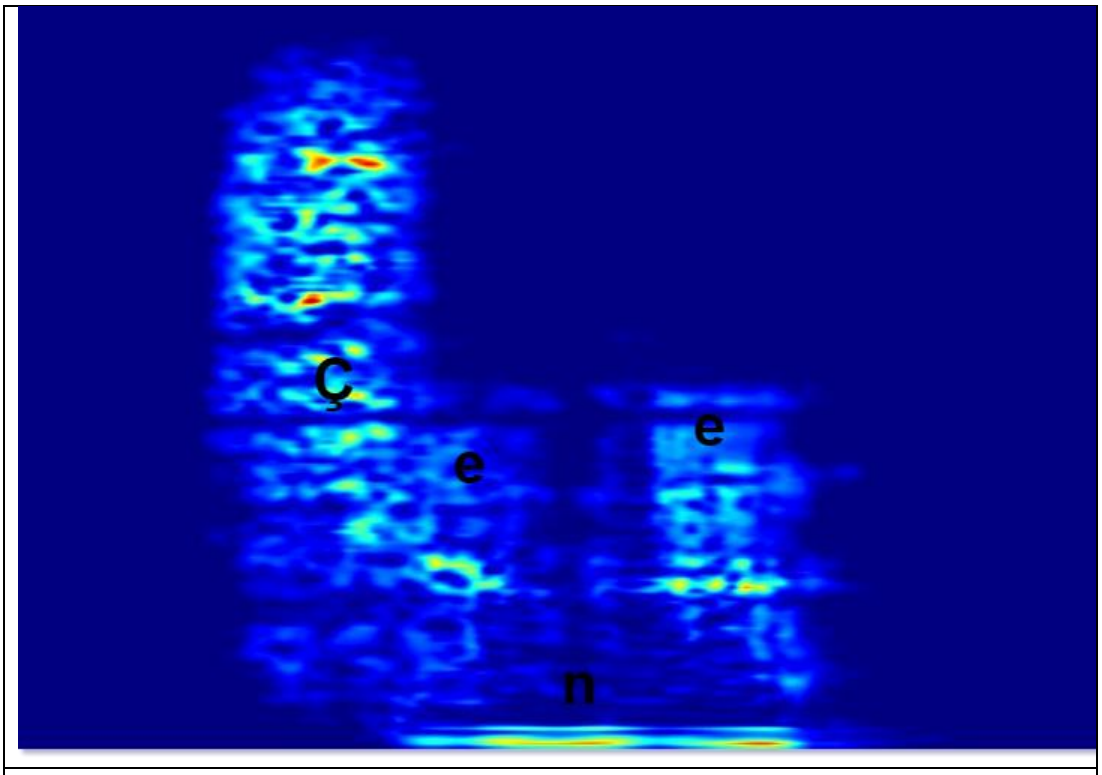

Figure 3: Heat map generated from the vocalization of the word *çene* in emotional state angry

3.3.2 Python and Its Libraries

Python is a high-level, general-purpose language which has been used for various purposes from data processing to web design. Python is currently in its third iteration. Python and its modules have been extensively used in the thesis from the data processing to the neural network generation and training. All of the Python libraries and the tools that had been used in the thesis are freely available for public use. Anaconda version of the Conda package manager was used for organizing Python and its libraries (Anaconda, 2014). Interfaces used in the thesis were Jupyter Notebook and Spyder (DataCamp, 2016; Spyder, 2018). Jupyter Notebook and Spyder were used as an interface for the Python interpreter. All of the interfaces were used through Anaconda Navigator. These interfaces can be separately installed and run from other package managers. However, the author has chosen the Conda package manager for its ease of use.

SciPy and packages under the SciPy umbrella were extensively used in the study. These packages are Numpy, Pandas, and Matplotlib. The Convolutional Neural Network (CNN) architecture was built using Tensorflow as backhand and Keras as a frontend (Chollet François, 2015; GoogleResearch, 2015). cuDNN and libraries under

its umbrella had worked with a backhand to allow training on GPU (C, 2010). Lastly, OpenCV was used for manipulating the heat maps (OpenCV, 2010).

3.3.3 Preprocessing and Spectrogram Generation

The recordings in 32-bit Float PCM WAV format were passed to a script designed to generate spectrograms specifically tailored for requirements of neural networks. The loaded recordings were subjected to a low-pass filter at 8000 hertz then transformed using the short time Fourier transformation. The resulting data was saved as images in PNG format. In Figure 2 one such example of a spectrogram image is presented.

The low-pass filter was designed as a digital Butterworth filter with an order of 8 and 8000-hertz cut point. The recordings were not down-sampled in order not to decrease Nyquist frequency, which is half of the sampling rate (Yao, 2014). This decision was made to have the highest possible resolution in the sound signal. The resulting signal did not contain any noise on frequencies 8000 and higher due to the order to the filter and its cut point.

The short time Fourier transformation (STFT) was applied to the filtered sound with the Hamming window as its window type (Podder, Zaman Khan, Haque Khan, & Muktadir Rahman, 2014). Each window was 2205 samples long with a sampling frequency of 44100 Hertz. Overlapping points were kept in 2100 to achieve stronger resolution in the expense of computational power. The effect of sample size and windowing on spectrograms is presented in Figure 4.

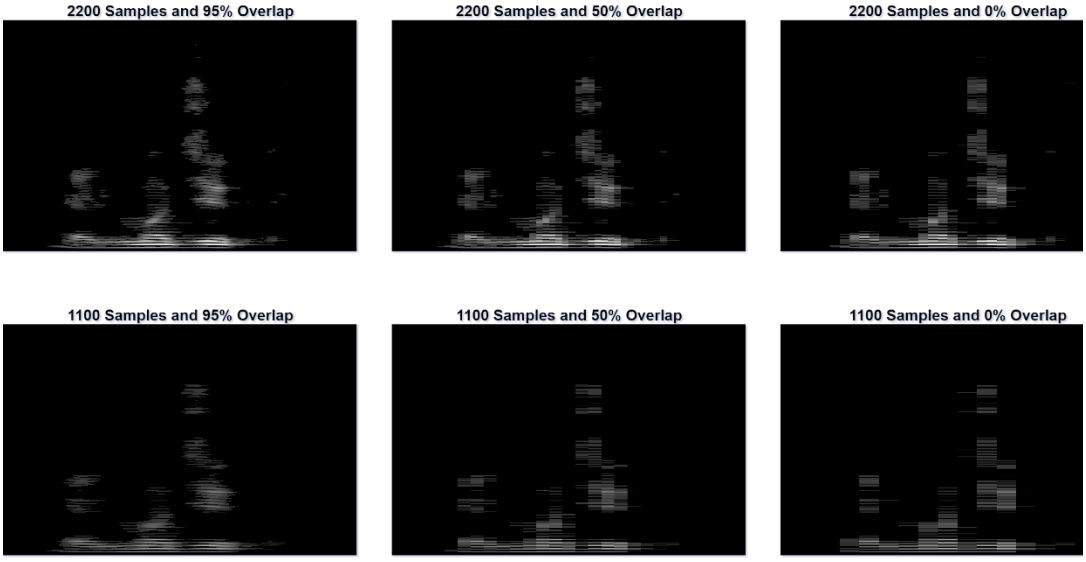

Figure 4: The effect of sample size and windowing on spectrograms

The resulting array was used to generate a pseudo color plot using Matplotlib. However, the grayscale color map was used while mapping colors to save final spectrogram in grayscale. The spectrogram was configured to have three by two aspect ratio with 320 dots per inch (dpi) resolution, and its frequency axis (y-axis) was cut at 8000-hertz mark. The final spectrogram was saved to the local disc in PNG format.

3.3.4 Convolutional Neural Network Model

3.3.4.1 Comparison Between The CNN and The SVM Models

The CNN architecture was selected for the machine learning experiment instead of the SVM architecture. Firstly, although the SVM architecture produces high accuracy scores and it is generally faster than the CNN architecture, it requires hand tailored features, such as F_0, pitch and energy values of different bands (Alghowinem et al., 2016). Hand crafted features allow fast computation but they can't be used for exploratory purposes other than brute force approaches. On the other hand, the CNN model can be exploited for its self-feature-extracting properties (Trigeorgis et al., 2016). Secondly, an ablation study similar to the one conducted in this thesis would either not produce any results or would not affect the SVM model. For instance, an SVM model that uses Max F_0 which resides in the 0-500 hertz band as its feature would not produce any results at the 500-8000 hertz frequency band and would not be affected at all by the removal of the 5000-8000 hertz band because it does not use any of the frequencies over 500 hertz. Moreover, the CNN architecture produces heat maps, which are essential components of the TurEV Database that could not be produced by an SVM model. Lastly, the CNN model is more robust to the noise and changes in the sound signal than an SVM model. The CNN model uses an image, the spectrogram, as input and performs an object classification task on it. This operation causes the CNN to be more robust.

3.3.4.2 Convolutional Neural Network Architecture

The convolutional neural network consisted of one input layer, one softmax output layer, four convolutional layers, four dropout layers, two max-pooling layers, one flatting layer, and a dense layer.

Input Layer

The input layer is technically not a layer; its primary role is to hold tensors and transfer them to the first layer in the model.

The input layer was configured to accept grayscale images. It had the input shape of 400x600 that can accept images 600 pixels wide and 400 pixels high and flexible in batch size.

Convolutional Layer

The convolutional layer is the central part of neural network architecture; convolution is defined as a continuous dot product of one function over another, which expresses

the similarity between them (M. Cohen, 2014; Owens & Murphy, 1988). From the neural network perspective, a convolutional layer takes a signal and subjects it to the layer's kernel and passes the result to the next layer.

Convolutional layers all had a 3x3 kernel size and rectified linear unit (ReLU) activation function.

Dropout Layer

The dropout layer partially sets the input of the previous layer to 0, effectively disabling set tensors for that data pass. It lowers the chance of overfitting.

Maxpooling Layer

The maxpooling layer down samples previous layer considering all of the values in its pool size then transcribes the maximum value to the next layer. A maxpooling layer with pool size four checks for 4x4 (in 2-dimensional data) and transcribes the maximum value to the next layer. A maxpooling layer effectively reduces the previous layer it is applied by a factor of maxpooling layer's pool size.

Flatten Layer

The Flatten layer flattens the previous layer to 1 dimension. However, it does not effect the batch size. A convolutional layer with 5x5x3 shape will turn into 75.

Dense Layer

The dense layer is a standard layer type of neural networks.

Softmax Layer

The softmax layer is a dense layer with softmax activation. It normalizes outputs between 1 and 0, and the sum of all outputs always equal to 1.

3.3.4.3 Activations in a Neural Network

In a neural network, the data flow from one node to the other. Each input data goes through a series of weights then it goes through an activation function. The result of this activation function is called *activation* and the result of the activations in a layer is called *layer activation*. It is also possible to collect the activations of a filter or the result of convolution operation performed by a filter. If activations are collected with respect to a category, it is called the *gradient activation*. Neural networks in general enable users to observe these activations and collect them but in most cases these activations do not carry any meaning. Convolutional Neural Networks, on the other hand, result in meaningful activations that show the parts of image where the model had "focused on" or the features which the model uses.

3.3.4.4 Convolutional Neural Network Build

The final architecture of the neural network was an input layer accepting 600x400 pixel sized grayscale images then three pairs of 2-dimensional convolutional layers

and dropout layers. Each convolutional layer had 3x3 kernel size, and each dropout layer had 0.25 probability. Filter sizes were 32, 64, and 64, respectively. This body of pairs followed by a max pooling layer with pooling factor of 2 then followed by another pair of 2-dimensional convolutional layers and a dropout layer. This was the last pair in the architecture and heat maps were generated from this layer. Kernel size and drop out probability did not change but the number of filters were increased to 256 in order to capture as many latent features as possible. This pair was followed by a max pooling layer with pooling factor of 8, a flattening layer and a dense layer with 256 inputs. The dense layer was followed by softmax activation layer with four outputs, one for each emotional category.

All layers that require an activation function in the model used ReLU activation. The only exception was the softmax layer, which used a softmax activation to categorize the result of the network. Resulting network architecture contained 56827844 parameters, all of which were trainable.

The model was compiled using categorical cross-entropy as its loss function. Optimizer that was used for the model was Adam optimizer with a learning rate of 0.0001. The accuracy rating was set to be recorded in the training history, but it wasn't used for training the parameters. The parameter training was handled by categorical cross-entropy loss function and the Adam optimizer. The overall architecture of the model is presented in Figure 5.

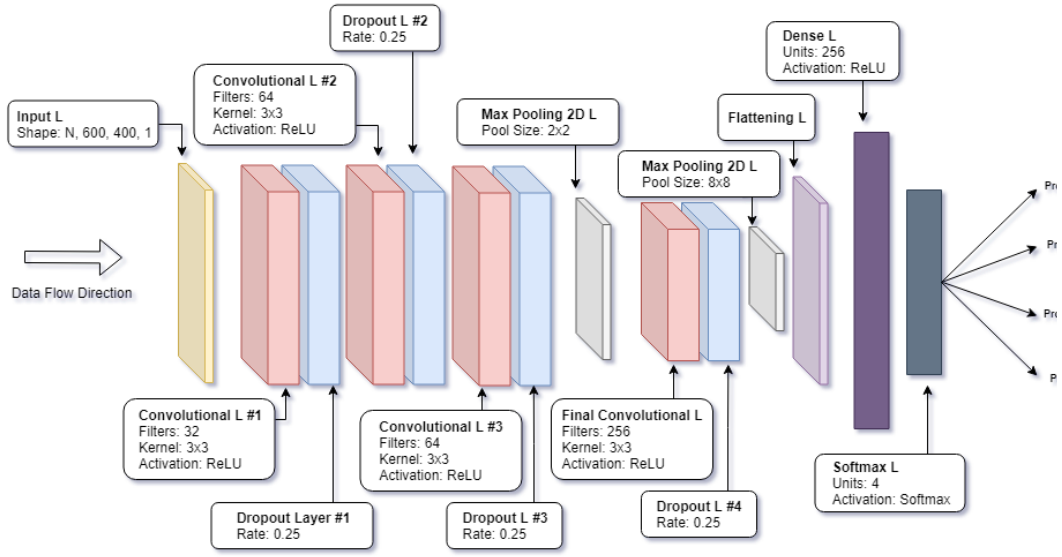

Figure 5: Overall architecture of the CNN model

3.3.5 Train and Validation Data Split

The spectrogram data was split using a 0.20 ratio between the training set and the validation set. 80% of the data was used for training the model, and 20% of data was

used for validation purposes. The process of splitting included adding randomly generated 6-digit number to the beginning of the file name. The validation set was generated from this pool. This procedure created a pool of files with random file order. Generally, due to the data collection guidelines, each data was saved in ID_EMOTION_WORD format. After shuffling, the order of words and IDs changed. The change allowed the model to be trained without being subject to order effect. This process had practically allowed each batch to consist of random words from random actors.

3.3.6 Six-Fold Cross-Validation

The spectrogram data was shuffled within emotion categories using the shuffling method introduced in Section 3.3.5. The shuffled data was subjected to a six-fold cross-validation. Each fold was trained for 9 iterations on a newly initialized model. The mean accuracy (72.5%) and the mean of maximum accuracy (74.9%) were calculated. The results were within the boundaries of the original model (see 4.2.1.1). Thus, the original model (cf Section 3.3.4.4) is considered valid. Accuracy values of the six-fold cross-validation study is presented in Table 2.

Table 2: Accuracy values for the cross-validation study

Fold Number	**Accuracy**	
	Maximum	**9th Iteration**
1	0.70%	0.65%
2	0.74%	0.71%
3	0.75%	0.72%
4	0.77%	0.77%
5	0.75%	0.72%
6	0.79%	0.78%

3.3.7 The Data Flow

The internal image data generation module of Keras was used for the model dataflow procedure. The spectrograms were loaded into memory through directories; two main data generators were used for loading the spectrograms. These data generators were the training data generator and the validation data generator. Mainly both the training and the validation data generators used the same parameters for loading spectrograms as grayscale images, rescaling the data format from 0-255 to 0-1 scale, then resizing the images to 600x400 pixels. Batch sizes for these models were 4 for the training generator and 1 for the validation generator. The training data generator loaded the images from the training folder, whereas the validation generator loaded the images from the validation folder.

The training and the validation generators primarily generated data using the same module; however, the data loaded by the validation generator was not used in training the parameters and did not affect the training regime in any way. It was used to validate the training regime and measure its ability through the change in loss and accuracy. On overview of dataflow is presented at Figure 6.

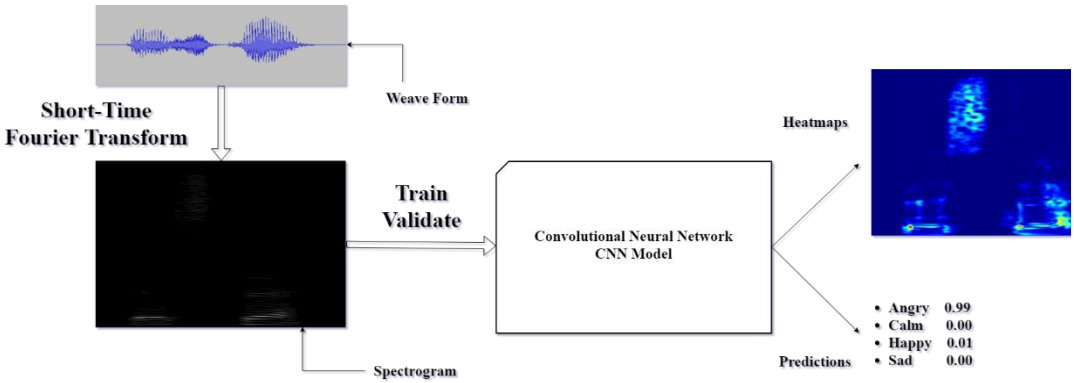

Figure 6: An overview of data flow

3.3.8 Model Training

The model was trained for 9 epochs using the data loaded by the trained generator, the validation data was also fed to the model at this stage, and the validation accuracy and loss were recorded. The resulting accuracy on the training data was 93%, and validation accuracy was 76%. Chance accuracy was 25%, which indicates that the model had performed more than three times better than the chance accuracy.

The training history was saved in CSV format, whereas the model was saved using the h5 format. The resulting model was later used for the analysis and for generating the heat maps.

3.3.9 Heat Map Generation

The heat map generation followed the steps given below.

1- The source image was loaded into the Python environment.
   a. The image was loaded in the color format with red, green blue (RGB) layers containing the same information.
   b. The image was loaded in the grayscale format with no RGB layers. This grayscale image was used for the generation of the heat map in the model.
2- The model was loaded into the Python environment.
3- The last convolutional layer was extracted from the model.
4- The model output relative to the target image was extracted from the model.
5- A new model was produced using the last convolutional layer, and the model output.
   a. Gradients from the model output and the last convolutional layer were extracted.
   b. Average of these gradients calculated using the mean method.

    c. The new model was produced using the original model's input and the pooled gradients as the input, and the output of the last convolutional layer as the output.
6- The image loaded at 1b was fed to this model, and the outputs were collected.
    a. The last convolutional layer's output was collected; it will be named convolutional output for the simplicity sake.
    b. The pooled gradient's output was collected; it will be named gradient output for the simplicity sake.
7- Each filter of the convolutional output was multiplied with the gradient output.
8- The average of the convolutional output was calculated using the mean method.

To generate the heat maps, the target image and the model were loaded into the working environment. A new model was created using the outputs relative to the target image and the last convolutional layer. The image was fed to this new model, and the gradient activations were collected from the last convolutional layer. These activations were resized to the original size of the target image, and the pseudo color was applied to it using the jet colormap. The Heat map's alpha value was lowered to 40%, and then the heat map was applied on the original image creating a mixed image where the gradient activations were visible over the spectrogram itself.

3.4 Statistical Analysis

The recordings that have been preprocessed and cleaned of noise were used for calculation of their F_0 values. The minimum, maximum, range, and standard deviation of F_0 values were calculated. These values were subjected to Bayesian ANOVA 4x (Emotional Category: Angry, Calm, Happy, Sad) with words themselves and gender used as a random factor.

Fundamental frequency (F_0) is the base formant and resides between 0 and 500 hertz ("Fundamental Frequency of Continuous Signals," 2011; Lemmetty & Sami, 1999). Fundamental frequency is the frequency range with highest amplitude that resides within 0-500 hertz. It defines the intonation contour and is one of the main features used in emotion-voice studies.

Due to the changes in frequency ranges in the speech signal, in this thesis, the data were divided into different frequency ranges up to formant five that reaches 5000 hertz. For more information on formants refer to 2.4.1.

3.5 Summary

In this chapter, we presented our methodology starting with our data processing method moving on to the machine learning experiment. We ended the chapter with the procedure of statistical analysis. In next chapter we present the results of the analyses described in this chapter.

Chapter IV

4. Analysis and Results

4.1 Introduction

In this chapter, we present and then analyze the results that were derived from the model, the judges, and compare the results of the model with the judges. The results will be presented in a series of analysis on the model, the judges, and the results of the comparative analysis. Each section will include the results derived from the analysis and an assessment.

4.2 Machine Learning Results

In this section, we present the machine learning model results. The subsections will contain the results of model training, model validation, and the classifications in contingency tables.

4.2.1 Results of Training

In this section, we present the results of the model through accuracy and loss. The results will be presented both for the training set and the validation set (information regarding the training and the validation sets were presented in Section 3.3.5). Epoch numbers in the context of training represent one cycle of training, where every single data point in the training set is subjected to the model.

Training accuracy is calculated by dividing the hits by the sum of the hits and the misses. The formula of accuracy is presented below in Equation 4.1. Training loss is calculated by the formula presented in Equation 4.2. Unlike accuracy, it is also sensitive to the indecisions made by the model. For instance, in a categorization task with A and B categories, categorizing A as A with 0.51 confidence is penalized for 0.49 loss in confidence (indecision) in categorical crossentropy function where as it is not penalized in accuracy calculation.

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Equation 4.1: Accuracy

$$f(s)_i = \frac{e^{s_i}}{\sum_j^C e^{s_j}}, CE = -\sum_i^C ti \log(f(s)_i)$$

Equation 4.2: Categorical Crossentropy

4.2.1.1 Training Accuracy

The results presented below in Table 3 and Figure 7 indicate that the validation and the testing accuracy steadily increased. The training accuracy reached a plateau at epoch 7, whereas the validation accuracy showed a steady increase.

Table 3: Accuracy score for training and validation sets

	Set	
Epoch Number	**Training**	**Validation**
1	41.76%	42.41%
2	55.56%	58.17%
3	67.41%	64.18%
4	76.01%	67.34%
5	81.21%	69.05%
6	87.14%	70.49%
7	91.33%	71.06%
8	93.35%	72.21%
9	93.71%	76.36%

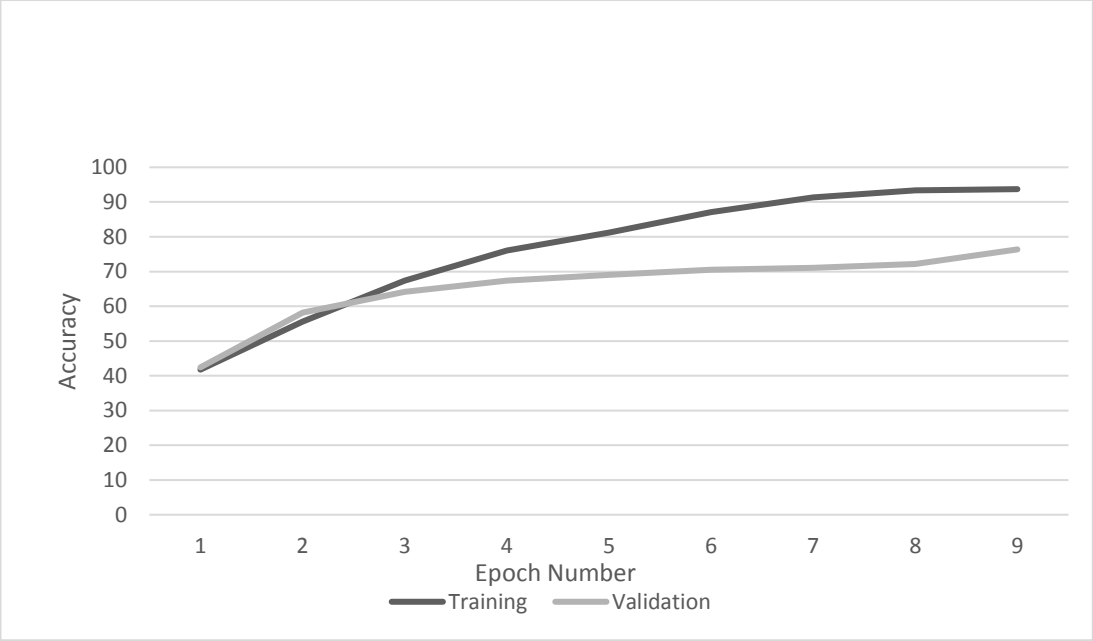

Figure 7: Accuracy score for training and validation sets

4.2.1.2 Training Loss

The results presented below in Table 4 and Figure 8 indicate that both the training and the validation loss function had a steady decrease. The small jump in the validation loss at the 9th epoch corresponded to the jump in the validation accuracy.

Table 4: Loss value during for training and validation sets

	Set	
Epoch Number	**Training**	**Validation**
1	1.2694	1.2828
2	1.0748	1.1738
3	0.8396	0.968
4	0.6298	0.8849
5	0.4737	0.8214
6	0.3555	0.7717
7	0.2679	0.7874
8	0.201	0.7588
9	0.1726	0.6649

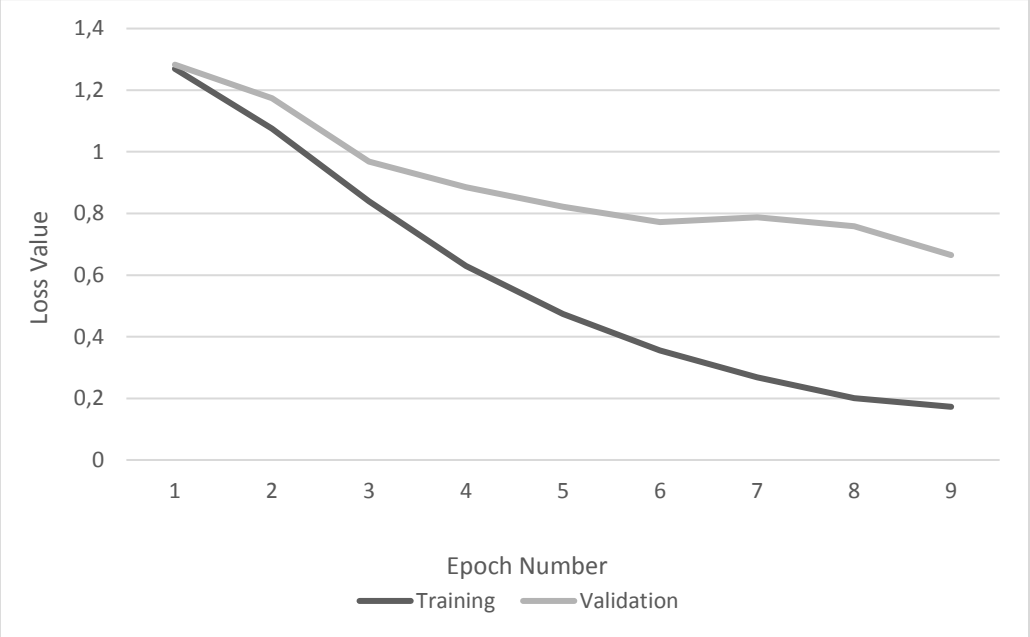

Figure 8: Loss value during for training and validation sets

4.2.1.3 Summary of Training Results

The steady decline in loss function and increase in accuracy for both training and validation sets indicate a healthy training regime. This result is tied to the early stopper function used in the training regime.

4.2.2 Results of the Validation Study

In this section, we present the classification scores in terms of precision, recall, and F1. We present these scores for three different frequency bands; 0-8000 hertz band, 0-5000 hertz band, and 500-8000 hertz band.

The precision score represents the classifier's (the model's or the judges') ability to avoid false positives (FP); on the other hand, the recall score represents the classifier's ability to avoid false negatives (FN). The F1 score is the compound metric that measures the performance of both precision and recall. However, the F1 score is not very sensitive and is the average of both metrics. For example, the precision metric can be represented as an email spam filter that prevents spams (TP) but tries not to mark any non-spam mail as spam mail (FP). The recall metric, on the other hand, is like a medical test, tries to find if there is any ailment (TP) even if it means marking non-ill cases as ill (FP).

Computations of the scores are presented in Equation 4.3, Equation 4.4, and Equation 4.5.

$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

Equation 4.3: Precision

$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

Equation 4.4: Recall

$$\frac{2*(\mathrm{Precision}*\ \mathrm{Recall})}{\mathrm{Precision}+\mathrm{Recall}}$$

Equation 4.5: F1

4.2.2.1 0-8000 Hertz Band (Full Spectrum)

The results presented below in Table 5 and Figure 9 indicate that the category angry has the highest classification scores in precision, recall, and the F1 metric. The category calm, on the other hand, has the lowest classification scores in precision, recall, and the F1 metric. Overall, the category sad has the least amount of variation in the classification scores.

Table 5: Classification metrics for 0-8000 hertz frequency band

Emotional Category	Classification Metric		
	Precision	**Recall**	**F1**
Angry	.84	.76	.80
Calm	.69	.74	.71
Happy	.71	.79	.75
Sad	.79	.75	.77

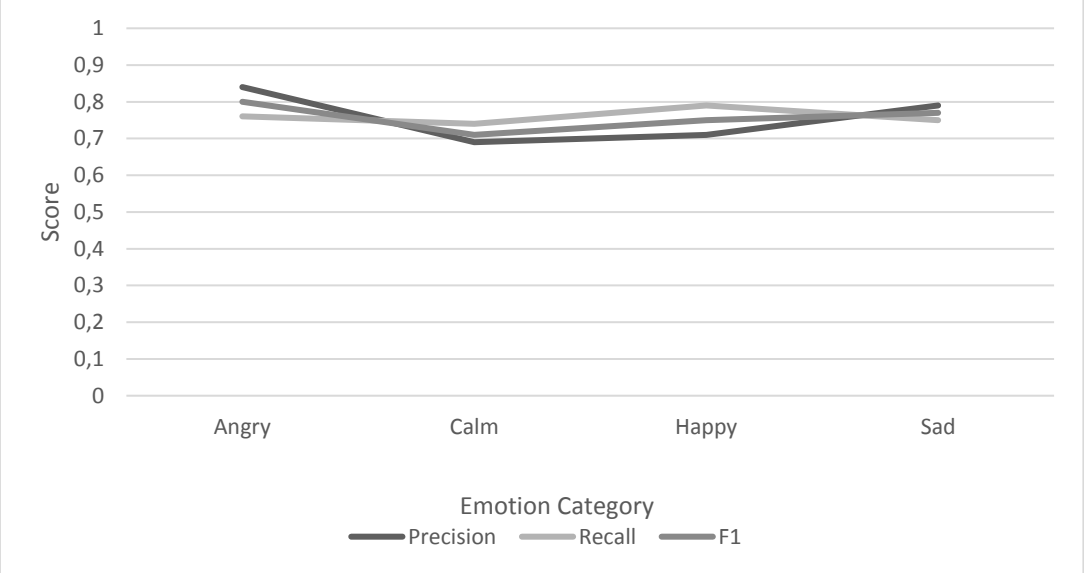

Figure 9: Classification metrics for 0-8000 hertz frequency band

4.2.2.2 0-5000 Hertz Band (Formant Only Spectrum)

The results presented below in Table 6 and Figure 10 indicate that the category angry has the highest precision score, whereas it has the lowest recall score. The number of false negatives made in the categorization of the category angry increased significantly. Overall, happy is effected most by the loss presented by the 0-5000 hertz frequency band and the category angry has lost most of its recall score.

Table 6: Classification metrics for 0-5000 hertz frequency band

Emotional Category	Classification Metric		
	Precision	**Recall**	**F1**
Angry	.93	.51	.66
Calm	.67	.73	.70
Happy	.60	.85	.71
Sad	.77	.82	.80

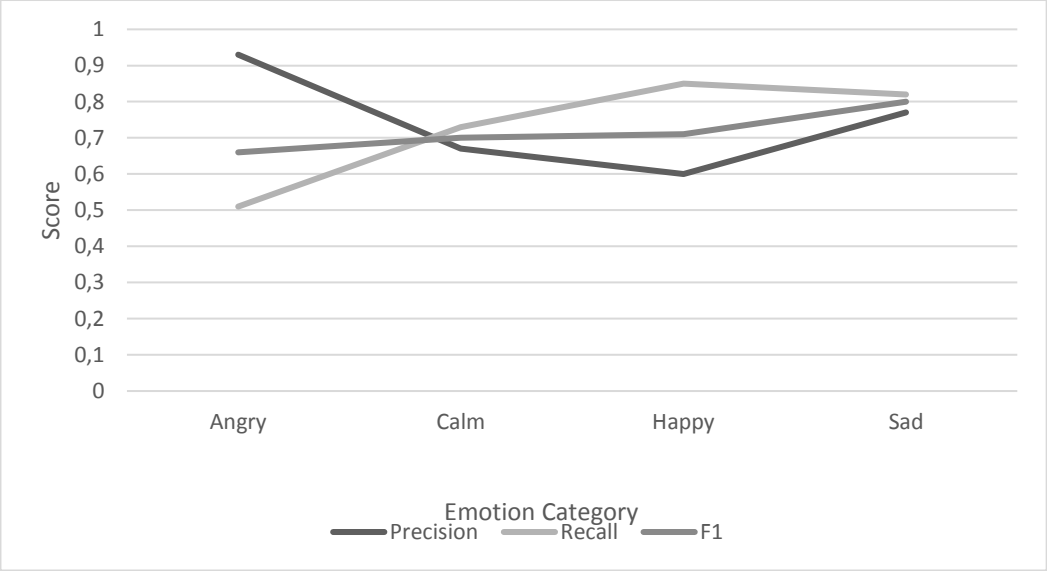

Figure 10: Classification metrics for 0-5000 hertz frequency band

4.2.2.3 500-8000 Hertz Band (Spectrum That Lacks Fundamental Frequency)

The results presented below in Table 7 and Figure 11 indicate that the precision score and the recall score of the category angry switched places with the 0-5000 hertz band. The category angry with high recall score and low precision score indicates that the number of false negatives has fallen, whereas the number of false positives has increased. The category calm has the lowest of all classification scores except for a relatively high precision score. In other words, in this frequency band, the model stopped differentiating between angry and other categories. This caused the model to flag other categories as angry.

Table 7: Classification metrics for 500-8000 hertz frequency band

Emotional Category	Classification Metric		
	Precision	**Recall**	**F1**
Angry	.46	.84	.60
Calm	.53	.11	.18
Happy	.41	.64	.50
Sad	.70	.31	.43

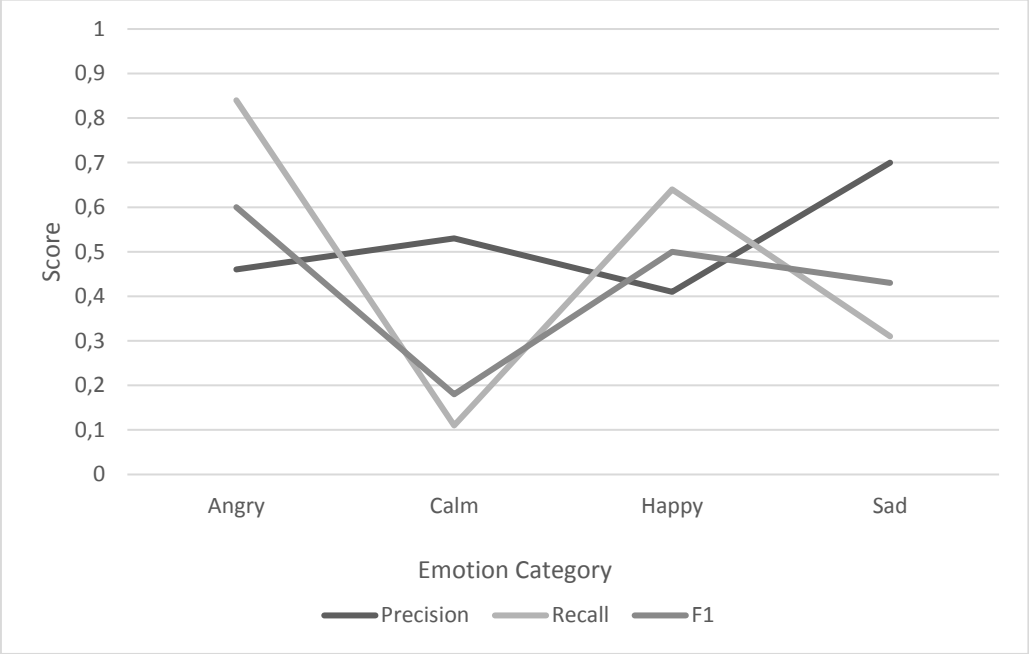

Figure 11: Classification metrics for 500-8000 hertz frequency band

4.2.3 Results of the Classifications for the Model in Contingency Tables

In this section, we present the classification results in contingency tables. A contingency table is a representation of the distribution of one variable within another. While the classification metrics offer a birds-eye view of the classification and the misclassification ratios, the contingency tables offer a detailed look into the exact category where the classifications and the misclassifications were made.

The contingency tables presented below will offer the number of matches between the emotional categories and present the percentile results of how much of the column is represented at the column-row intersection.

4.2.3.1 Contingency Tables for the 0-8000 Hertz Band

The results presented below in Table 8 indicate that the classifications and the misclassifications are mostly evenly distributed. Majority of misclassifications for the true category angry is made as calm with a 10.2% misclassification rate.

Table 8: Contingency tables for the 0-8000 hertz band

Model's Prediction		True Category: Angry	Calm	Happy	Sad
Angry	Count	**74.00**	5.00	4.00	5.00
	Percent	**75.5 %**	6.1 %	5.6 %	5.2 %
Calm	Count	10.00	**61.00**	6.00	12.00
	Percent	10.2 %	**74.4 %**	8.3 %	12.4 %
Happy	Count	9.00	7.00	**57.00**	7.00
	Percent	9.2 %	8.5 %	**79.2 %**	7.2 %
Sad	Count	5.00	9.00	5.00	**73.00**
	Percent	5.1 %	11.0 %	6.9 %	**75.3 %**

4.2.3.2 Contingency Tables for the 0-5000 Hertz Band

The results presented below in Table 9 indicate that classifications and the misclassifications are not evenly distributed; the majority of the misclassifications for the true category angry is made as to the category happy with a 27.6% misclassification rate.

Table 9: Contingency tables for the 0-5000 hertz band

Model's Prediction		True Category: Angry	Calm	Happy	Sad
Angry	Count	**50.00**	3.00	0.00	1.00
	Percent	**51.0 %**	3.7 %	0.0 %	1.0 %
Calm	Count	13.00	**60.00**	5.00	12.00
	Percent	13.3 %	**73.2 %**	6.9 %	12.4 %
Happy	Count	27.00	9.00	**61.00**	4.00
	Percent	27.6 %	11.0 %	**84.7 %**	4.1 %
Sad	Count	8.00	10.00	6.00	**80.00**
	Percent	8.2 %	12.2 %	8.3 %	**82.5 %**

4.2.3.3 Contingency Tables for the 500-8000 Hertz Band

The results presented below in Table 10 indicate that classifications and the misclassifications are not evenly distributed. The lowest categorization rate belongs to the category calm with 11% prediction rate. The highest categorization rate belongs to the category angry with 83.7% prediction rate. However, the following emotional categories were misclassified as the category angry; calm (42.7%), happy (27.8%), and sad (41.2%).

Table 10: Contingency tables for the 500-8000 hertz band

Model's Prediction		True Category			
		Angry	**Calm**	**Happy**	**Sad**
Angry	Count	**82.00**	35.00	20.00	40.00
	Percent	**83.7 %**	42.7 %	27.8 %	41.2 %
Calm	Count	1.00	**9.00**	3.00	4.00
	Percent	1.0 %	**11.0 %**	4.2 %	4.1 %
Happy	Count	14.00	29.00	**46.00**	23.00
	Percent	14.3 %	35.4 %	**63.9 %**	23.7 %
Sad	Count	1.00	9.00	3.00	**30.00**
	Percent	1.0 %	11.0 %	4.2 %	**30.9 %**

4.2.4 Assessment of the Machine Learning Results

So far, we have mainly concluded that;

1. The category angry has more false negatives in the 0-5000 hertz band, whereas it has more false positives in 500-8000 hertz band. The model is using 5000-8000 hertz band to flag the category angry whereas it uses 0-500 hertz band to flag other categories.
2. The category angry is mostly misclassified as the category happy in 27.6% of the cases in 0-5000 hertz band.
3. The category angry is misclassified as the category calm, happy, and sad in 42.7%, 27.8%, and 41.2% of the cases respectively in 500-8000 hertz band.
4. The category calm has the lowest categorization rating at 11.0% for 500-8000 hertz band.

Moreover, other conclusions are;

- The category angry has the highest classification scores for the 0-8000 hertz band.
- The category calm has mostly low classification scores for all three bands.
- The category calm has less than chance F1 score for the 500-8000 hertz band.
- The category angry is misclassified as the category calm for 10.2% in 0-8000 hertz band.
- The category angry is misclassified as the category happy for 27.6% in 0-5000 hertz band.

According to the findings, we can assume that the category angry has the highest interaction with other categories in classification. In the 0-5000 hertz band, where the category angry has shown to have high false negatives (finding 1), most of the vocalizations with the category angry is classified as the category happy (finding 2). This result indicates that the information in the 5000-8000 hertz range carries essential information for the model to differentiate between the category angry and category happy.

The model tends to classify the emotions as angry in the 500-8000 hertz band (finding 3). The category happy has the lowest misclassification as the category angry in the 500-8000 hertz band (finding 3). It can be assumed that the features that separate the category calm and sad reside in the 0-500 hertz range, whereas the category happy has some other informative features in higher frequencies.

The category calm has mostly low classification scores and has an accuracy of 11% for the 500-8000 hertz range (finding 4). It can be assumed that the features that define calm are not as strong as other emotions, and they reside within the 0-500 hertz range.

4.3 Results of the Judges

In this section, we present the results of the judges and then we asses these results. The results are presented under two main topics; the classification reports and the contingency tables.

4.3.1 Results of the Classifications for the Judges in Classification Reports

In this section, we present the classification scores of the judges in terms of precision, recall, and the F1. We present these scores for three different frequency bands; 0-8000 hertz band, 0-5000 hertz band, and 500-8000 hertz band.

4.3.1.1 0-8000 Hertz Band (Full Spectrum)

The results presented below in Table 11 and Figure 12 indicate that the category angry has the strongest and most stable results in the classification metrics precision, recall, and the F1. The categories happy and sad followed the trend of high precision and low recall scores; therefore, committing less false positive errors but more false negative errors.

Table 11: Classification Metrics for the 0-8000 hertz Frequency Band

Emotional Category	Classification Metric		
	Precision	Recall	F1
Angry	.74	.80	.77
Calm	.43	.68	.53
Happy	.80	.56	.66
Sad	.83	.54	.65

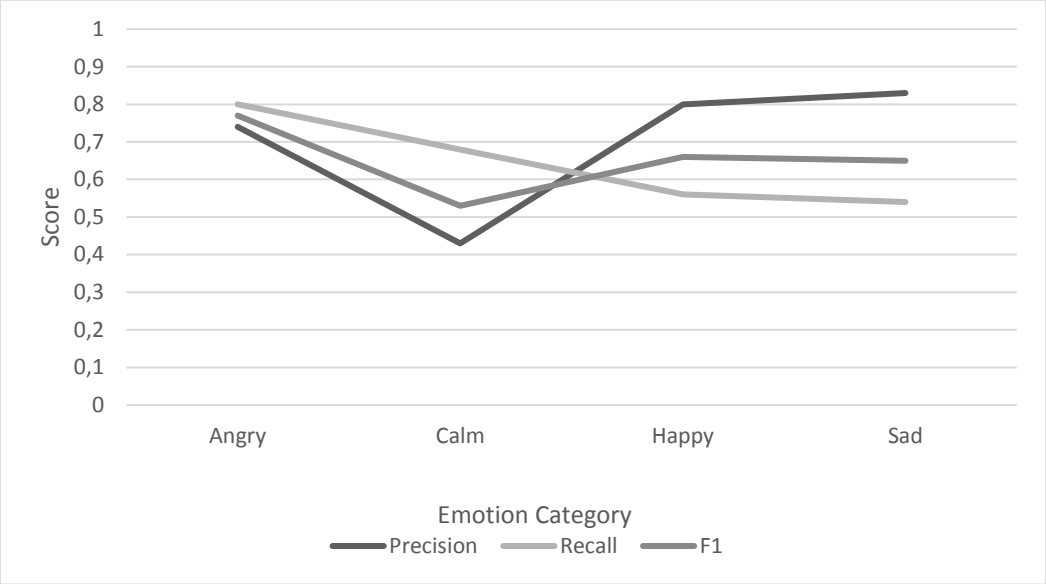

Figure 12: Classification metrics for the 0-8000 hertz frequency band

4.3.1.2 0-5000 Hertz Band (Formant Only Spectrum)

The results presented below in Table 12 and Figure 13Figure 13 indicate that the category angry has the highest precision and F1 classification scores. The category calm, on the other hand, has the lowest ratings in all three classification scores; precision, recall, and F1.

Table 12: Classification metrics for the 0-5000 hertz frequency band

Emotional Category	**Classification Metric**		
	Precision	**Recall**	**F1**
Angry	.90	.63	.74
Calm	.38	.60	.46
Happy	.84	.61	.70
Sad	.69	.71	.70

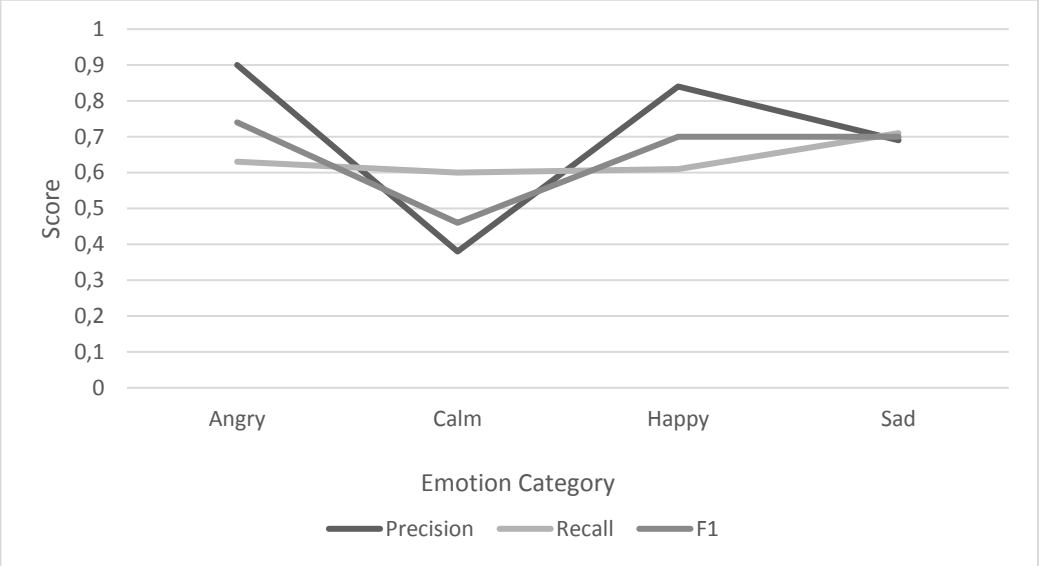

Figure 13: Classification metrics for the 0-5000 hertz frequency band

4.3.1.3 500-8000 Hertz Band (Spectrum That Lacks Fundamental Frequency)

The results presented below in Table 13 and Figure 14 indicate that the category happy has the highest precision score; however, it also has the lowest recall score. This pattern indicates a high false negative error rate. The category calm keeps the trends of having the lowest scores in all of the classification metrics.

Table 13: Classification metrics for 500-8000 hertz frequency band

Emotional Category	Classification Metric		
	Precision	**Recall**	**F1**
Angry	.71	.79	.74
Calm	.41	.57	.48
Happy	.92	.49	.64
Sad	.74	.67	.70

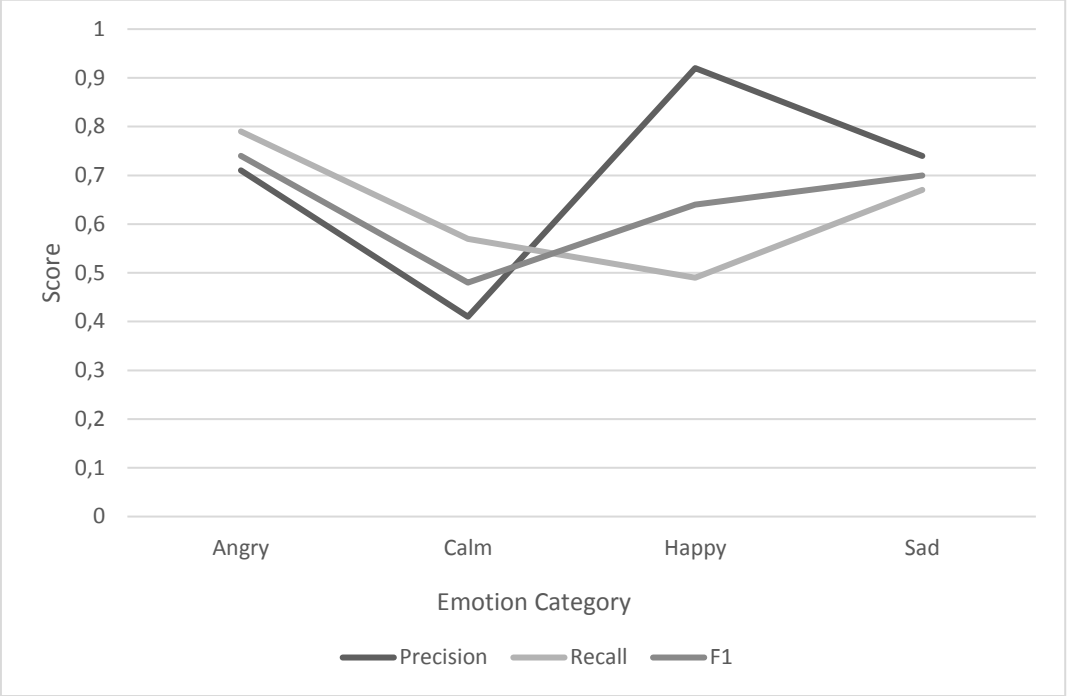

Figure 14: Classification metrics for 500-8000 hertz frequency band

4.3.2 Results of the Classifications for The Judges in Contingency Tables

In this section, we present the classification results in contingency tables. The contingency tables presented below will offer the number of matches between the emotional categories and present the percentile results of how much of the column is represented at the column-row intersection.

4.3.2.1 Contingency Tables for 0-8000 Hertz Band

The results presented below in Table 14 indicate that the category sad is misclassified 42.3% of the time as the category calm.

Table 14: Contingency tables for the 0-8000 hertz band

Judge's Prediction		True Category			
		Angry	**Calm**	**Happy**	**Sad**
Angry	Count	**78.00**	11.00	12.00	4.00
	Percent	**79.6 %**	13.4 %	16.7 %	4.1 %
Calm	Count	16.00	**56.00**	18.00	41.00
	Percent	16.3 %	**68.3 %**	25.0 %	42.3 %
Happy	Count	4.00	6.00	**40.00**	0.00
	Percent	4.1 %	7.3 %	**55.6 %**	0.0 %
Sad	Count	0.00	9.00	2.00	**52.00**
	Percent	0.0 %	11.0 %	2.8 %	**53.6 %**

4.3.2.2 Contingency Tables for 0-5000 Hertz Band

The results presented below in Table 15 indicate that the emotional category sad follows the trend of misclassification into the category calm with 40.2% of misclassification rate. Other notable misclassifications are: the emotional category calm is misclassified as the category angry (18.3%) and the emotional category angry is misclassified as the category calm (22.4%).

Table 15: Contingency tables for the 0-5000 hertz band

Judge's Prediction		True Category			
		Angry	**Calm**	**Happy**	**Sad**
Angry	Count	**73.00**	15.00	15.00	6.00
	Percent	**74.5 %**	18.3 %	20.8 %	6.2 %
Calm	Count	22.00	**48.00**	28.00	39.00
	Percent	22.4 %	**58.5 %**	38.9 %	40.2 %
Happy	Count	1.00	3.00	**29.00**	1.00
	Percent	1.0 %	3.7 %	**40.3 %**	1.0 %
Sad	Count	2.00	16.00	0.00	**51.00**
	Percent	2.0 %	19.5 %	0.0 %	**52.6 %**

4.3.2.3 Contingency Tables for 500-8000 Hertz Band

The results presented below in Table 16 indicate that although the majority of the misclassifications still exist, they are no longer notable.

Table 16: Contingency tables for the 500-8000 hertz band

Judge's Prediction		True Category			
		Angry	**Calm**	**Happy**	**Sad**
Angry	Count	**77.00**	16.00	12.00	4.00
	Percent	**78.6 %**	19.5 %	16.7 %	4.1 %
Calm	Count	17.00	**47.00**	22.00	28.00
	Percent	17.3 %	**57.3 %**	30.6 %	28.9 %
Happy	Count	1.00	2.00	**35.00**	0.00
	Percent	1.0 %	2.4 %	**48.6 %**	0.0 %
Sad	Count	3.00	17.00	3.00	**65.00**
	Percent	3.1 %	20.7 %	4.2 %	**67.0 %**

4.3.3 Assessment of the Judges' Results

So far, we have mainly concluded that;

1. The category angry has the most stable, and relatively high categorization results for the 0-8000 hertz band.
2. The categories happy and sad have high precision scores (80%, 83% respectively) and low recall scores (56%, 54% respectively) for the 0-8000 hertz band.
3. The category sad has been misclassified as the category calm for 42.3% in 0-8000 hertz band.
4. The category sad has been misclassified as the category calm 40.2% in the 0-5000 hertz band.
5. The category angry has been misclassified as the category calm for 22.4% in 0-5000 hertz band.
6. The category calm has the lowest F1 scores (53%, 46%, and 48%) for the frequency bands 0-8000, 0-5000, and 500-8000 hertz, respectively.

Moreover, the other conclusions are;

- The category angry has the highest precision score (90%), but it has a relatively lower recall score (63%) for the 0-5000 hertz band.
- The category happy has the highest precision score (92%), but it has the lowest recall score (49%) for the 500-8000 hertz band.
- The category angry and calm has relatively close precision and recall scores for the 500-8000 hertz band.

According to the findings, we can assume that the category angry is stable when all of the frequencies between 0 and 8000 hertz is presented to the judge (finding 1).

The judges have high false negatives (they have high precision scores and low recall scores) for the category happy and sad in 0-8000 hertz band (finding 2).

The judges misclassify the category of sad as calm at 42.3% in 0-8000 hertz band (finding 2) and 40.2% for 0-5000 hertz band (finding 3). It can be assumed that without external cues (i.e, visual cues), the judges rate vocalizations with low energy and low variability as the category calm. Energy and variation of a sound signal is explained at 2.4.1.

The category angry is misclassified as the category calm at 22.4% in the 0-5000 hertz band (finding 5), this misclassification was on the direction of the category happy for 27.6% in the same frequency band (4.2.2.2) for the model. It can be assumed that human judges used the frequency band 5000-8000 for the emotion of different energy profile (2.2.1) whereas the model used the same band to differentiate the emotion of the same energy profile.

The category calm has the lowest F1 ratings for all frequency ranges (finding 6). It can be assumed that the judges were classifying low-energy, low-variance vocalizations in the category calm as sad and vice-versa.

4.4 Comparative Results

In this section, we compare the results of the model with the judges. Comparisons will be presented in terms of accuracy, precision, and recall values. Also, contingency tables that compare the model and the judges will be presented. This section aims to investigate the parallelisms and the deviations between the model and the judges.

4.4.1 Comparative Results in Classification Reports

4.4.1.1 The Accuracy Rating

The results presented below in Table 17 and Figure 15 indicate that the accuracy rating of the judges is lower than the models on the 0-8000 hertz the 0-5000 hertz bands. However, the accuracy rating of the judges is higher than the model at the 500-8000 hertz band. Importantly, the accuracy rating of the judges is robust throughout the bands, whereas the model experiences a sharp drop at the 500-8000 hertz band.

Table 17: Accuracy rating for the model and the judges

	Frequency Bands			
	0-8000	**0-5000 hertz**	**500-8000 hertz**	
The Judges	0.65	0.64	0.64	**Accruacy Rating**
The Model	0.76	0.72	0.48	

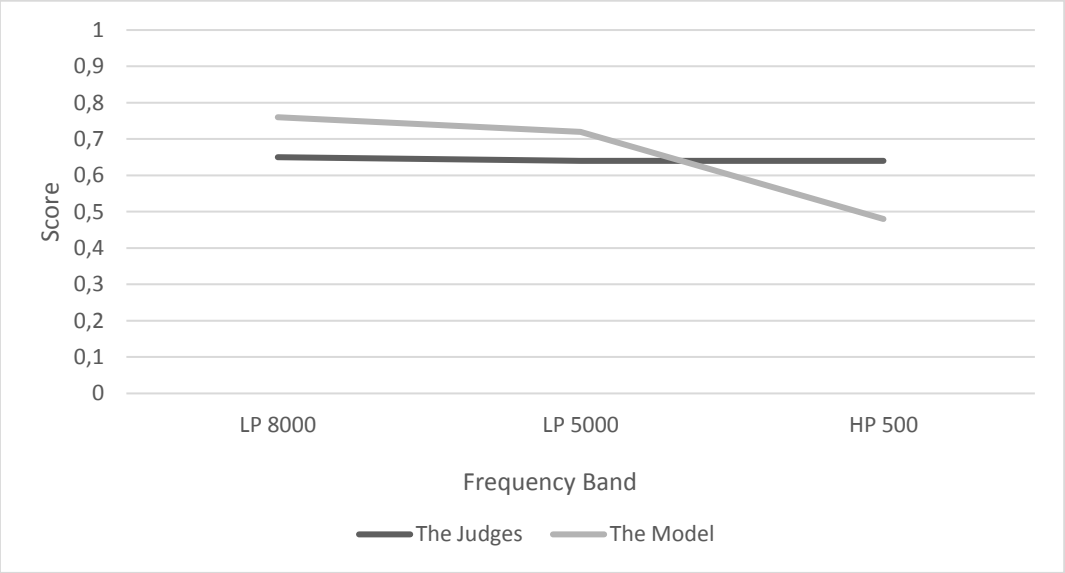

Figure 15: Accuracy rating for the model and the judges

4.4.1.2 Precision Scores of the Model and the Judges for Emotional Category Angry

The results presented below in Table 18 and Figure 16 indicate that the model and the judges follow the same trend in the 0-8000 hertz the 0-5000 hertz bands. However, the judges did not get effected as much as the model at the 500-8000 hertz band and managed to stay over 70% accuracy.

Table 18: Precision scores of the model and the judges for the category angry

	Frequency Bands			
	0-8000	**0-5000 hertz**	**500-8000 hertz**	
The Judges	0.74	0.90	0.71	**Precision Score**
The Model	0.84	0.93	0.46	

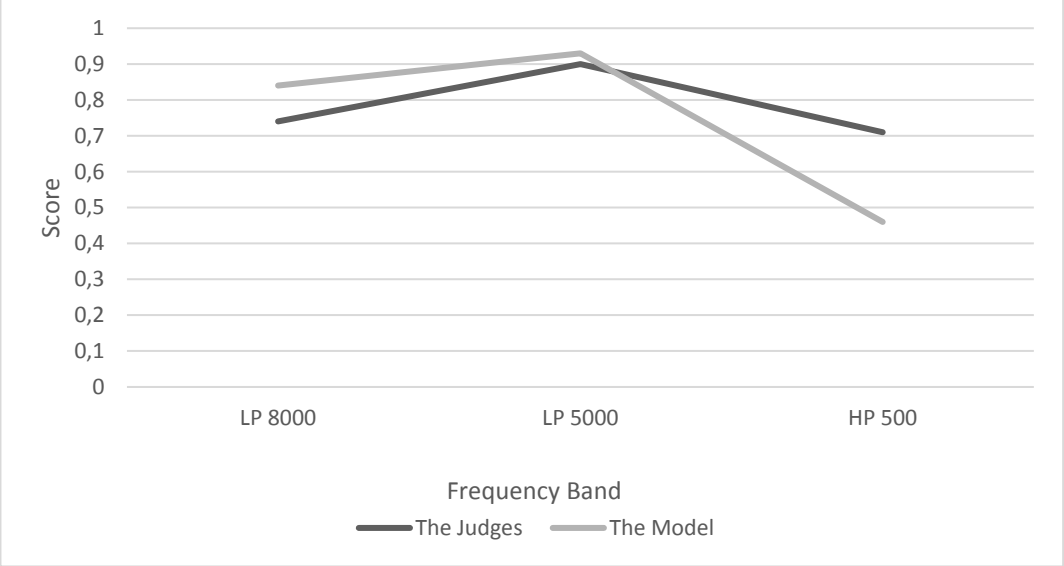

Figure 16: Precision scores of the model and the judges for the category angry

4.4.1.3 The Recall Scores of the Model and the Judges for Emotional Category Angry

The results presented below in Table 19 and Figure 17 and 4.4.1.2 in Table 18 and Figure 16 indicate that the discrepancy of the model's scores at 0-5000 hertz band and 500-8000 hertz band between the precision scores and the recall scores suggest a high level of false negatives.

Table 19: Recall scores of the model and the judges for the category angry

	Frequency Bands			
	0-8000 hertz	**0-5000 hertz**	**500-8000 hertz**	
The Judges	0.80	0.63	0.79	**Recall Score**
The Model	0.76	0.51	0.84	

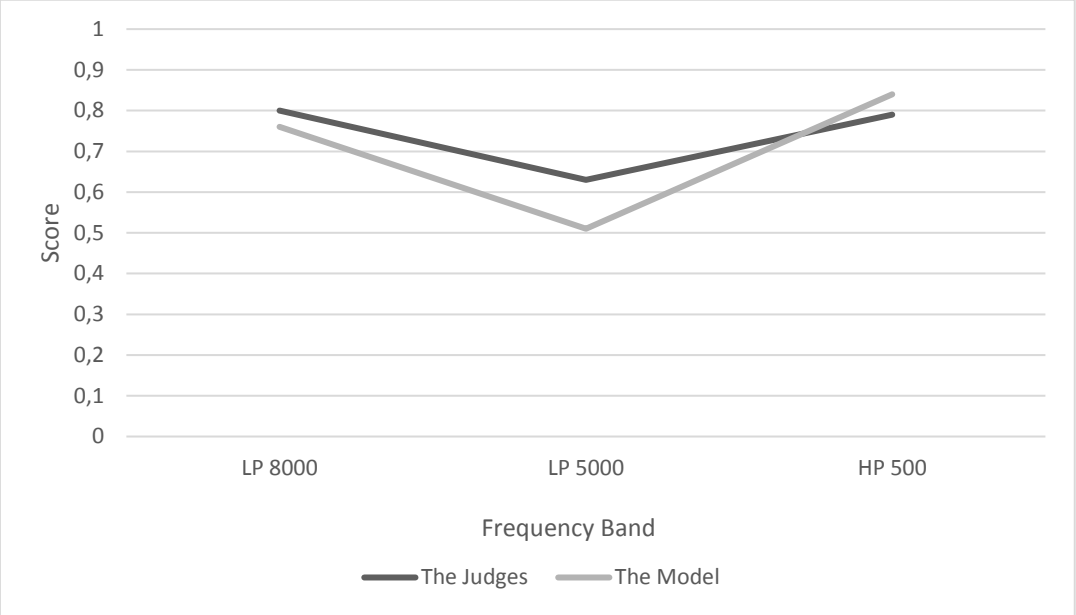

Figure 17: Recall scores of the model and the judges for the category angry

4.4.2 Results of the Classifications for The Comparative Analysis in Contingency Tables

In this section, we present the classification results in contingency tables.The contingency tables presented below will offer the number of the matches between the emotional categories and present the percentile results of how much of the column is represented at the column-row intersection.

4.4.2.1 Contingency Tables for 0-8000 Hertz Band

The results presented below in Table 21 and Table 20 indicate that the judges have agreed with the model more than half of the time (53.6%), the chance agreement is 25%. The judges have assessed single-word vocalizations more into the category angry than the model. For the category angry the model has a higher precision score of 84% (4.2.2.1) than the judges 74% (4.3.1.1) moreover the model has lower recall score (76%) than the judges (80%). The F1 scores indicate that the model with 80% F1 score is more successful in the classification of the category angry than the judges with 76% F1 score.

Table 20: Contingency tables for 0-8000 hertz band with the judges as the key

Model's Prediction		Judge's Prediction			
		Angry	**Calm**	**Happy**	**Sad**
Angry	Count	**65.00**	14.00	5.00	4.00
	Percent	**61.9 %**	10.7 %	10.0 %	6.3 %
Calm	Count	20.00	**50.00**	7.00	12.00
	Percent	19.0 %	**38.2 %**	14.0 %	19.0 %
Happy	Count	15.00	22.00	**34.00**	9.00
	Percent	14.3 %	16.8 %	**68.0 %**	14.3 %
Sad	Count	5.00	45.00	4.00	**38.00**
	Percent	4.8 %	34.4 %	8.0 %	**60.3 %**

Table 21: Contingency tables for 0-8000 hertz band with the model as the key

Judge's Prediction		Model's Prediction			
		Angry	**Calm**	**Happy**	**Sad**
Angry	Count	**65.00**	20.00	15.00	5.00
	Percent	**73.9 %**	22.5 %	18.8 %	5.4 %
Calm	Count	14.00	**50.00**	22.00	45.00
	Percent	15.9 %	**56.2 %**	27.5 %	48.9 %
Happy	Count	5.00	7.00	**34.00**	4.00
	Percent	5.7 %	7.9 %	**42.5 %**	4.3 %
Sad	Count	4.00	12.00	9.00	**38.00**
	Percent	4.5 %	13.5 %	11.3 %	**41.3 %**

4.4.2.2 Contingency Tables for 0-5000 Hertz Band

The results presented below in Table 22 and Table 23 indicate that the judges have agreed with the model less than half of the time (45.8%), the chance agreement is 25%. The classification trend of the category angry observed in 0-8000 hertz band is still existent, however in this frequency band the recall score of the category angry for the model (51%) is much lower than the judges' (63%) whereas their precision scores are very similar to the 93% for the model and 90% for the judges (the information can be found at 4.2.2.2 and 4.3.1.2). Majority of the model's predictions of the category happy are assessed as the category calm by the judges (40.6%) whereas the agreement on calm by the model is 48.9% and for the judges it is 32.1%. The majority of the judges' assessments of the emotional state calm is also predicted as the emotional state happy (29.9%) by the model. The agreement on the judges' assessments of the category calm is 32.1%, and the model's prediction of happy is 25.7%.

Table 22: Contingency tables for 0-5000 hertz band with the model as the key

Judge's Prediction		Model's Prediction			
		Angry	**Calm**	**Happy**	**Sad**
Angry	Count	**44.00**	25.00	29.00	11.00
	Percent	**81.5 %**	27.8 %	28.7 %	10.6 %
Calm	Count	7.00	**44.00**	41.00	45.00
	Percent	13.0 %	**48.9 %**	40.6 %	43.3 %
Happy	Count	1.00	5.00	**26.00**	2.00
	Percent	1.9 %	5.6 %	**25.7 %**	1.9 %
Sad	Count	2.00	16.00	5.00	**46.00**
	Percent	3.7 %	17.8 %	5.0 %	**44.2 %**

Table 23: Contingency tables for 0-5000 hertz band with the judges as the key

Model's Prediction		Judge's Prediction			
		Angry	**Calm**	**Happy**	**Sad**
Angry	Count	**44.00**	7.00	1.00	2.00
	Percent	**40.4 %**	5.1 %	2.9 %	2.9 %
Calm	Count	25.00	**44.00**	5.00	16.00
	Percent	22.9 %	**32.1 %**	14.7 %	23.2 %
Happy	Count	29.00	41.00	**26.00**	5.00
	Percent	26.6 %	29.9 %	**76.5 %**	7.2 %
Sad	Count	11.00	45.00	2.00	**46.00**
	Percent	10.1 %	32.8 %	5.9 %	**66.7 %**

4.4.2.3 Contingency Tables for 500-8000 Hertz Band

The results presented below in Table 24 and Table 25 indicate that the judges have agreed with the model less than half of the time (44.4%), the chance agreement is 25%. The pattern of a discrepancy between the judge's assessment of happy and model's prediction of calm remains in 0-5000 hertz band (4.4.2.2). The judges' assessment of the emotional state calm is predicted as the emotional state happy for 36.0% by the model. However, the judge no longer assesses the model's prediction of the emotional state calm as the emotional state happy.

Table 24: Contingency tables for 500-8000 hertz band with the model as the key

Judge's Prediction		Model's Prediction			
		Angry	**Calm**	**Happy**	**Sad**
Angry	Count	**82.00**	2.00	18.00	7.00
	Percent	**46.3 %**	11.8 %	16.1 %	16.3 %
Calm	Count	49.00	**11.00**	41.00	13.00
	Percent	27.7 %	**64.7 %**	36.6 %	30.2 %
Happy	Count	8.00	1.00	**27.00**	2.00
	Percent	4.5 %	5.9 %	**24.1 %**	4.7 %
Sad	Count	38.00	3.00	26.00	**21.00**
	Percent	21.5 %	17.6 %	23.2 %	**48.8 %**

Table 25: Contingency tables for 5000-8000 hertz band with the judges as the key

Model's Prediction		Judge's Prediction			
		Angry	**Calm**	**Happy**	**Sad**
Angry	Count	**82.00**	49.00	8.00	38.00
	Percent	**75.2 %**	43.0 %	21.1 %	43.2 %
Calm	Count	2.00	**11.00**	1.00	3.00
	Percent	1.8 %	**9.6 %**	2.6 %	3.4 %
Happy	Count	18.00	41.00	**27.00**	26.00
	Percent	16.5 %	36.0 %	**71.1 %**	29.5 %
Sad	Count	7.00	13.00	2.00	**21.00**
	Percent	6.4 %	11.4 %	5.3 %	**23.9 %**

4.4.3 Assessment of Comparative Results

So far, we mainly concluded that;

1. The accuracy rating for the model is higher for the 0-8000 hertz and the 0-5000 hertz band (76%, and 71% respectively). However, the accuracy rating sharply drops for the model at 500-8000 hertz band to 48% whereas the accuracy rating for the judges stays at 65%, 64%, and 64% respectively for 0-8000 hertz, 0-5000 hertz, and 500-8000 hertz band.
2. The agreement between the model and the judges is 53.6%, 45.8% and 44.4% for frequency bands 0-8000 hertz, 0-5000 hertz, and 500-8000 hertz bands, respectively. The chance agreement is 25%.
3. The judges displayed more false positives than the model for the category angry for 0-8000 hertz band.
4. The model displayed more false negatives than the model for the category angry for 0-5000 hertz band.
5. The model has high recall ratings for the categories happy (85%) and calm (73%) for 0-5000 hertz band.

6. The majority of the model's predictions of the category happy are assessed as the category calm at 40.6% by the judges for 0-5000 hertz band.
7. The judges' assessment of the category calm is predicted as the category happy at 36.0% by the model for 500-8000 hertz band.

Moreover, other conclusions are;

- The precision score for the category angry for the model and the judges have a weak correlation between different bands. However, the judges are more robust in this regard.
- The majority of the judges' predictions of the emotional state calm are assessed as the emotional state happy at 29.9% by the model for 0-5000 hertz band.

According to the findings, we can assume that although the model has stronger accuracy for the 0-8000 hertz and 0-5000 hertz bands, human judges are more robust to the changes in 500-8000 hertz (finding 1).

The agreement between the model and the judges is much higher than the chance agreement (finding 2). However, the model's and the judges' predictive power also affect this agreement lowering its significance.

The judges are more prone to making false positive errors for the category angry at 0-8000 hertz band (finding 3). The model has very high precision but very low recall for the emotional category angry at the 0-5000 hertz band (finding 4); therefore, the model is prone to making false negative errors at this frequency band. Although the model has high recall rating for the emotional states happy (85%) and calm (73%) (finding 5), the judges have assessed the category happy as the category calm for 40.6% for 0-5000 hertz band (finding 6). This can be assumed to be caused by the vocalizations lacking high pitched sounds that reside at 5000-8000 Hertz. This can be evidenced by this phenomenon disappearing at 500-8000 hertz band (finding 7).

4.5 Summary

In this chapter, we presented and investigated the results of the analysis. The results were derived from the model, the judges, and the comparison of both. The results were presented in terms of training results, classification reports, and contingency tables.

The main results of the study can be summarized as follows:

- The model uses the 5000-8000 hertz frequency band to differentiate between angry and happy, which are both high arousal emotions with high energy.
- The judges use the 5000-8000 hertz frequency band to differentiate between angry and calm, which shares neither valence nor arousal.
- The calm and sad have the lowest classification scores, it is probably because they both have low energy.

- The model has higher score than the judges at all frequency bands except the 500-8000 hertz.
- The model loses much of its predictive power at 500-8000 hertz band, i.e. the band that lacks F0. This result confirms the importance of F0 and shows why the contemporary literature uses F0 to recognize different emotion categories.
- The judges do not lose any classification power due to the changes in frequency band manipulations.

We can safely conclude that the model follows a mechanical approach of following the high energy signal at various frequency bands and using fundamental frequency as its guideline. But the judges follow a holistic approach making use of all frequency bands to distinguish between the four emotional categories. In particular, humans can compensate loses in all three main frequency bands used in the thesis with little to no loss in predictive power. The model, on the other hand, fails completely upon losing 0-500 hertz band and its features.

Chapter V

5. The Turkish Emotional Voice Database (TurEV Database)

5.1 Introduction

In this chapter, we present The Turkish Emotional Voice Database or TurEV Database. We first present the database then investigate the fundamental frequency (F0) statistics of the tokens in the database. Lastly, we evaluate the corpus through different statistical methods.

5.2 Database Coverage

In this section, we describe the database components; that is, the corpus and the peripheries. We present the corpus coverage. Then we present the peripheral components, including the results of audio statistics by F0 values and the demographic statistics on the amateur actors. Finally, we perform reliability analysis of the judges.

5.2.1 Corpus Coverage

Number of Unique Words

Eighty-two words have been selected from *Türkçe'nin Ses Dizgesi* (Ergenç & Bekar Uzun, 2017). As mentioned in section 3.1.2, these words were selected for their representativeness of the phonological properties of Turkish sounds. The list of selected words is presented at APPENDIX B.

Emotional Categories

Four emotional categories were chosen for the study; angry, calm, happy, and sad. These emotion categories were chosen because they are widely studied and they are easily distinguishable as well as producible (Busso et al., 2008; El Ayadi et al., 2011; Y. Kim et al., 2013; Liscombe et al., 2003; McGilloway et al., 2000). Moreover, these emotions reside in different axes of the arousal-valence space in which sad is negative and low arousal, calm is positive and low arousal, angry is negative and high arousal, and happy is positive and high arousal (Barrett, 1998). The position of emotions on the valence arousal axis is presented in Figure 18.

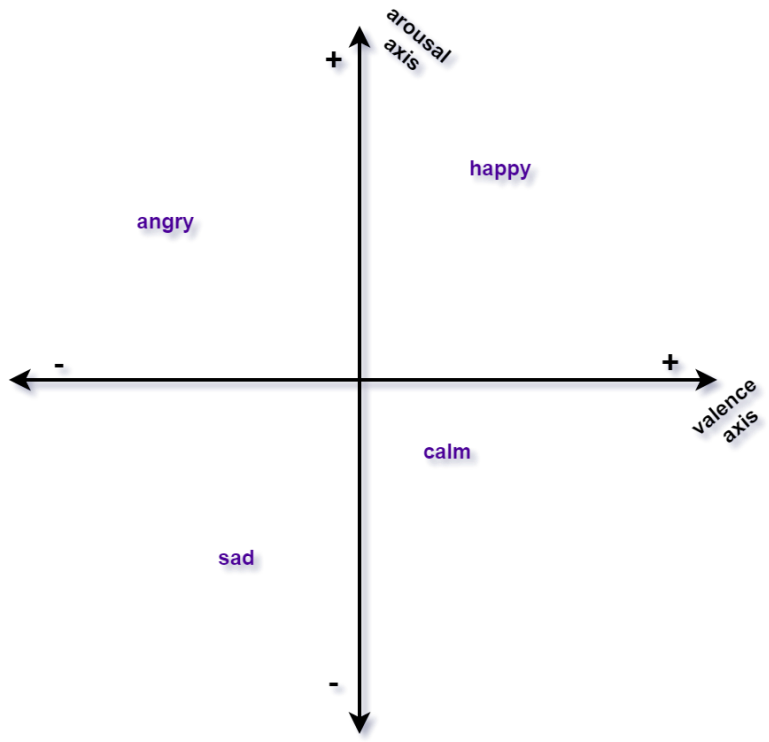

Figure 18: Position of the emotion categories on valence arousal axes

Vocalizations

Eighty-two words were voiced once for each emotional category for a total of four times by each one of the six amateur actors. Thus the corpus contains a total of 1735 words. 20% of the vocalizations were randomly selected for the validation set, whereas 80% of the vocalizations were randomly selected for the training set. Each vocalization is accompanied by a spectrogram. The vocalizations that belong to the validation set are accompanied by their versions in 0-8000 hertz band (the full spectrum), 0-5000 hertz band, and 500-8000 hertz band. Additionally, they have heat maps for these versions.

Table 26 presents the total number of vocalizations performed by actors, Table 27 presents the number of vocalizations in the training set, and Table 28 presents the number of vocalizations in the validation set.

Table 26: Total number of vocalizations performed by actors

Emotion	**Actors**						**Total**
	7895	**1984**	**1234**	**1358**	**1157**	**6783**	
Angry	82	82	82	82	77	82	487
Calm	80	82	82	82	0	82	408
Happy	29	82	82	82	0	82	357
Sad	82	82	82	82	73	82	483
Total	273	328	328	328	150	328	1735

Table 27: Number of vocalizations in the training set

Emotion	Actors						Total
	7895	1984	1234	1358	1157	6783	
Angry	66	62	67	64	58	72	389
Calm	61	67	62	69	0	67	329
Happy	25	64	64	65	0	67	285
Sad	68	69	66	71	55	57	386
Total	221	263	260	269	113	263	1389

Table 28: Number of vocalizations in the test set

Emotion	Actors						Total
	7895	1984	1234	1358	1157	6783	
Angry	16	20	15	18	19	10	98
Calm	19	15	20	13	0	15	82
Happy	4	18	18	17	0	15	72
Sad	14	13	16	11	18	25	97
Total	53	66	69	59	37	65	349

5.2.2 Statistics for Peripheral Components

We present the minimum, maximum, mean, standard deviation, and the range of the fundamental frequency (F_0) for each gender and emotional category. We expect F_0 values to vary with respect to gender and emotion. Therefore, Bayesian ANOVA and Bayesian t-test are conducted in order to determine the separability of the fundamental frequency statistics in terms of gender and the emotional category.

5.2.2.1 Fundamental Frequency Statistics in Terms of Gender

According to the results of the Bayesian independent t-test analysis, it is found that only F_0MIN yielded a significant difference (BF_{10}=1.2e+6, *Error*=6.4e-10) showing strong evidence for the separability of male (M=188.1, STD=288.6) and female (M=115.7, STD=224.4) actors in the overall corpus. See Table 29 for a summary of the fundamental frequency statistics in terms of gender.

Table 29: F_0 statistics in terms of gender

	F_0MIN		**F_0MAX**		**F_0MEAN**		**F_0STD**		**F_0RANGE**	
	F	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**
Valid	929	806	929	806	929	806	929	806	929	806
Missing	0	0	0	0	0	0	0	0	0	0
Mean	115.7	188.1	4787	4793	1737	1840	1498	1469	4671	4605
Std. Deviation	224.4	288.6	2910	4257	1139	1688	885.0	1375	2831	4181
Minimum	0.08807	0.1786	139.7	393.3	47.73	118.4	0.000	119.8	0.000	373.9
Maximum	2597	3228	1.508e+4	3.078e+4	7222	1.167e+4	4827	1.034e+4	1.467e+4	3.077e+4

5.2.2.2 Fundamental Frequency Statistics in Terms of Emotions

Maximum Fundamental Frequency (F_0MAX) Statistics in Terms of Emotions

According to the results of the Bayesian ANOVA test, when the random effects of gender and word type are controlled, F_0MAX is separable (*P(M)*=.5, BF_{10}=2.2e+87) in terms of emotions. According to the posthoc tests, it is found that emotional category pairs angry and happy are (BF_{10}=4.342, *Error*=2.4e-5) mildly separable while rest of the emotional categories are (BF_{10}>2.1e+24, *Error*<*1*.3e-10) strongly separable. Statistics are provided in Table 30.

Table 30: Maximum fundamental frequency (F_0MAX) statistics in terms of emotions

	F_0MAX			
	Angry	**Calm**	**Happy**	**Sad**
Valid	487	408	357	483
Missing	0	0	0	0
Mean	6113	3727	6872	2814
Std. Deviation	2939	2824	4728	2185
Minimum	1037	166.6	1236	139.7
Maximum	2.175e+4	1.634e+4	3.078e+4	1.168e+4

Minimum Fundamental Frequency (F_0MIN) Statistics in Terms of Emotions

According to the results of the Bayesian ANOVA test, when the random effects of gender and word type are controlled, F_0MAX is separable (*P(M)*=.5, BF_{10}=2.9e+14) in terms of emotions. According to the posthoc tests, it is found that only the emotion angry (BF_{10}>5836, *Error*<*2.3*e-8) is strongly separable from other emotions in terms of F_0MAX. Descriptive statistics of F0MIN are provided in Table 31.

Table 31: Minimum fundamental frequency (F_0MIN) statistics in terms of emotions

	F_0MIN			
	Angry	**Calm**	**Happy**	**Sad**
Valid	487	408	357	483
Missing	0	0	0	0
Mean	231.4	128.1	129.6	99.15
Std. Deviation	346.3	203.2	234.1	185.9
Minimum	0.4317	0.08807	0.1786	0.1601
Maximum	3228	1854	1735	1897

Mean Fundamental Frequency (F_0MEAN) Statistics in Terms of Emotions

According to the results of the Bayesian ANOVA test, when the random effects of gender and word type are controlled, F_0MEAN is separable (*P(M)*=.5, BF_{10}=9.9e+79) in terms of emotions. According to the posthoc tests, it is found that emotional category pairs angry and happy are (BF_{10}=.612, *Error*=1.5e-4) weakly separable while rest of the emotional categories are (BF_{10}>4991, *Error*<*2*.5e-9) strongly separable. Detailed statistics are provided in Table 32.

Table 32: Mean fundamental frequency (F_0MEAN) statistics in terms of emotions

	F_0MEAN			
	Angry	**Calm**	**Happy**	**Sad**
Valid	487	408	357	483
Missing	0	0	0	0
Mean	2302	1375	2524	1063
Std. Deviation	1182	1082	1941	865.5
Minimum	399.8	47.73	544.3	49.37
Maximum	7744	5874	1.167e+4	5482

Standard Deviation of Fundamental Frequency (F_0STD) Statistics in Terms of Emotions

According to the results of the Bayesian ANOVA test, when the random effects of gender and word type are controlled, F_0STD is separable (*P(M)*=.5, BF_{10}=9.9e+88) in terms of emotions. According to the posthoc tests, it is found that emotional category pairs angry and happy are (BF_{10}=13.39, *Error*=8.1e-6) moderately separable while rest of the emotional categories are (BF_{10}>28e+3, *Error*<*4*.6e-10) strongly separable. Detailed statistics are provided in Table 33.

Table 33: STD of fundamental frequency (F_0STD) statistics in terms of emotions

	F_0STD			
	Angry	**Calm**	**Happy**	**Sad**
Valid	487	408	357	483
Missing	0	0	0	0
Mean	1889	1126	2169	872.6
Std. Deviation	908.7	804.3	1583	663.8
Minimum	320.1	52.57	415.4	0.000
Maximum	6353	4245	1.034e+4	3372

Range of Fundamental Frequency (F_0RANGE) Statistics in Terms of Emotions

According to the results of the Bayesian ANOVA test, when the random effects of gender and word type are controlled, F_0RANGE is separable (*P(M)*=.5, BF_{10}=3.3e+87) in terms of emotions. According to the posthoc tests, it is found that emotional category pairs angry and happy are (BF_{10}=6.62, *Error*=6.9e-6) moderately separable while rest of the emotional categories are (BF10>58e+3, Error<1.6e-10) strongly separable. Detailed statistical information is provided in Table 34.

Table 34: Range of fundamental frequency (F_0RANGE) statistics in terms of emotions

	F_0RANGE			
	Angry	**Calm**	**Happy**	**Sad**
Valid	487	408	357	483
Missing	0	0	0	0
Mean	5881	3599	6742	2715
Std. Deviation	2858	2704	4687	2128
Minimum	991.8	164.5	1220	0.000
Maximum	2.149e+4	1.531e+4	3.077e+4	1.139e+4

5.2.3 Actor Statistics

Within the corpus study, six actors volunteered to join the study; all of them were amateur actors with little to no prior acting experience. The actors were between the ages of 23 and 35 with a mean, and standard deviation of 26.83 and 4.35 respectively. Three of the actors were female, and three of the actors were male. The female actors were between the ages of 23 and 28 with a mean and standard deviation of 25.3 and 2.5, respectively. The male actors were between the ages of 24 and 35 with a mean and standard deviation of 28.3 and 5.9, respectively (see Table 35). Actors were presented with the Amateur Actor Guidelines presented in the APPENDIX A.

Table 35: Actor Statistics

Actor ID	Gender	Age
7895	Female	28
1984	Female	25
1234	Female	23
1358	Male	24
1157	Male	35
6783	Male	26

5.3 Corpus Evaluation

5.3.1 Sample-Population Evaluation

As already mentioned, the population of the corpus consists of 1735 vocalizations. For a 95% confidence interval, the ideal sample size would be 315. In the study, we used 349 samples as the validation set. In other words, the corpus can be assumed to have an adequate sample size for its validation set.

5.3.2 The Reliability Study

A reliability study was conducted to evaluate the consistency of the categorization of vocalizations into emotions. One judge rated each frequency band. The accuracy calculated by using the original categories as the key. According to the results of the reliability study; the judge of 0-8000 hertz frequency band attained 65% accuracy. The respective judges of 0-5000 and 500-8000 hertz bands achieved 64% accuracy. We consider results over 60% a feasible accuracy rating considering that the judges only listened to single-word-vocalizations rather than full sentences.

5.4 Conclusion

In this chapter, we investigated the TurEV database through different methods. The results of the earlier chapters indicate that the corpus has an adequate number of words and vocalizations to enable a machine learning experiment and further analysis. The fundamental frequency was adequate to separate most of the emotional categories. However, F_0 yielded low confidence for the separation of the category happy and angry.

5.5 Summary

In this chapter, we have investigated the corpus through statistical and descriptive methods. In the next chapter, we will evaluate the previous chapters and conclude the study.

Chapter VI

6. Conclusion

The thesis aimed at developing a novel model for studying emotion in the speech signal. This novel model was intended to be robust to noise, and trained on Turkish words. Moreover, the model was inspected with techniques such as gradient visualizations. In the validation process of the model, our ablation study allowed us to contrast the model with the judges in different frequency bands and observe the parallelisms and differences. The observations we have made allowed us to gain insight into the cognitive processes of the human mind as opposed to the calculations of the model. To accomplish all these, we had to develop a Turkish emotion-voice database from the ground up. Lastly, we aim at making public not only the results of the study but also the TurEV database and the source codes used in the study.

6.1 Contributions

6.1.1 The Methodological Contributions

The model developed through the study is one of its kind. The CNN models are developed for object detection. The results indicate it is feasible to use the CNN models in emotion-voice studies. The model is found to be robust to some changes in the speech signal.

The Grad-CAM model was developed on top of the CNN model. Heat maps were successfully extracted. Heat maps themselves show that it is possible to extract visual information from the voice signal.

Overall, the CNN model allows visual patterns to be incorporated into studies and engender new research methodologies.

6.1.2 The Empirical Contributions

A corpus of emotion in voice appeared as one of the gaps in Turkish emotion-voice studies. During the thesis study, we developed the Turkish Emotion Voice Database (TurEV). TurEV is a multipurpose database that integrates an emotion-voice corpus for four different emotional states, train and test sets, spectrograms, and statistics and heat maps for the test set.

TurEV will be open to the public. It can be exploited by phonologists and emotion researchers, and future studies can use and improve the database.

6.1.3 The Cognitive Contributions

The results we obtained from the model, the judges, and the comparative study allowed us to investigate the emotion in voice from multiple perspectives. Moreover, the ablation study allowed us to investigate the differences when certain frequency bands are removed from the sound signal.

To sum up, the judges were robust to the changes in the speech signal, which implies the holistic processing of the sound signal in general, but the model lost most of its categorization power with the removal of low pitch frequencies. The most exciting discovery was that the model used the 5000-8000 hertz band to differentiate between happy and angry, whereas the judges used this same band to differentiate between angry and calm. This differentiation implies that the model is more focused on energy and differentiates high energy emotions (angry and happy) through the variance in the energy, whereas humans focus on any change such as subtle variations in the speech signal.

6.2 Limitations

One of the limitations of the study was related to data collection. The vocalizations were limited by the lack of proper hardware. The number of man-hours required for the collection and processing of the vocalizations was enormous. As a result, the number of vocalized emotion categories were limited to four. Yet, each category represented one main area of the arousal-valence space. Also, only single-word vocalizations were recorded instead of phrases or sentences. Finally, three judges were recruited, each of whom evaluated one frequency band to decide which vocalization fits which of the four emotion categories.

The second limitation was that the CNN model could not be utilized to its full potential. The heat maps could only be briefly inspected, and intermediate activations could not be extracted due to various constraints. With more data and more rigorous training regime, the model could learn more features. However, it is important to note that This limitation was also a feature in the study; we needed to learn how the model failed. For our purposes, a model with a very high classification rate would defeat the aim of the study.

6.3 Future Work

The constraints and the findings of the thesis engender a wide variety of studies possible. For example, the database and the model developed through the study has many different unexplored aspects. The heat maps and intermediate activations will allow a much larger phonological study to be conducted. In particular, further study is needed to understand what exactly is present in the 500-8000 hertz band that allow humans to make the correct predictions to distinguish among the four emotion categories while the model loses much of its predictive power when it loses the information in the 0-500 hertz band.

A more comprehensive Turkish emotion-voice database can be built using professional actors, equipment, and a more comprehensive validation study. Nevertheless, we hope that the TurEV database is both a starting point for many different studies and a reference for future ones.

The CNN architecture was found to be viable for sound-signal processing. The sound signal can be used for emotion recognition in various new ways. Moreover, the sound signal can be used for various linguistic tasks such as sentence segmentation, speech-to-text processing, or to enable diagnosis in psychology.

VII.References

Alghowinem, S., Goecke, R., Epps, J., Wagner, M., & Cohn, J. (2016). Cross-cultural depression recognition from vocal biomarkers. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *08-12-Sept*, 1943–1947. https://doi.org/10.21437/Interspeech.2016-1339

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders : DSM-5. American Psychiatric Association. DSM*. https://doi.org/10.1176/appi.books.9780890425596.893619

Anaconda. (2014). Anaconda Software Distribution. *Computer Software*.

Audacity. (2018). Audacity ® | Free, open source, cross-platform audio software for multi-track recording and editing.

Audcaity Team. (n.d.). Audacity Manual. Retrieved February 4, 2019, from https://manual.audacityteam.org/man/noise_reduction.html

Barrett, L. F. (1998). Discrete Emotions or Dimensions? The Role of Valence Focus and Arousal Focus. *Cognition and Emotion*. https://doi.org/10.1080/026999398379574

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B., Berlin, T. U., … Berlin, H. U. (2005). A Database of German Emotional Speech. In *Proc. of INTERSPEECH 2005*.

Busso, C., Bulut, M., Lee, S., & Narayanan, S. (2008). Fundamental frequency analysis for speech emotion processing (pp. 309–337).

C, N. C. (2010). Nvidia cuda. *Compare A Journal Of Comparative Education*.

Cai, Z., Fan, Q., Feris, R. S., & Vasconcelos, N. (2016). A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision -- ECCV 2016* (pp. 354–370). Cham: Springer International Publishing.

Chollet François. (2015). Keras: The Python Deep Learning library. *Keras.Io*. https://doi.org/10.1086/316861

Cohen, A. S., Najolia, G. M., Kim, Y., & Dinzeo, T. J. (2012). On the boundaries of blunt affect/alogia across severe mental illness: Implications for Research Domain Criteria. *Schizophrenia Research*. https://doi.org/10.1016/j.schres.2012.07.001

Cohen, M. (2014). *Analyzing Neural Time Series Data: Theory and Practice*. *MIT Press*. https://doi.org/2008-2045

Cohn, J. F., Kruez, T. S., Matthews, I., Yang, Y., Nguyen, M. H., Padilla, M. T., … Torre, F. De. (2009). Detecting Depression from Facial Actions and Vocal

Prosody.pdf. *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference On*.

Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*. https://doi.org/10.1016/S0167-6393(02)00071-7

DataCamp. (2016). Jupyter Notebook. *DataCamp*. https://doi.org/10.1023/a:1009769707641

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium On Motivation*. https://doi.org/10.1037/0022-3514.53.4.712

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., … Tzavaras, A. (1987). Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/0022-3514.53.4.712

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*. https://doi.org/10.1016/j.patcog.2010.09.020

Erdem, E. S. (2014). *Emotion Recognition and Retrieval*. Başkent University.

Ergenç, İ., & Bekar Uzun, İ. P. (2017). *Türkçenin Ses Dizgesi* (1st ed.). Ankara: Seçkin Yayıncılık.

Fidan, D. (2007). Türkçe ezgi örüntüsünde duygudurum ve sözedim görünümü, (2005).

Fundamental Frequency of Continuous Signals. (2011). Retrieved from http://fourier.eng.hmc.edu/e101/lectures/Fundamental_Frequency.pdf

GoogleResearch. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. *Google Research*. https://doi.org/10.1207/s15326985ep4001

Hansen, J. H. L., & Bou-Ghazale, S. E. (1997). Getting started with SUSAS: a speech under simulated and actual stress database. In *EUROSPEECH*.

Hoy, M. (2018). Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Medical Reference Services Quarterly*, *37*, 81–88. https://doi.org/10.1080/02763869.2018.1404391

Hozjan, V., Zdravko, K., Asuncion, M., Antonio, B., & Albino, N. (2002). Interface Databases: Design and Collection of a Multilingual Emotional Speech Database. In *LREC*.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, *9*(3), 99–104. https://doi.org/10.1109/MCSE.2007.55

Johnstone, T., & Scherer, K. (1999). The effects of emotions on voice quality. *Proceedings of the 14th International Conference of Phonetic Sciences*,

(January), 2029–2032. Retrieved from http://eclub.unige.ch/system/files/biblio/1999_Johnstone_ICPS.pdf

Kim, E. H., Hyun, K. H., Kim, S. H., & Kwak, Y. K. (2007). Speech Emotion Recognition Using Eigen-FFT in Clean and Noisy Environments. In *RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 689–694). https://doi.org/10.1109/ROMAN.2007.4415174

Kim, Y., Lee, H., & Provost, E. M. (2013). Deep learning for robust feature generation in audiovisual emotion recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. https://doi.org/10.1109/ICASSP.2013.6638346

Konstantareas, M. M. (2006). Social skills training in high functioning autism and Asperger's Disorder. *Hellenic Journal of Psychology*.

Lassalle, A., Pigat, D., O'Reilly, H., Berggen, S., Fridenson-Hayo, S., Tal, S., … Lundqvist, D. (2018). The EU-Emotion Voice Database. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-1048-1

Lemmetty, & Sami. (1999). Phonetics and Theory of Speech Production. Retrieved from http://research.spa.aalto.fi/publications/theses/lemmetty_mst/chap3.html

Liscombe, J., Venditti, J. J., & Hirschberg, J. (2003). Classifying subject ratings of emotional speech using acoustic features. *Proceedings of Eurospeech*. https://doi.org/10.1.1.117.5988

Mao, S., Tao, D., Zhang, G., Ching, P. C., & Lee, T. (2019). Revisiting Hidden Markov Models for Speech Emotion Recognition. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6715–6719). https://doi.org/10.1109/ICASSP.2019.8683172

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*.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, *1697900*(Scipy), 51–56. Retrieved from http://conference.scipy.org/proceedings/scipy2010/mckinney.html

Microsoft. (n.d.). PCM Stream Data Format. Retrieved February 4, 2019, from https://docs.microsoft.com/en-us/windows-hardware/drivers/audio/pcm-stream-data-format

Mitra, V., Shriberg, E., Vergyri, D., Knoth, B., & Salomon, R. M. (2015). Cross-corpus depression prediction from speech. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2015-Augus*, 4769–4773. https://doi.org/10.1109/ICASSP.2015.7178876

Morrison, D., Wang, R., & De Silva, L. C. (2007). Ensemble methods for spoken

emotion recognition in call-centres. *Speech Communication*. https://doi.org/10.1016/j.specom.2006.11.004

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

Oliphant, T. E. (2007). SciPy: Open source scientific tools for Python. *Computing in Science and Engineering*, *9*, 10–20. https://doi.org/10.1109/MCSE.2007.58

OpenCV. (2010). Open Source Computer Vision Library.

Owens, F. J., & Murphy, M. S. (1988). A short-time Fourier transform. *Signal Processing*. https://doi.org/10.1016/0165-1684(88)90040-0

Pachet, F., & Roy, P. (2009). Analytical features: A knowledge-based approach to audio feature generation. *Eurasip Journal on Audio, Speech, and Music Processing*. https://doi.org/10.1155/2009/153017

Pacific Northwest Seismic Network. (n.d.). What is a Spectrogram? Retrieved February 4, 2019, from https://pnsn.org/spectrograms/what-is-a-spectrogram

Paeschke, A., Kienast, M., & Sendlmeier, W. F. (1999). F0 -Contours in emotional speech. In *Proceedings of 14th International Congress of Phonetic Science*.

Pereira, C. (2000). Dimensions of emitional meaning in speech. *ISCA Archive*. https://doi.org/10.1161/CIRCOUTCOMES.116.002749

Podder, P., Zaman Khan, T., Haque Khan, M., & Muktadir Rahman, M. (2014). Comparative Performance Analysis of Hamming, Hanning and Blackman Window. *International Journal of Computer Applications*. https://doi.org/10.5120/16891-6927

Rabiner, L., & Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, *3*(1), 4–16. https://doi.org/10.1109/MASSP.1986.1165342

Scherer, K. R. (1982). Emotion as a process: Function, origin and regulation. *Social Science Information*. https://doi.org/10.1177/053901882021004004

Scherer, K. R. (1987). Toward a dynamic theory of emotion: The component process model of affective states. *Geneva Studies in Emotion and Communication*. https://doi.org/10.1080/02699930902928969

Scherer, K. R. (2001). Appraisal Considered as a Process of Multilevel Sequential Checking. In *Appraisal process in emotion: Theory, Methods, Research*. https://doi.org/10.1007/s10566-008-9057-3

Spyder. (2018). SPYDER IDE.

Tang, H., Chu, S. M., Hasegawa-Johnson, M., & Huang, T. S. (2009). Emotion recognition from speech VIA boosted Gaussian mixture models. In *2009 IEEE International Conference on Multimedia and Expo* (pp. 294–297). https://doi.org/10.1109/ICME.2009.5202493

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller,

B., & Zafeiriou, S. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. https://doi.org/10.1109/ICASSP.2016.7472669

Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, *13*(2), 22–30. https://doi.org/10.1109/MCSE.2011.37

Vogt, T., & André, E. (2005). Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *IEEE International Conference on Multimedia and Expo, ICME 2005*. https://doi.org/10.1109/ICME.2005.1521463

Wang, K.-C. C. (2014). The feature extraction based on texture image information for emotion sensing in speech. *Sensors (Switzerland)*, *14*(9), 16692–16714. https://doi.org/10.3390/s140916692

Weißkirchen, N., Böck, R., & Wendemuth, A. (2018). Recognition of emotional speech with convolutional neural networks by means of spectral estimates. In *2017 7th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW 2017* (Vol. 2018-Janua, pp. 50–55). https://doi.org/10.1109/ACIIW.2017.8272585

Wilkinson, L., & Friendly, M. (2009). History of the Cluster Heat Map. *The American Statistician*. https://doi.org/10.1198/tas.2009.0033

Yao, Y.-C. (2014). Nyquist Frequency. In *Wiley StatsRef: Statistics Reference Online*. https://doi.org/10.1002/9781118445112.stat03517

APPENDICES

APPENDIX A

Amateur Actor Data Collection Manual

Kelime Bazlı Duygu Verisi
Amatör Aktör Klavuzu

Çalışma esnasında sizlere sunulan kelimeleri mutlu, hüzünlü, sakin, ve kızgın olmak üzere dört farklı duygu durumunda dile getirmeniz rica edilmektedir. Her bir kelimeyi, her bir duygu durumu için dile getirip ayrı ses dosyaları halinde kaydetmeniz gerekmektedir.

Dikkat Edilmesi Gereken Hususlar

- Dört basamaklı Amatör Aktör Numaranız (AAN) size ayrıca verilecektir.
- Seslendirilecek kelimeler size kelimeler.xls adlı bir dosya içerisinde size sunulacaktır.
- Kayıt durumundan önce yankı bulunmayan ve sessiz bir ortamda olduğunuzdan emin olun.
- Android cihazlarda yapacağınız kayıt için Sony Mobile Productions'dan Audio Recorder adlı programı kullanınız.
- Aldığınız kayıtları mono olarak ve wav formatında alınız.
- Kelimeleri her bir duygu durumu için seslendiriniz, her bir kelime için dört farklı ses dosyanız olacak.
- Kelimeleri seslendirirken o duyguyu hissediyormuş gibi seslendirin.
- Kelimeleri seslendirirken sesinizi yükseltmeden fakat duyguyu (sakin duygu durumu harici) yoğun şekilde hissediyormuş gibi seslendirin.
- Tercihen bir oturumda tek bir duyu durumu üstünde çalışın, duygu durumları arasında sıkça geçiş yapmamaya dikkat edin.
- Seslendirmeleri o duyguyu yoğun olarak hissediyormuşçasına yapın lakin ses tonunuzu normal tonda tutun.

Kayıt Süreci

- Ses kaydı için kullanacağınız cihazı (telefon, mikrofon, vb) ile aranızda 30cm yahut kayıt mikrofonları için cihazın tavsiye edilen uzaklığı kadar mesafe bırakın.
- Kayıt düğmesine basın, 1 saniye bekleyin, sözcüğü ilgili duygu durumuna uygun olarak seslendirin, 1 saniye bekledin ve kaydı durdurun.

- Kaydı dinleyin, mümkün olduğu durumda üçüncü bir kişiye sunun. Kayıt ilgili duydu durumunu net olarak yansıtıyorsa aşağıda verilen adda kaydedin. Yansıtmıyorsa o kelime-duygu durumu çifti için süreci tekrarlayın.
- Ses dosyalarını AAN_DK_KELIME.wav formatında kaydediniz.
    - AAN (Amatör Aktör Numarası) size sunulan dört basamaklı rakamdır.
    - DK (Duygu Kodu) Her bir duygu için belirtilen kod.
        - Mutlu: MT
        - Hüzünlü: HL
        - Kızgın: KZ
        - Sakin: SK
    - Kelime: Kelimenin kendisi. Size sunulan kelimeler.xls dosyası içinde yer almaktadır.

APPENDIX B

LIST OF WORDS USED FOR VOCALIZATION

açık	demet	fırıldak	kağıt	kulaç	oğul	şarap	tahta	yığın
algı	demir	gezi	karton	kuşluk	okul	şarj	tembel	zeytin
beste	deve	gıcık	kedi	küstah	ölçüt	satır	tırtıl	
buğu	deyim	göz	kemik	lokma	ördek	şenlik	türev	
çay	dil	güneş	kibir	lüfer	örgü	sepet	ünlü	
çekirge	dizge	gürbüz	kiler	melek	pamuk	serin	utanç	
çene	düğün	hekim	kırgın	müjde	pas	sevgi	vurgu	
ceren	eğer	iğde	koku	nasıl	pıhtı	simit	yankı	
çilek	elmacık	ısırgan	konuk	niçin	rıhtım	soba	yazım	
defne	eski	japon	köy	ödenek	saf	sucuk	yığılmak	