Université de Montréal

Creation of a Vocal Emotional Profile (VEP) and Measurement Tools.

Par

Mahsa Aghajani

Département d'informatique et de recherche opérationnelle

Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de Maître ès sciences (M.Sc.)

en 217510, avec mémoire

Octobre, 2021

Université de Montréal

Département d'informatique et de recherche opérationnelle

Faculté des arts et des sciences

Ce mémoire intitulé

Creation of a Vocal Emotional Profile (VEP) and Measurement Tools.

Présenté par

Mahsa Aghajani

A été évalué(e) par un jury composé des personnes suivantes

Esma Aïmeur
Président-rapporteur

Claude Frasson
Directeur de recherche

Aishwarya Agrawal
Membre du jury

Résumé

La parole est le moyen de communication dominant chez les humains. Les signaux vocaux véhiculent à la fois des informations et des émotions du locuteur. La combinaison de ces informations aide le récepteur à mieux comprendre ce que veut dire le locuteur et diminue la probabilité de malentendus. Les robots et les ordinateurs peuvent également bénéficier de ce mode de communication. La capacité de reconnaître les émotions dans la voix des locuteurs aide les ordinateurs à mieux répondre aux besoins humains. Cette amélioration de la communication entre les humains et les ordinateurs conduit à une satisfaction accrue des utilisateurs. Dans cette étude, nous avons proposé plusieurs approches pour détecter les émotions de la parole ou de la voix par ordinateur. Nous avons étudié comment différentes techniques et classificateurs d'apprentissage automatique et d'apprentissage profond permettent de détecter les émotions de la parole. Les classificateurs sont entraînés avec des ensembles de données d'émotions audio couramment utilisés et bien connus, ainsi qu'un ensemble de données personnalisé. Cet ensemble de données personnalisé a été enregistré à partir de personnes non-acteurs et non-experts tout en essayant de déclencher des émotions associées. La raison de considérer cet ensemble de données important est de rendre le modèle compétent pour reconnaître les émotions chez les personnes qui ne sont pas aussi parfaites que les acteurs pour refléter leurs émotions dans leur voix. Les résultats de plusieurs classificateurs d'apprentissage automatique et d'apprentissage profond tout en reconnaissant sept émotions de colère, de bonheur, de tristesse, de neutralité, de surprise, de peur et de dégoût sont rapportés et analysés. Les modèles ont été évalués avec et sans prise en compte de l'ensemble de données personnalisé pour montrer l'effet de l'utilisation d'un ensemble de données imparfait. Dans cette étude, tirer parti des techniques d'apprentissage en profondeur et des méthodes d'apprentissage en ensemble a dépassé les autres techniques. Nos meilleurs classificateurs pourraient obtenir des précisions de 90,41 % et 91,96 %, tout en étant entraînés par des réseaux de neurones récurrents et des classificateurs d'ensemble à vote majoritaire, respectivement.

Mots-clés: Reconnaissance vocale des émotions, Machine Learning, Deep Learning, Ensemble Learning, Interface cerveau-ordinateur, Informatique affective.

Abstract

Speech is the dominant way of communication among humans. Voice signals carry both information and emotion of the speaker. The combination of this information helps the receiver to get a better understanding of what the speaker means and decreases the probability of misunderstandings. Robots and computers can also benefit from this way of communication. The capability of recognizing emotions in speakers voice, helps the computers to serve the human need better. This improvement in communication between humans and computers leads to increased user satisfaction. In this study we have proposed several approaches to detect the emotions from speech or voice computationally. We have investigated how different machine learning and deep learning techniques and classifiers perform in detecting the emotions from speech. The classifiers are trained with some commonly used and well-known audio emotion datasets together with a custom dataset. This custom dataset was recorded from non-actor and non-expert people while trying to trigger related emotions in them. The reason for considering this important dataset is to make the model proficient in recognizing emotions in people who are not as perfect as actors in reflecting their emotions in their voices. The results from several machine learning and deep learning classifiers while recognizing seven emotions of anger, happiness, sadness, neutrality, surprise, fear and disgust are reported and analyzed. Models were evaluated with and without considering the custom data set to show the effect of employing an imperfect dataset. In this study, leveraging deep learning techniques and ensemble learning methods has surpassed the other techniques. Our best classifiers could obtain accuracies of 90.41% and 91.96%, while being trained by recurrent neural networks and majority voting ensemble classifiers, respectively.

Keywords: Voice Emotion Recognition, Machine Learning, Deep Learning, Ensemble Learning, Brain Computer Interface, Affective Computing.

Table of contents

List of tables

List of figures

List of acronyms and abbreviations

CNN: Convolutional Neural Network

EEG: Electroencephalogram

GSVM: Gaussian kernel Support Vector Machine

HMM: Hidden Markov Model

HNR: Harmonic-to-Noise Ratio

IVR: Interactive Voice Response

LLD: Low-Level Descriptors

LSVM: Linear Support Vector Machine

MFCC: Mel Frequency Cepstral Coefficient

RAVDESS: Ryerson Audio-Visual Database of Emotional Speech and Song

RNN: Recurrent Neural Network

SAVEE: Surrey Audio-Visual Expressed Emotion

SER: Speech Emotion System

SVM: Support Vector Machine

TESS: Toronto Emotional Speech Database

Acknowledgments

The completion of this study could not have been possible without the invaluable guidance of my research supervisor, Prof. Claude Frasson. I would like to express my sincere gratitude to him for all his support, empathy and motivation. His advice carried me through all the stages of my research and writing of this thesis.

In addition, a debt of gratitude is owed to Dr. Hamdi Ben Abdessalem, for his keen interest on me at every stage of my research. I’m truly inspired by his dynamism, kindness and patience in helping me with different aspects of my research.

Last but not the least, I would like to thank my parents, who have devoted their life their children. Without you none of this would indeed be possible.

Chapter 1

Introduction

Speech is the dominant way of communication among humans. One of the reasons behind this dominance is that speech carries both information and emotions. Embedding the emotions in this way of communication is so important. This importance is especially noticed when we choose other ways of communication, such as text messages. Lack of emotions in the messages may lead to misinterpretation. Therefore, using speech is more efficient, regarding the information that the other ends of communication receive. It’s not only humans who can benefit from this information; Computers can also use this method of communication to gain a better understanding of humans needs. This goal requires computers to be able to recognize emotions from the speech or the voice of the human, whom they are interacting with. The process of associating human properties such as observing, interpreting and generating affective features to computers is known as affective computing (J. Tao & Tan, 2005).

1.1 Affective Computing

Affective computing goal is to enhance the interaction between humans and computers. While leveraging affective computing, a computer can detect the user’s emotion and generate a counter response for increasing user’s satisfaction. For example, in today's applications, recognizing emotions in users' voices leads to a better user experience and more user-friendly applications. Different applications can benefit from detecting the current emotion of the user by analyzing the input voice. In applications where the user’s satisfaction is an important factor or the competitive advantage of that application, detecting the emotion of the user through interacting with the application becomes of great value. Examples of these applications are customer support applications, artificially intelligent virtual assistants, etc. (Petrushin, 2000).

Along with these mentioned applications, many other applications can also benefit from voice emotion detection hugely but have some specific requirements too. One of these specific requirements for this set of applications is that the emotion detection should be done in real time. Instances of such applications are educational applications, video games, driving assistant applications, etc. The need for real time emotion detection in these applications originates from enabling the application to behave differently based on the current emotion of the user. For example, if the student is stressed, angry or sad, providing the regular material in an educational application may not be the best practice. For Alzheimer patients detecting negative emotions would be important to trigger some relaxing mechanisms.

Detecting the emotion using the speaker’s voice in real time applications encounters several challenges indicated below:

First the nature of this detection is a completed task, even for humans. In different scenarios, based on the person's ability to reflect his emotions in his voice, this recognition task can get even harder. For example, when the speaker suffers from the Alzheimer disease or when the speaker has autism. Even with people who do not have these conditions, detecting several pairs of emotions from each other in their voice, such as happiness and surprise, or neutrality and sadness requires high proficiency. In all these cases, the predictor model should already be trained with similar data.

The second challenge is about identifying the most relevant and important characteristics of voice signals which are proper for emotion detection. For recognizing the emotions, any speech emotion detection system, should analyze the received voice signals. Voice or speech analysis is the process of analyzing the speech signal to obtain compact relevant information of the signal (Toledano et al., 2009). Speech signals contain tons of information and characteristics. Regarding to the different applications, different sets of these characteristics are necessary to be leveraged. Therefore, choosing the most relevant characteristics in voice signals, plays a vital role in building a successful voice emotion detection system.

The third challenge is related to the real time emotion detection requirement of the target applications. Considering this specific need, the emotion detection should be done in an acceptable amount of time.

1.2 Research objectives and questions

The objective of this work is to develop an accurate voice emotion detection system, capable of recognizing 7 emotions of anger, happiness, sadness, surprise, fear, disgust and neutrality from voice of speakers who are not necessarily actors or experts in reflecting the emotions in their voice. To reach this goal, we will investigate three following hypotheses:

The first hypothesis is that if it's possible to use machine learning techniques for training models capable of recognizing and classifying emotions from non-actor people's voices? The reason behind this hypothesis is that machine learning has progressed dramatically over the past two decades. Machine learning has emerged as the method of choice for building practical software for many applications including speech and voice classification(Jordan & Mitchell, 2015). Therefore, we want to explore the possibility of adopting these techniques in recognizing the emotions from speech.

The second hypothesis is if gathering more data and leveraging deep learning techniques may lead to obtaining higher accuracies for detecting the emotions from non-actor people's voices? The motivation behind this hypothesis is the tremendous progress of deep learning methods in recent years. Training accurate deep learning classifiers requires a very large collection of labeled data. Using these large-scale deep learning models has had a major effect in speech recognition and resulted in major improvements over previous approaches(Jordan & Mitchell, 2015).

The third and last hypothesis is what the effect of adopting ensemble learning classifiers can be on the performance measures of the final voice emotion detection system? If we are successful in training accurate classifiers while exploring the previous hypotheses, would it be possible to combine their strength into a more accurate one? We know that using ensemble classifiers is an emerging trend in artificial intelligence. These set of classifiers tend to combine several already trained classifiers and would combine their strength in recognizing the true label for each case.

Considering our main objective and the three described hypotheses that we will try to explore during this study, we must reach these objectives:

- Gathering a large scale and rich collection of labeled emotion voice files.
- Gathering voice samples from non-actor speakers, which should reflect their emotion.
- Developing accurate classifiers, while leveraging machine learning, deep learning or ensemble learning techniques, capable of recognizing 7 emotions in speaker's voice.
- Evaluating the developed classifiers regarding several performance measures

1.3 Outline of the thesis

- Chapter 2: contains fundamentals of recognizing emotions in speech. The reviewed fundamentals in this part include some basic definitions of emotions, the important characteristics of speech for recognizing the emotions and the well-known and commonly used speech emotion detection datasets and their properties.
- Chapter 3: provides an overview of the literature related to most of the previously developed speech emotion recognition systems.
- Chapter 4: aims to explore our first hypothesis. We will figure out the possibility of training classifiers with machine learning classification algorithms.
- Chapter 5: presents analyzing our second hypothesis. The approach and results of using deep learning techniques for making a voice emotion detection classifier are explained in detail in this chapter.
- Chapter 6: explains the answers to our third hypothesis. We will explain the steps and results of using ensemble techniques for building a voice emotion detection system.
- Chapter 7: concludes the thesis regarding our three hypotheses and describes our perspective for the future works.

Chapter 2

Fundamentals of Emotions in Speech

For discussing the voice emotion recognition systems, first we should get familiar with the underlying elements in these systems. From an abstract view, any recognition system is composed of three parts: input, processing module and the output. The input in a SER system is voice or speech signals. The processing unit is responsible for mapping the input speech signals to one of the output classes. Finally, the output in a SER system is the recognized emotion in the input speech. Before getting deeper about how these recognition systems work, it's necessary to discuss the definition of input and output, in this case, speech signals and emotions. In the following sections, first we will provide a brief definition for emotion. After that, we will focus on the input, or speech signals. In this part, the important characteristics of speech signals for recognizing emotions are introduced.

After discussing the characteristics of speech signals, we will go over databases that contain this information and are commonly used in speech emotion recognition studies.

2.1 Models of Emotions

A voice emotion recognition system is going to recognize emotions; Thus, we should define emotions first. Of course, defining emotions is still an open problem in psychology. Based on (Plutchik, 2001) there are more ninety definitions for emotions. Based on these definitions, two models are commonly used among in voice emotion recognition tasks: discrete emotional model and dimensional emotional model.

2.1.1 Discrete Emotional Model

In this theory, 6 basic independent categories of emotions are defined: happiness, sadness, fear, disgust, anger and surprised (Ekman et al., 2013). Other emotions are result of combining these 6 basic emotions. This discrete model is used commonly for addressing everyday emotions, since it’s easier to map observed emotions to one of the categories; On the other hand, this model is unable to define some of more complex emotions.

2.1.2 Dimensional Emotional Model

In this model, emotions are addressed by using a small number of latent dimensions. Examples of these dimensions are valence, arousal, control and power(Russell & Mehrabian, 1977). In contradiction to the discrete model, in the dimensional model, emotions are not independent of each other, but analogous. It should be noted that mapping emotions to the dimensional model, is not intuitive and requires training(Zeng et al., 2009).

2.2 Fundamentals in Voice Emotion Recognition

In this part, first we will briefly introduce waveforms with a few examples. Then we will explain extracting which features from these waveforms helps us in recognizing the emotions. After describing these features, several commonly used databases for voice emotion recognition are reported. These databases contain speech audio files that are labeled by the emotion of the speaker.

2.2.1 Voice Features in Speech Emotion Recognition

In a voice emotion recognition system, emotions are recognized based on voice features; therefore, choosing appropriate features that represent emotions properly is essential. In general, voice features can be divided into two categories: global features and local features. Global or long-term or supra-segmental features contain information about gross statistics such as mean, maximum and minimum values and standard deviation. On the other hand, local or short-term or segmental values show the temporal dynamics(Akçay & Oğuz, 2020). These short-term features are used for approximating a stationary state, since the emotional features may not be uniformly distributed over all parts of the voice signal(Rao et al., 2013).

Speech signals are variations of air pressure over time. For digitalizing this information, we will capture samples of the air pressure over time at a specific sample rate. This captured information, which is called waveform, is the first form of the input for an emotion recognition system. Figures 2.1 to 2.7 show waveforms for 7 different sample voices, associated with 7 emotions of anger, happiness, sadness, fear, disgust, surprise and neutrality, respectively.

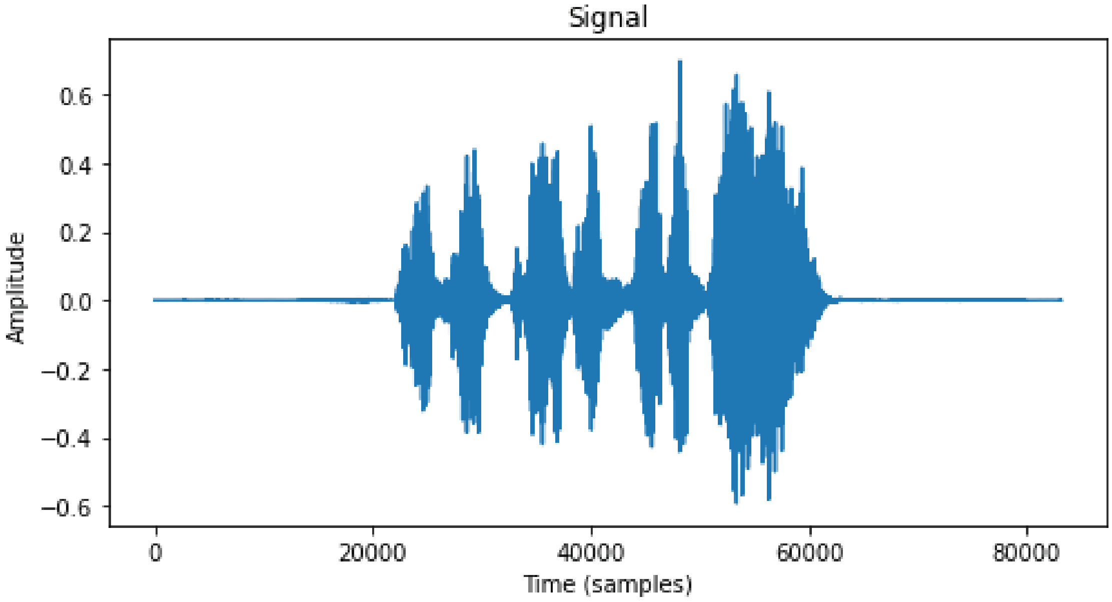


Figure 2.1 The waveform of a sample angry voice, showing amplitude over time.

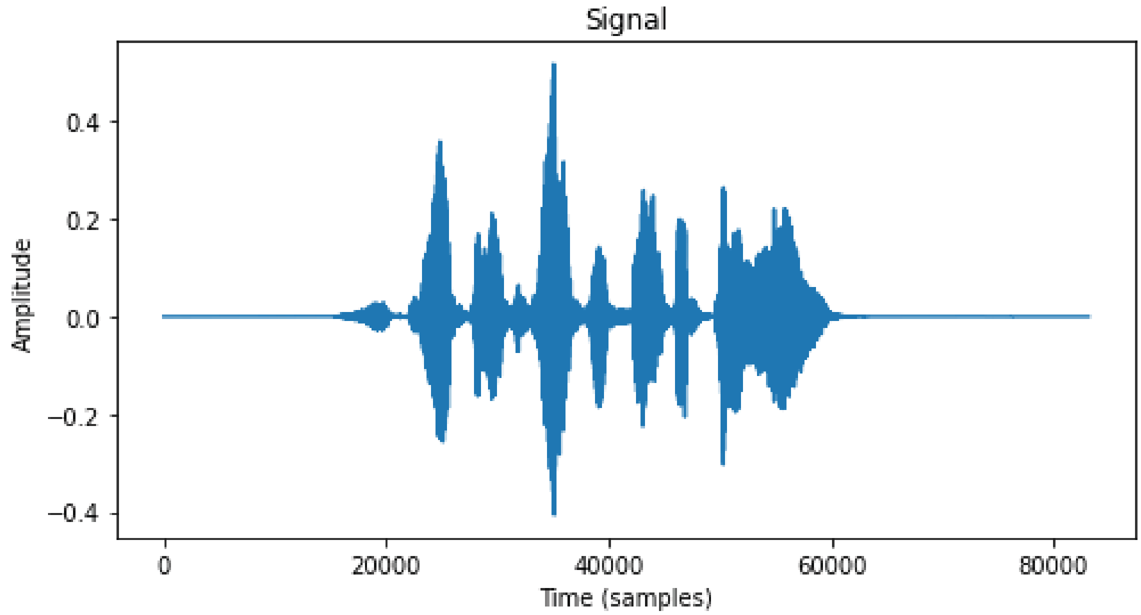


Figure 2.2 The waveform of a sample happy voice, showing amplitude over time.

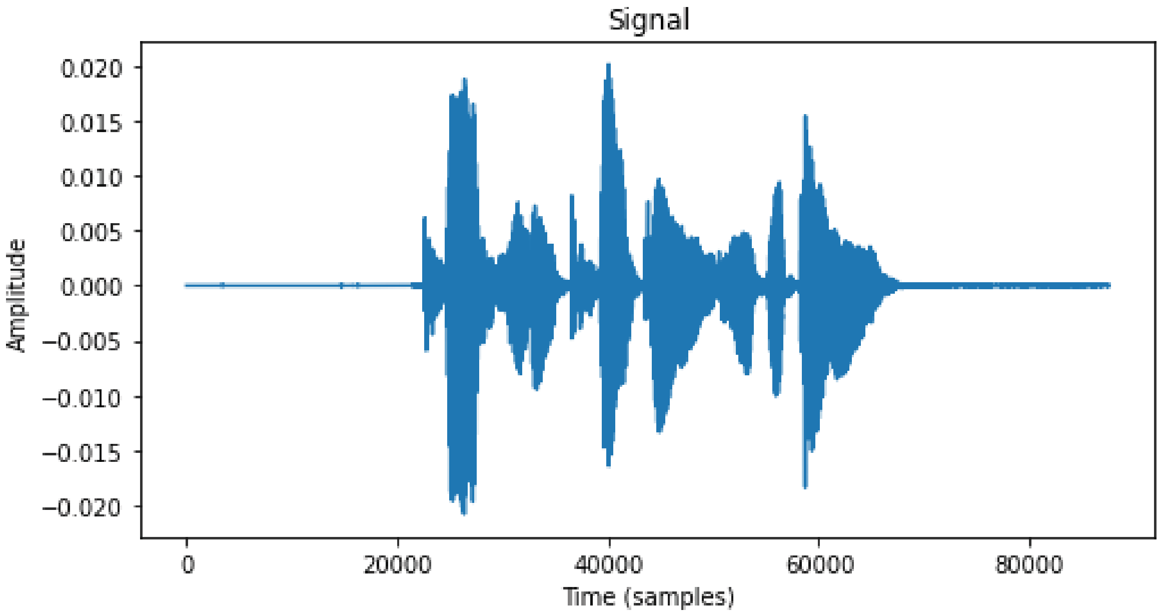


Figure 2.3 The waveform of a sample sad voice, showing amplitude over time.

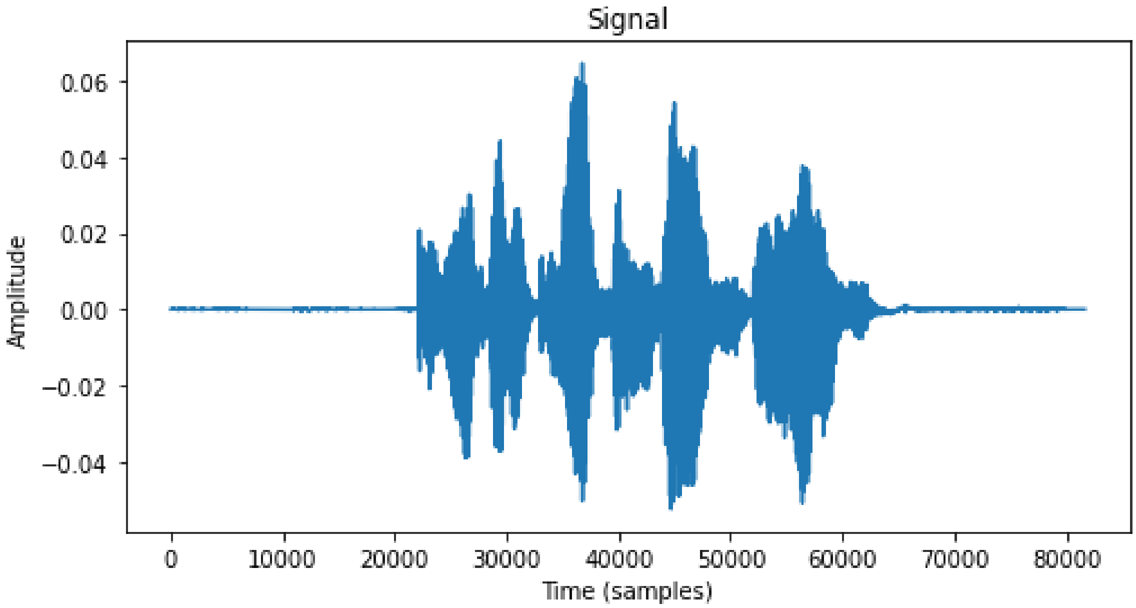


Figure 2.4 The waveform of a sample feared voice, showing amplitude over time.

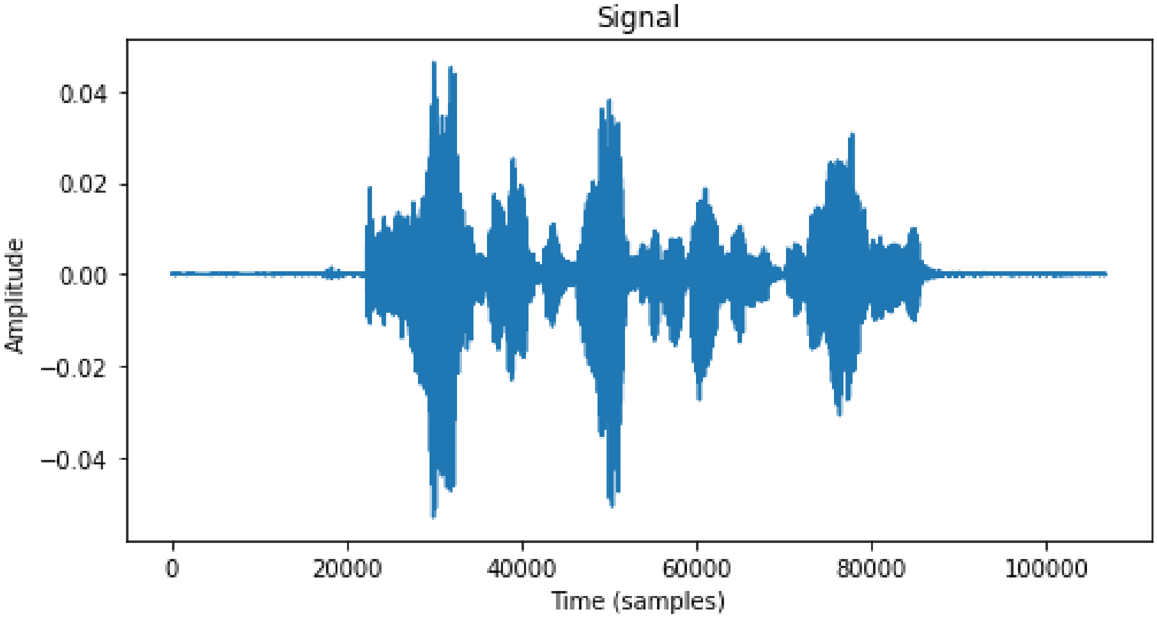


Figure 2.5 The waveform of a sample disgusted voice, showing amplitude over time.

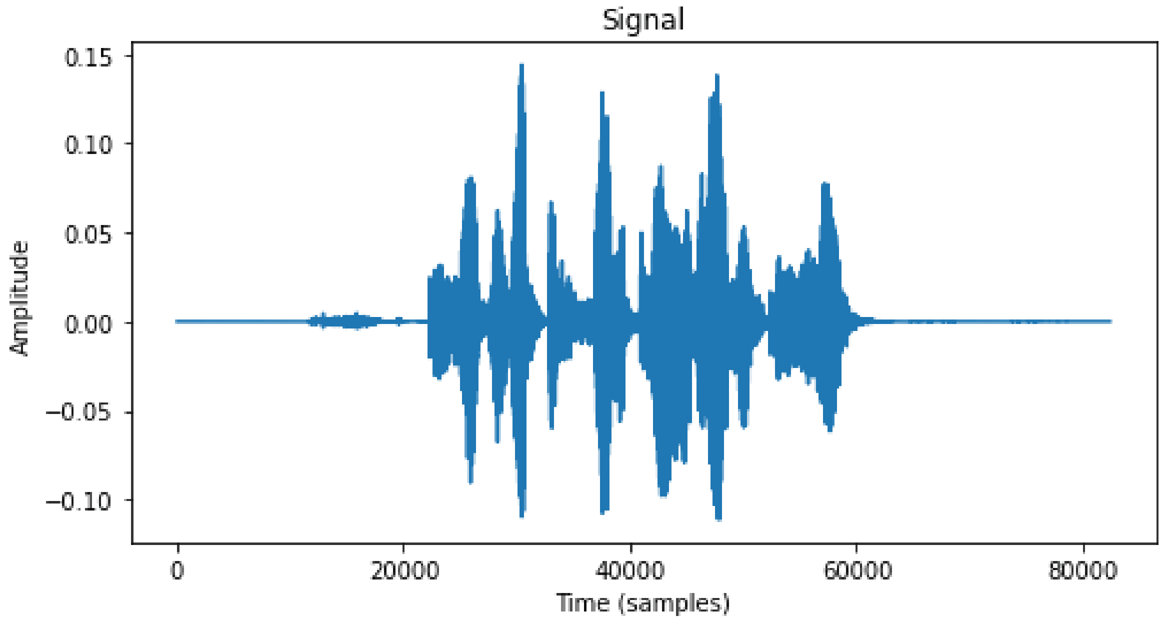


Figure 2.6 The waveform of a sample surprised voice, showing amplitude over time.

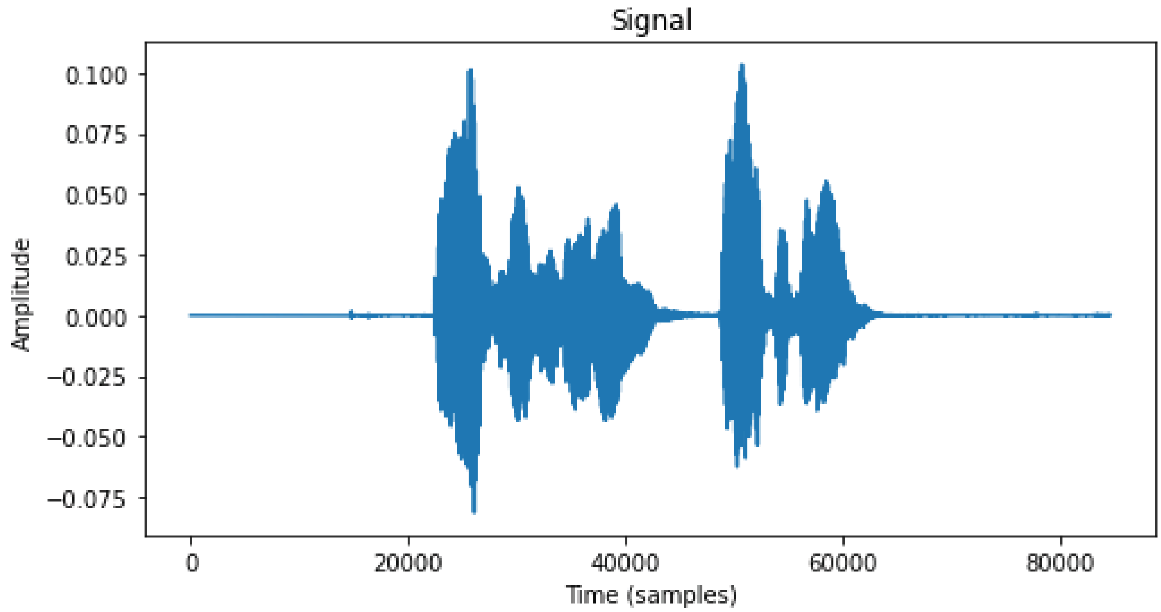


Figure 2.7 The waveform of a sample neutral voice, showing amplitude over time.

Each of the audio signals for these waveforms are composed of several single-frequency sound waves. For extracting important information for emotion recognition from these waveforms, we need to decompose these signals into their individual frequencies and the frequency’s amplitudes. For this, we have used a library named “Librosa”(McFee et al., 2015). There are several commonly used acoustic features for voice emotion recognition that can be calculated for each waveform. The most popular features which are used for emotion recognition are:

Pitch

Pitch is the ear’s perception of the tone height(Davenport & Hannahs, 2020). Pitch is an obvious property of speech, even for non-experts. In a happy voice, pitch is expected to be higher and variant. On the other hand, in a neutral voice, pitch is not much variant. In general, a rise in pitch shows high arousal.

Formant

Formants are concentration of acoustic energy around a particular frequency in the speech wave. Therefore, formants are local maxima in the frequency spectrum caused by resonance during speech production(Davenport & Hannahs, 2020). f_0 is considered as the fundamental frequency and is related to the pitch. The local maxima in the frequency spectrum are known as $f_1, f_2, f_3, \ldots$ which are the formants.

Loudness and Energy

Loudness is about the strength of sound and is perceived by the human ear. Since directly measuring this strength is hard, the energy of the signal is used alternatively. For calculating energy, a Fourier transformation should be done on the original signal.

Mel-Frequency Cepstral Coefficients

Mel-frequency cepstral coefficients (MFCCs) are the other way of representing signals parametrically. For calculating MFCCs, a Mel-scale filter bank should be applied to the Fourier transform of a windowed signal. This should be followed by transforming the logarithmised spectrum into a cepstrum. The amplitudes of this cepstrum are MFCCs. Figures 2.8 and 2.9 show the spectrogram and Mel spectrogram, respectively, for the sample angry voice, presented in figure 2.1. As another example, figures 2.10 and 2.11 show the spectrogram and Mel spectrogram, respectively, for the sample sad voice, presented in figure 2.3. It can be observed how the Mel spectrogram for a sample angry and sad voice differs by looking at figures 2.9 and 2.11.

Duration and Speaking Rate

The duration of speech units such as utterance length and average word length in an utterance together with speaking rate can help in recognizing emotions from speech. Speaking rate can be calculated by counting energy peaks.

Voice Quality

In the context of voice analysis, voice quality is about being sounded breathy, creaky, harsh or whispery. One way to relate these emotions to qualities is to associate breathiness with excitement, harshness with anger and whisper with begin frightened. There are several ways of measuring voice quality. Calculating jitter, shimmer and harmonics-to-noise-ratio (HNR) are most common ways of measuring quality of voice.

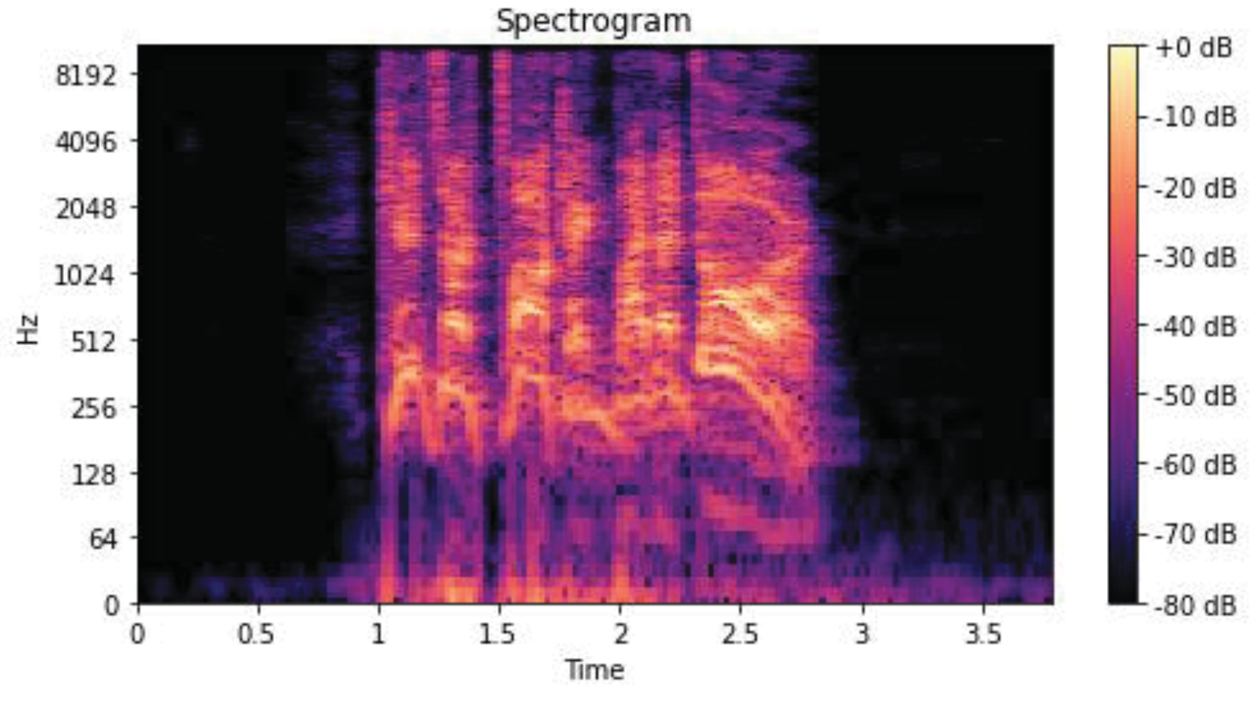


Figure 2.8 The spectrogram for a sample angry voice.

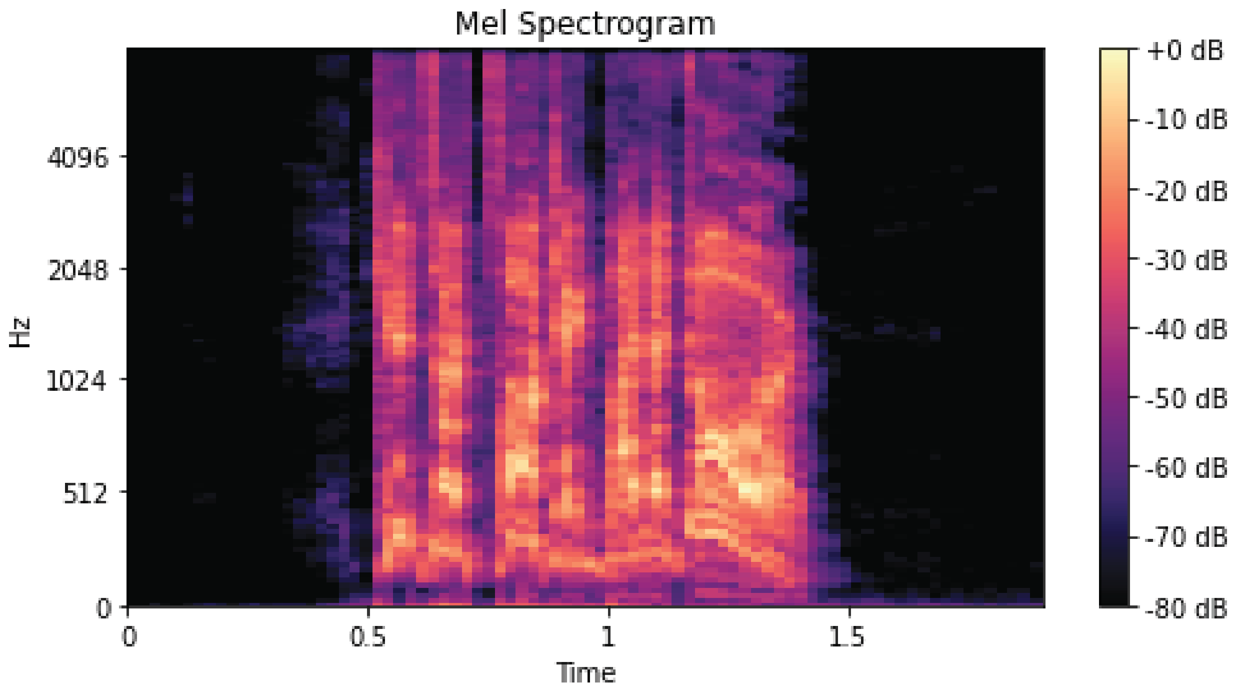


Figure 2.9 The Mel spectrogram for a sample angry voice.

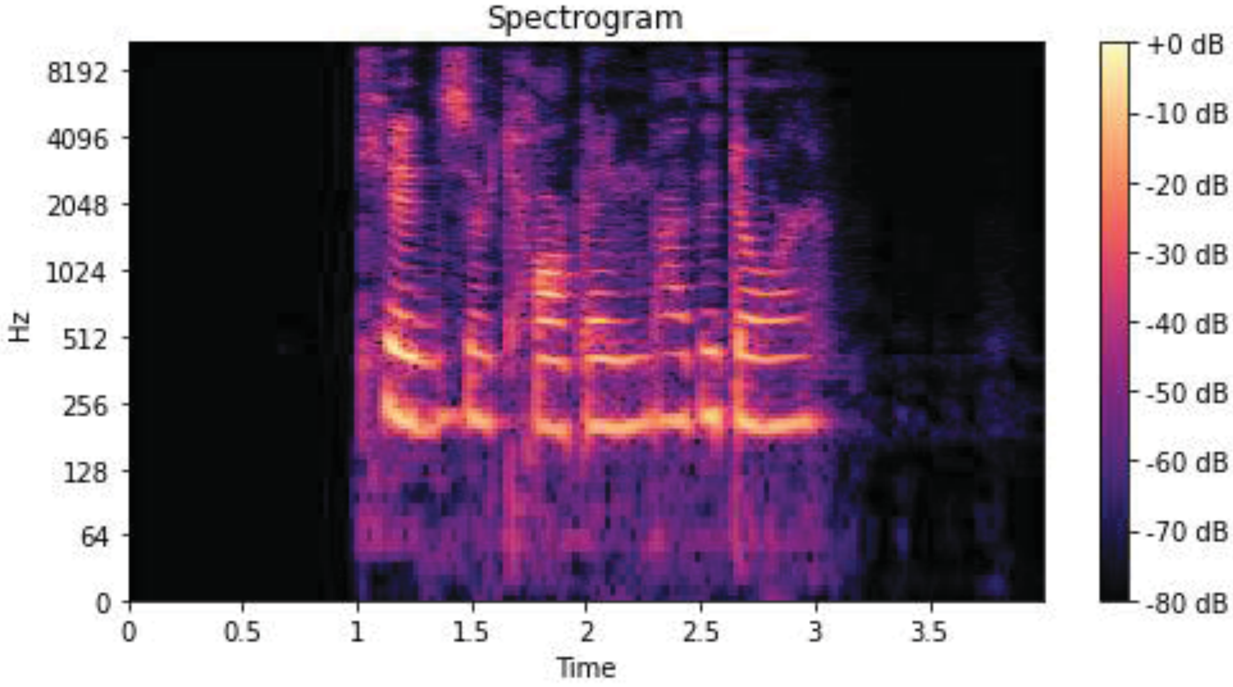


Figure 2.10 The spectrogram for a sample sad voice.

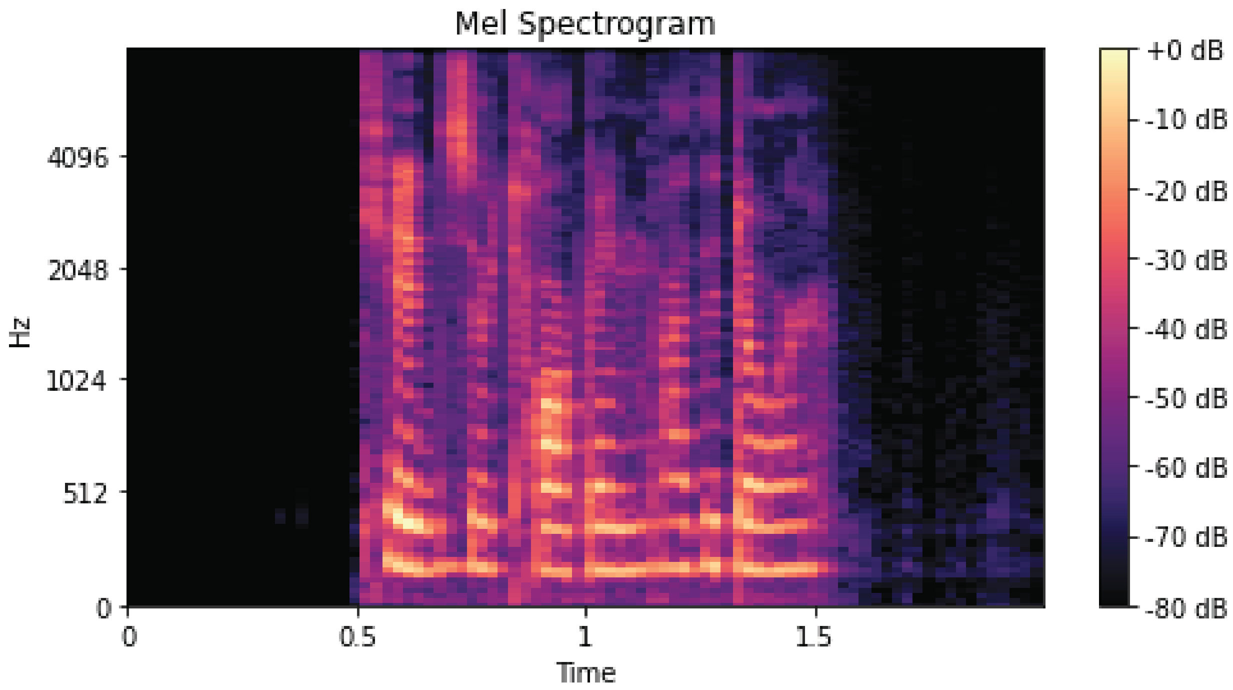


Figure 2.11 The Mel spectrogram for a sample sad voice.

2.2.2 Databases

One of the most important parts in building a voice emotion recognition system is using proper labeled data for training and validating the classifiers. Databases that are used for voice emotion recognition can be divided to three categories:

- Acted or simulated speech emotion databases: In this type of databases, utterances are recorded from professional or semi-professional actors in soundproof and noise free studios. Gathering this kind of database is relatively easier. The disadvantage of using such database is that the emotions may be exaggerated, since usually non-actor people cannot reflect their emotions in their voices as actors. Therefore, relying solely on these kinds of databases may have the potential risk of building voice emotion systems that are not well tuned for recognizing emotions in non-actors' voices. Figure 8 shows an actor while recording the utterances for an acted database in (Burkhardt et al., 2005).
- Elicited or induced speech emotion database: These kinds of databases contain voices that are recorded by placing the speaker in a simulated situation that probably may trigger the desired emotion. In this method, it's possible that the emotion is not

triggered enough in the speaker, but still the exposed emotions in the voices, are more real and more similar to the everyday scenarios.

- Natural speech emotion database: This category of databases contains real recorded voices, from call centers, media recordings such as talk shows or radio talks and other resources. Gathering these kind of databases faces a few challenges; First, the already recorded data needs to be labeled carefully; second, there are usually legal or ethical considerations for using and distributing the real-life recorded data.

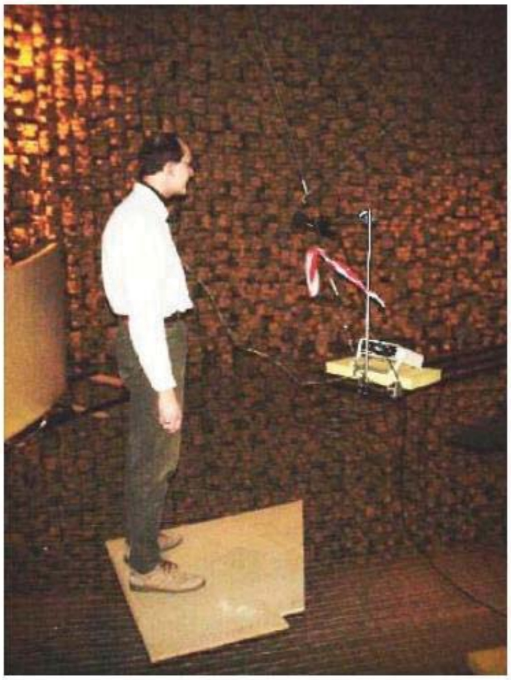


Figure 2.12 One of the actors during the recordings in (Burkhardt et al., 2005)

Based on the target of the analysis, after choosing the appropriate database, there are several other criteria that should be defined, for example the age and gender of speakers. Also, it’s possible to choose different utterances representing different emotions or to repeat the same utterances for different emotions.

Table 2.1 shows several most used speech emotion databases, ordered based on the language. The number of audio files, presented emotions, access type and type of the databases are shown in this table.

Table 2.1 Several commonly used databases in voice emotions recognition systems, ((Akçay & Oğuz, 2020) modified).

Database	Language	Size	Emotions	Access Type	Type
Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)(Livingstone & Russo, 2018)	English	24 professional actors (12 male, 12 female), 7356 speech and song utterances	Anger, disgust, neutral, fear, happiness, sadness, pleasant, surprise	Free	Acted
Toronto Emotional Speech Database (TESS)(Pichora-Fuller & Dupuis, 2020)	English	2 speakers (female), 2800 utterances	Anger, disgust, neutral, fear, happiness, sadness, pleasant, surprise	Free	Acted
Surrey Audio-Visual Expressed Emotion (SAVEE)(Haq & Jackson, 2010)	English	14 speakers (male)x 120 utterances	Anger, disgust, fear, happiness, sadness, surprise, neutral, common	Free	Acted
The Interactive Emotional Dyadic Motion Capture Database (IEMOCAP)(Busso et al., 2008)	English	10 speakers (5 male- 5 female) 1150 utterances	Happiness, anger, sadness, frustration, neutral	Available with license	Acted
Electromagnetic Articulography Database (EMA)(S. Lee et al., 2005)	English	3 speakers (1 male, 2 female) 14 sentences for male, 10 sentences for female	Anger, happiness, sadness, neutral	Free to research use	Acted
eNTERFACE'05 Audio-Visual Emotion Database(Martin et al., 2006)	English	42 speakers (34 male, 8 female) from 14 nationalities, 1116 video sequences	Anger, disgust, fear, happiness, sadness, surprise	Free	Elicited
LDC Emotional Speech	English	7 speakers (4 male,	Hot anger, cold	Commer	Acted

Database(Liberman, Mark et al., 2002)		3 female), 470 utterances	anger, disgust, fear, contempt, happiness, sadness, neutral, panic, pride, despair, elation, interest, shame, boredom	cially available	
Speech Under Simulated and Actual Stress Database (SUSAS)(Hansen & Bou-Ghazale, 1997)	English	32 speakers (19 male, 13 female), 16,000 utterances also include speech of Apache Helicopter pilots	Four states of speech under stress: Neutral, Angry, Loud, and Lombard	Commercially available	Natural Acted
TUM AVIC Database(Schuller et al., 2009)	English	21 speakers (11 male, 10 female), 3901 utterances	Five level of interest; 5 non-linguistic vocalizations (breathing, consent,	Free	Natural
AFEW Database(Kossaifi et al., 2017)	English	330 speakers, 1426 utterances from movies, TV-shows	Anger, disgust, surprise, fear, happiness, neutral, sadness	Free	Natural
SAMAINE Database(McKeown et al., 2012)	English Greek Hebrew	150 speakers, 959 conversations	Valence, activation, power, expectation, overall emotional intensity	Free	Natural
RECOLA Speech Database(Ringeval et al., 2013)	French	46 speakers (19 males, 27 females) 7 h of speech	Five social behaviors (agreement,	Free	Natural

			dominance, engagement, performance, rapport); arousal and valence		
Danish Emotional Speech Database (DES)(Engberg et al., n.d.)	Danish	4 speakers (2 male, 2 female)10 min of speech	Neutral, surprise, anger, happiness, sadness	Free	Acted
Berlin Emotional Database (EmoDB)(Burkhardt et al., 2005)	German	7 Emotions x 10 speakers (5 male, 5 female) x 10 utterances	Anger, boredom, disgust, fear, happiness, sadness, neutral	Open access	Acted
Vera Am Mittag Database (VAM)(Grimm et al., 2008)	German	47 speakers from talk-show, 947 utterances	Valence, activation and dominance	Free	Natural
FAU Aibo Emotion Corpus(Batliner et al., 2008)	German	51 children talking to robot dog Aibo, 9 h of speech	Anger, bored, emphatic, helpless, joyful, motherese, neutral, reprimanding, rest, surprised, touchy	Commercially available	Natural
Italian Emotional Speech Database (EMOVO)(Costantini et al., 2014)	Italian	6 speakers (3 male, 3 female) x 14 sentences x 7 emotions = 588 utterances	Disgust, happiness, fear, anger, surprise, sadness, neutral	Free	Acted
Keio University Japanese Emotional Speech	Japanese	71 speaker (male) 940 utterances	Anger, happiness, disgusting,	Free	Acted

Database (Keio-ESD)(Mori et al., 2006)			downgrading, funny, worried, gentle, relief, indignation, shameful, etc.(47emotions)		
Chinese Emotional Speech Corpus (CASIA)(J. Tao et al., 2008)	Mandarin	6 Emotions x 4 Speakers (2 male, 2 female) x 500 utterances (300 parallel, 200 non-parallel texts)	Surprise, happiness, sadness, anger, fear, neutral	Commercially available	Acted
Beihang University Database of Emotional Speech (BHUDES)(X. Mao et al., 2009)	Mandarin	5 speakers (2 male, 3 female),323 utterances	Anger, happiness, fear, disgust, surprise		Acted
Chinese Annotated Spontaneous Speech corpus (CASS)(Aijun et al., n.d.)	Mandarin	7 speakers (2 male, 5 female),6 h of speech	Anger, fear, happiness, sadness, surprise, neutral	Commercially available	Natural
Chinese Natural Emotional Audio–Visual Database (CHEAVD)(Y. Li et al., 2017)	Mandarin	238 speakers (child to elderly) 140 min emotional segments from movies, TV-shows.	Anger, anxious, disgust, happiness, neutral, sadness, surprise and worried	Free to research use	Acted Natural
Chinese Elderly Emotional Speech Database (EESDB)(K. Wang, 2018)	Mandarin	16 speakers (8 male, 8 female),400 utterances from teleplay	Anger, disgust, fear, happiness, neutral, sadness, surprise	Free to research use	Acted
Turkish Emotional Speech Database	Turkish	582 speakers (394 male,188 female)	Happiness, surprised,	Free to research use	Acted

(TURES)(Oflazoglu & Yildirim, 2013)		from movies, 5100 utterances	sadness, anger, fear, neutral, valence, activation, and dominance		
BAUM-1 Speech Database (Zhalehpour et al., 2017)	Turkish	31 speakers (18 male, 13 female) 288 acted, 1222 spontaneous video clip	Happiness, anger, sadness, disgust, fear, surprise, bothered, boredom, contempt unsure, being thoughtful, concentration, interest	Free to research use	Acted Natural

2.3 Conclusion

In this chapter we discussed and introduced basic building blocks of emotion recognition systems. We went over the definition of emotion and two suggested models of it. We also briefly introduced speech signal characteristics that are widely used in emotion recognition tasks. After these definitions, we introduced several commonly used voice emotion databases, regarding the size of the database, the covered emotions and a few other factors.

In the next chapter, we will review several related studies in the field of speech emotion recognition. We will explain and discuss the most important and related studies, regarding their approaches and results. Also, a wider range of studies in this field, are described at the end of next chapter.

Chapter 3

Voice Emotion Recognition Approaches

Voice emotion recognition has been around for decades(Busso et al., 2004). Researchers have used different machine learning techniques. Feature sets and databases for making accurate voice emotion recognizing systems. Like any other complicated classifying problem, there is not one combination that outperforms all the others in all the scenarios. The studies for building voice emotion detection systems can be divided based on the type of machine learning techniques that they have leveraged. Generally, we can categorize these studies in three groups:

- The studies that have utilized machine learning algorithms.
- The studies that have adopted deep learning algorithms.
- The studies that have applied any other classification algorithms.

In the following, first we will explain several multi-class classification performance measures such as accuracy, precision, recall and Fscore. After that we will discuss most related, important and recent studies in the field of voice emotion recognition in each category. These studies and their results are briefly introduced. Table 10 contains information about several other studies about voice emotion recognition systems. This table shows the datasets, used feature sets and classifiers and the results in each study.

3.1 Performance Measures for Multi-Class Classification

In multi-class classification, assuming we have n datapoints $x_1, \dots, x_n$ and l non-overlapping classes and class labels as $C_1, \dots, C_l$, each datapoint should be classified into exactly one of these classes. For defining performance measures for this type of classification, first we define a few notations that are used for performance evaluation in binary classification. In binary classification the input should be classified into one and only one of two non-overlapping classes C_1 and C_2. Assuming we have two classes of $positive$ and $negative$, we define a few notations as in table 3.1:

Table 3.1 Confusion matrix schema for binary classification

Data Class	Classified as pos	Classified as neg
pos	true positive (tp)	false negative (fn)
neg	false positive (fp)	true negative (tn)

In table 3.1, the first column shows the real class of the data and the second and third columns are about how the binary classifier, predicted the label. tp are the datapoints that were labeled $true$ and were also predicted as $true$ by the classifier. With the same approach, $tn$is used to refer to datapoints that were labeled $negative$ and the classifier predicted them to be $negative$ correctly. fp and fn are used to refer to the datapoints that were wrongly predicted by the classifier as $positive$ and $negative$ respectively.

After explaining these notations, we can define performance measures for binary classification as provided in table 3.2:

Table 3.2 Measures for binary classification using the notation in Table 3.1 (Sokolova & Lapalme, 2009).

Measure	Formula	Evaluation Focus
$Accuracy$	$\frac{tp + tn}{tp + fn + fp + tn}$	Overall effectiveness of a classifier
$Precision$	$\frac{tp}{tp + fp}$	Class agreement of the data labels with the positive labels given by the classifier
$Recall(Sensivity)$	$\frac{tp}{tp + fn}$	Effectiveness of a classifier to identify positive labels
$Fscore$	$\frac{(\beta^2 + 1)tp}{(\beta^2 + 1)tp + \ \beta^2 fn + fp}$	Relations between data’s positive labels and those given by a classifier

Now we are going to generalize the binary classification performance measures to multi-class classification measures. Table 3.3 contains these multi-class classification measures defined based on the notations in table 3.1.

Table 3.3 Measures for multi-class classification based on a generalization of the measures of Table 1 for many classes C_i: tp_i are true positive for C_i, and fp_i – false positive, fn_i– false negative, and tn_i– true negative counts respectively. μ and M indices represent micro- and macro-averaging (Sokolova & Lapalme, 2009).

Measure	Formula	Evaluation Focus
$Average\ Accuracy$	$\frac{\sum_{i=1}^{l}\frac{tp_i+tn_i}{tp_i+fn_i+fp_i+tn_i}}{l}$	The average per-class effectiveness of a classifier
$Error\ Rate$	$\frac{\sum_{i=1}^{l}\frac{fp_i+fn_i}{tp_i+fn_i+fp_i+tn_i}}{l}$	The average per-class classification error
$Precision_\mu$	$\frac{\sum_{i=1}^{l}tp_i}{\sum_{i=1}^{l}(tp_i+\ fp_i)}$	Agreement of the data class labels with those of a classifier if calculated from sums of per-text decisions
$Recall_\mu$	$\frac{\sum_{i=1}^{l}tp_i}{\sum_{i=1}^{l}(tp_i+\ fn_i)}$	Effectiveness of a classifier to identify class labels if calculated from sums of per-text decisions
$Fscore_\mu$	$\frac{(\beta^2+1)Precision_\mu Recall_\mu}{\beta^2 Precision_\mu+Recall_\mu}$	Relations between data's positive labels and those given by a classifier based on sums of per-text decisions
$Precision_M$	$\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fp_i}}{l}$	An average per-class agreement of the data class labels with those of a classifier
$Recall_M$	$\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fn_i}}{l}$	An average per-class effectiveness of a classifier to identify class labels
$Fscore_M$	$\frac{(\beta^2+1)Precision_M Recall_M}{\beta^2 Precision_M+Recall_M}$	Relations between data's positive labels and those given by a classifier based on a per-class average

Throughout the rest of this work, we are going to refer to these performance measures with the same definition as above.

3.2 Machine Learning Techniques

While reviewing related studies in voice emotion recognition systems, it should be noted that in most of these works, the data collection process was done in a studio environment, leading to using clear audio files for training the models.

H. Chen et al in (Chen et al., 2019) used CASIA Chinese Emotional Speech Corpus in training and evaluating several models such as SVM, Random Forest, NN and KNN. Feature sets used in this study are INTERSPEECH2009 and emobase2010 whare are two commonly used feature sets for voice emotion recognition. These feature sets contain 1582 and 384 voice features. Based on this dataset they reported the best features among the features they used, are the MFCC and LogMel FreqBand 0-7 features. After training and testing several classifiers with these feature sets on the CASIA dataset, the authors report SVM as the best model with the highest accuracy of 81.11% being trained on INTERSPEECH2009. It should also be noted that the dataset used in this work, which is CASIA is a perfect dataset, regarding reflecting the emotions by the speakers; therefore, the effect of using non-perfect data is not studied in this work. Table 3.4 compares the results of classifiers being trained with two described feature sets.

Table 3.4 Performance measures of classifiers trained on CASIA using INTERSPEECH2009 in (Chen et al., 2019).

Performance Measure	SVM	Random Forest	NN	KNN
Accuracy – using INTERSPEECH2009	0.8389	0.7056	0.8167	0.6778
Accuracy – using emobase2010	0.8056	0.5556	0.7611	0.6889

(Kwon et al., 2003) used voice features such as pitch, log energy, MFCCs, velocity and acceleration information. They trained several classifiers such as linear SVM(LSVM), binary Gaussian kernel SVM(GSVM) and HMM. Among these classifiers, GSVM and HMM had the best accuracies of 42.3% and 40.8% respectively.

3.3 Deep Learning Techniques

The base element in most of the deep learning algorithms is artificial neural networks. An artificial neural network is considered a deep learning classifier if is composed of many hidden layers. Deep learning techniques have significantly improved the accuracy of classification tasks including speech recognition. Recurrent neural networks (RNN) and convolutional neural networks (CNN) are most commonly used deep learning techniques in speech emotion recognition field(Akçay & Oğuz, 2020).

(Tian et al., 2016) have used a recurrent neural network classifier composed of three hidden layers with LSTM. Each layer is fed with different features from voice signal. First layer receives LLD and eGeMAPS features, the second layer consumes GP and DIS-NV features. Finally, the third layer uses PMI and CSA features. They have used AVEC 2012 and IEMOCAP databases for training and validating classifiers.

(Eyben, Wöllmer, Graves, et al., 2010) have also evaluated using recurrent neural networks with LSTM. First, they have used a three-dimensional emotion model. For voice signal features, they have chosen prosodic, spectral and voice quality features.

3.4 Other Classification algorithms

In (F. Tao et al., 2018) it is aimed to use a corpus that has background noise and is close to the real world. The authors proposed an ensemble framework consisting of 4 sub-systems. They used the Multimodal Emotion Challenge (MEC) 2017 corpus. This corpus contains clips from films and TV programs and the speakers are professional actors; therefore, the problem of having models that are trained with emotion reflective voice, recorded from actors, is remaining.

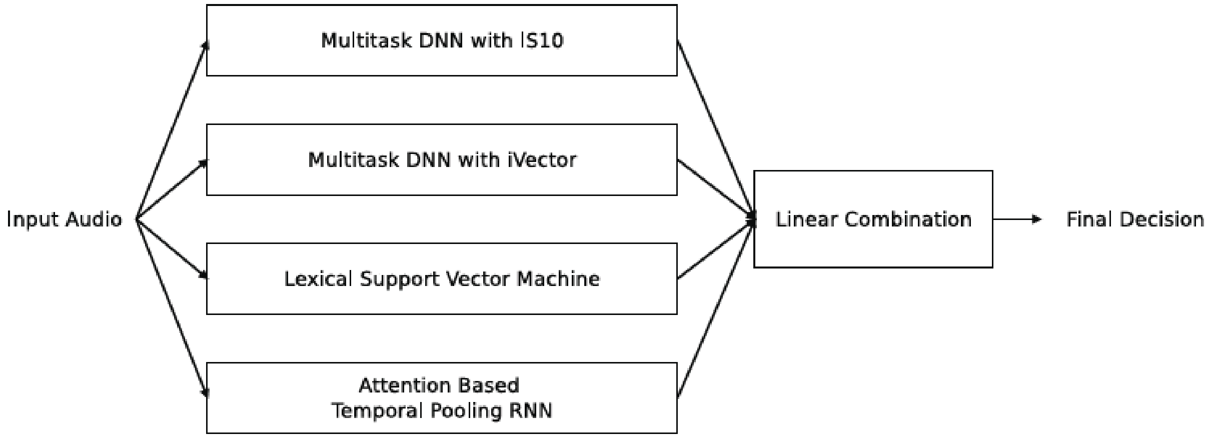


Figure 3.1 Four sub-systems in the framework proposed in (F. Tao et al., 2018)

The feature set in this study is the feature set provided in INTERSPEECH 2010 paralinguistic challenge (Schuller et al., 2010). This feature set is commonly used among voice emotion recognition tasks. As shown in figure 3.1, their proposed approach consists of 4 sub-modules. The result is the linear combination of results of all 4 sub-modules.

In (Shaqra et al., 2019) the relation between age and gender and the emotion recognition accuracy is studied. The authors have used the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) feature set which contains 88 features for energy, frequency, cepstral and dynamic information. The dataset used in this study was RAVDESS which is composed of 7356 audio files, covering eight emotional states of neutral, calm, happy, sad, angry, fearful, disgust and surprised. The audio files are recorded from 12 male and 12 female performers, as explained in table 2.1 in section 2.2.2. The ages of performers range between 21 and 33, with mean of 26.0 and standard deviation of 3.75 years. In this several classifiers using the described dataset and GeMAPS feature set are evaluated. Table shows the average accuracy of these classifiers:

Table 3.5 The average accuracies of the four proposed models in (Shaqra et al., 2019)

Classifier	Average Accuracy
Simple model	64.2
Gender-based model	70.59
Age-based model	65.1
Compound model	74

The first proposed model in this study or the Simple model is a multilayer perceptron neural network (MLP). This MLP has one input layer, two hidden layers and one output layer. The input layer has 88 units which is the number of features. Both hidden layers contain 50 perceptrons. Simple model is fed with RAVDESS data and reaches the accuracy of 64.2% in recognizing one of the eight mentioned emotions.

The second proposed approach in this study is to recognize the gender of the speaker first. Then if the speaker is male, the same input features are fed to a similar MLP (as the Simple model) which is trained only with male voices; On the contrary if the gender classifier recognizes female voice, the feature set is fed to another MLP which is trained solely with female speech files. The average accuracy of this two-level classifier is 70.59%.

The third proposed approach or the age-based model is again a two-level classifier, similar to the gender-based model. The difference is that the first level classifier is an age classifier with two output classes of young speaker (for ages less than 27 years) and old speaker (for ages more than 27 years). Based on the result of this classifier, the 88 speech features are fed to a MLP emotion recognizer (with the same architecture as simple model) of trained with young speakers or a MLP emotion recognizer of solely trained with old speaker's audio files. This two-level age-based classifier reaches the accuracy of 65.1% which is less than the accuracy of a gender-based approach. Authors address this accuracy decrease due to the close range of the actors' ages in RAVDESS dataset.

The fourth proposed model or compound model used the three previous proposed models. The final output in this approach is calculated by doing a soft majority voting over the results of simple, age-based and gender-based approaches. The accuracy of this this approach is 74%. The authors' experiments show that building a separate classifier for each gender and age group gives a better performance rather than having one model for both genders and ages.

S. Yacoub et. al in (Yacoub et al., 2003) have focused on extracting features from short utterances which are commonly used in Interactive Voice Response (IVR) applications. Authors have recognized the anger and neutral emotions from each other with an accuracy of 90% with models trained over the Linguistic Data Consortium at University of Pennsylvania. In our work, we obtained the accuracy of 90% and 87% (without and with our custom dataset respectively) while predicting five emotions at the same time.

Our work's novelty is in gathering and using non-expert and non-actor speaker voices. One of the challenges in gathering such a dataset, was to trigger the emotions in the speakers before recording their voice. We used The International Affective Picture System (IAPS) for triggering the desired emotion in the speakers. IAPS is a standardized database of pictures for studying emotion and attention(Lang et al., 1999). The details about IAPS and the gathered custom dataset are explained in section 4.1.4.

After collecting this dataset, we evaluated several machine learning models, trying to predict 7 emotions of anger, happiness, sadness, fear, disgust, surprise and neutrality at the same time.

After training and testing machine learning classifiers, we aimed to validate their predicted emotions by using electroencephalography. Electroencephalography or EEG is a physiological method to record brain's electrical activity via electrodes that are placed on the scalp surface. EEG is a reliable and cost-effective technology used to measure brain activity and to detect human emotions (Gannouni et al., 2021). One way to validate our classifiers, is to compare the predicted emotion with the emotion that EEG detects. During this experiment, the participants should be wearing the EEG cap, which encompasses the electrodes, while speaking. Their voice should be recorded and fed to the classifiers.

Due to the covid-19 restrictions, while conducting our experiments, we were not able to extend our validation techniques with EEG experiments able to directly provide emotions assessments. After returning to the normal situation, we will resume our validation procedures with EEG experiments.

Table 3.6 Several voice emotion recognition studies, ((Akçay & Oğuz, 2020) modified)

Study	Dataset	Features	Classifier	Results
(Albornoz et al., 2011)	Berlin Emo DB	Mean of log spectrum, MFCC, and prosodic features	Hierarchical classifier using HMM, GMM, and MLP	71.5% average recognition rate
(Bitouk et al., 2010)	LDC, Berlin Emo DB	Spectral features	SVM	46.1% recognition rate for LDC by using group-wise feature selection with class level spectral features, 81.3% recognition rate for EMODB by rank search subset evaluation feature selection with combined class level spectral features and utterance level prosodic features.
(Borchert & Dusterhoft, 2005)	Berlin Emo DB	Formants, spectral energy distribution in different frequency bands, HNR, jitter, and shimmer.	SVM, J48	90% recognition rate for single emotion recognition, 70% for all emotions
(Busso et al., 2009)	EPSAT, EMA, GES, SES, WSJ	Features derived from the F0 contour.	GMM	77% average recognition rate
(Deng et al., 2013)	AVIC, EMODB, eNTERFACE, SUSAS, VAM	LLDs such as ZCR, RMS, energy, pitch frequency, HNR, MFCC	Denoising autoencoder	For AVIC 62.7% recognition rate, for EMODB 57.9%, for eNTERFACE 59.1% for SUSAS 59.5%, for VAM 60.2%

(Deng et al., 2014)	AIBO DB, ABC DB, SUSAS DB	Low-Level Descriptors	Denoising autoencoders and SVM	64.18% average recognition rate for ABC DB, 62.74% average recognition rate for SUSAS DB
(Grimm et al., 2007)	EMA DB, VAM I-II DBs	Pitch related features, speaking rate related features, spectral features	Rule based fuzzy estimator and SVM	0.27, and 0.23 mean errors for VAMI, and VAMII, respectively for both genders. 0.19 mean error for EMA DB.
(Han et al., 2014)	IEMOCAP DB	MFCC features, pitch-based features, and their delta feature across time frames	DNN and Extreme Learning Machine	54,3% average recognition rate
(Hu et al., 2007)	8 native Chinese speakers (4 females and 4 males) uttered each sentence in five simulated emotional states, resulting in 1600 utterances in total.	Spectral features	GMM supervector based SVM	82.5% recognition rate for mixed gender, 91.4% for male,93.6% for female

(Kwon et al., 2003)	SUSAS DB and AIBO DB	Prosodic and spectral features	GSVM and HMM	For SUSAS DB using GSVM 90% and 92.2% recognition rates are obtained for neutral and stress speech, respectively. Using HMM 96.3% recognition rate is obtained. Recognition rate is 70.1% for 4-class style classification with HMM. For multiclass classification on AIBO DB, GSVM achieved an average recognition rate of 42.3%. The average recognition rate is 40.8% using HMM.
(C.-C. Lee et al., 2011)	AIBO DBUSC IEMOCAP DB.	ZCR, root mean square energy, pitch, harmonics-to-noise ratio, and 12 MFCCs and their deltas.	Decision tree	For AIBO, 48.37% using leave-one speaker out cross validation on the training dataset. For IEMOCAP, average unweighted recall of 58.46% using leave-one speaker out cross-validation
(Luengo et al., 2005)	Emotional speech database for Basque, recorded by the University of the Basque Country, with single actress	Prosodic and spectral features	SVM, GMM	98% accuracy for GMM-MFCC, 92.32% with SVM & prosodic features, 86.71% with GMM& prosodic feature

(Mirsamadi et al., 2017)	IEMOCAP corpus	Automatically learned by Deep RNN, as well as hand-crafted LLDs consisting of F0, voicing probability, frame energy, ZCR, and MFCC	Deep RNN	Proposed system with raw spectral features has 61.8% recognition rate Proposed system with LLD features have 63.5% recognition rate
(Q. Mao et al., 2014)	SAVEE DB, Berlin EMO DB, DES DB, MES DB	Automatically learned by CNN	CNN	73.6% accuracy for SAVEE DB,85.2% for EMODB,79.9% for DES DB 78.3% for MES DB
(Nakatsu et al., 1999)	100 utterance, 50 males, 50 females.	Speech power, pitch, LPC	Neural networks	50% recognition rate
(Nogueiras et al., 2001)	Spanish corpus of INTERFACE, Emotional Speech Synthesis Database	Prosodic and spectral features	HMM	Recognition rate higher than 70%for all emotions
(Nwe et al., 2003)	3 female, 3 male for Burmese language, 3 female, 3 males for Mandarin language	LFPC	HMM	Average recognition rates of classification for the Burmese and the Mandarin utterances are 78.5% and 75.7%, respectively.

(Rao et al., 2013)	Telugu emotion speech corpus	Prosodic features	SVM	66% average recognition rate with sentence level prosodic features,65.38% average recognition rate with word level prosodic features, 63% average recognition rate with syllable level prosodic features
Rong et al. (2009)	One natural and one acted speech corpora in Mandarin	Pitch, energy, ZCR, and spectral features	kNN	66.24% average recognition rate with all 84 features, 61.18% with PCA/MDS, 60.40% with ISOMAP, and 69.21% with proposed ERFTrees method
Sato and Obuchi (2007)	Database from Linguistic Data Consortium	MFCC	HMM	66.4% recognition rate
Schuller et al. (2003)	German and English 5 speaker 5250 sample. Acted and natural data	Energy and Pitch based features	continuous HMM	86.8% average recognition rate with global prosodic features and 77.8% average recognition rate for instantaneous features
Schuller et al. (2005b)	3947 movie and automotive interaction dialog-turns database consisting of 35 speakers.	Pitch, energy, and duration related features	StackingCSVM NB C4.5 kNN	63.51% recognition rate for 276dimensional features and 71.62% for 100 dimensional features

Schuller et al. (2005a)	Berlin EmoDb	The raw contours of ZCR, pitch, first seven formants, energy, spectral development, and HNR and linguistic features	StackingC MLR NB 1NN SVM C4.5	76.23% recognition rate with all 276 features, 80.53% with top 75 features selected by SVM SFFS
Schuller (2011)	VAM DB	Low level descriptors such as signal contour, spectral pitch, formants, HNR, MFCCs, or energy of the signal and linguistic features	Support Vector Regression	Best result for Valence dimension is 66.7% using linguistic features. For Activation dimension 85.1% recognition rate with acoustic and bag of n-grams features. For Dominance acoustic and bag of character n-grams features recognition rate is 82.5%
Shen et al. (2011)	Berlin Emo DB	Energy, pitch, LPCC, MFCC, LPCMCC	SVM	Best results with energy and pitch features are 66.02%, 70.7% for only LPCMCC features, and 82.5% for using both.
Trigeorgis et al. (2016)	RECOLA DB.	Automatically learned by deep CNN	Deep CNN with LSTM	MSE in arousal dimension is 0.684, and MSE in valence domain is 0.261
Ververidis and Kotropoulos (2005)	1300 utterances from DES	Statistical properties of formants, pitch, and energy contours of the speech signal	GMM	48.5% recognition rate for GMM with one Gaussian density, 56% for males and 50.9% for females

Wang et al. (2015)	Berlin EMO DB, CASIA DB, Chinese elderly emotion database (EESDB)	Fourier Parameters, MFCC	SVM	For EMO DB 88.88% recognition rate, while for CASIA DB 79% recognition rate, and for EESDB 76% recognition rate
Wollmer et al. (2010)	Sensitive Artificial Listener (SAL) database	Low Level audio features such as pitch, MFCC, energy, HNR and linguistic features	BL STM, L STM, SVM, and conventional RNN	Quadrant prediction F1-measure of up to 51.3%,
Wu and Liang (2011)	Two corpora; corpora A and B consist of the utterances from six and two volunteers, total 2033 sentences	Pitch, intensity, formants1-4 and formant bandwidths1-4, four types of jitter-related features, six types of shimmer-related features, three types of harmonicity-related features, MFCCs	Meta Decision Tree (MDT) containing SVM, GMM, MLP classifiers	80% recognition rate with mixed utterances from corpora A and B.
Wu et al. (2011)	Berlin Emo DB, VAM DB	Prosodic features, speaking rate features, ZCR and TEO based features	SVM	91.3% by proposed modulation spectral features and prosodic features for EMODB, 86% by prosodic and spectral modulation features for VAM DB
Yang and Lugger (2010)	Berlin Emo DB	Prosodic, spectral and voice quality features	Bayesian classifier	73.5% average recognition rate

Zhang et al. (2011)	ABC, AVIC, DES, eNTERFACE, SAL, and VAM.	LLDs such as energy, pitch, voice quality, spectral, MFCC features	Unsupervised Learning	Mean unweighted recognition rate is 66.8% using Z-normalization of arousal classification, 58.2% for valence classification with centering normalization on cross-corpus emotion recognition

3.5 Conclusion

In this chapter, we reviewed several related studies in the field of voice or speech emotion detection. We explained several more related studies, considering their approach, used datasets, selected feature sets, leveraged machine learning or deep learning techniques and their results. Considering these related studies, we decided to target the need for recognizing the emotions when the speaker is not as capable as an actor in reflecting their emotions. Having this requirement, we gathered a dataset recorded from non-actor people while trying to trigger the emotions in them.

In the next chapters we will explain the approaches and methodologies that we followed for providing answers to our research questions and hypotheses that were mentioned in chapter 1.

Chapter 4

Recognizing Emotions in Voice Using Machine Learning Classifiers

In this chapter we are going to explain our first approach that we followed for recognizing the emotions in speech. First, we provide an abstract view of the whole process. Then we try to get deeper in each part and explain the steps in detail. The final section in this chapter contains the results that we gained by following the mentioned approach.

For classifying the emotions in the speech signals, like any other classifying task, first we must gather a rich dataset, containing labeled speech audio files. After that, we must decide about the set of features that will best suit our needs. When the features were ready, we have to train and test several machine learning models. This last step requires tunning the models and exploring the best hyperparameters. After having trained and evaluated models, we can embed them in any voice emotion recognition system.

In our work we used the databases that are explained in detail in next section. After gathering this rich dataset from mentioned emotional databases, we extracted appropriate features from the audio files. The extracted feature set was fed into several machine learning classifiers separately. For validating the performance of each classifier, the input data set was split into training and validation sets under two different approaches. In the first approach, 25% of the input data set was randomly chosen as the validation set.

4.1 Datasets

In all machine learning applications, selecting the proper dataset is extremely important. There are many different datasets for voice emotion recognition. These datasets were introduced in chapter 2, section 2.2.2. In our work, the following well-known and commonly used datasets are used for training and validating the models.

4.1.1 Toronto Emotional Speech Set (TESS)

This dataset contains 2800 audio files, each of them reflecting one of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness and neutral). In each file sentence “Say the word _” is filled with one word from a set of 200 target words. Speakers are two actresses who are 24 and 26 years old at the time of recording.

4.1.2 Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

This dataset contains 7356 speech and song files. The 7 emotions reflected in speech files are calmness, happiness, sadness, anger, neutrality, fear, surprise and disgust. Speakers are 24 professional actors (12 male, 12 female) vocalizing two lexically matched statements in a neutral North American accent.

4.1.3 Surrey Audio-Visual Expressed Emotion (SAVEE) Database

This database consists of 480 audio files. Speakers are 4 male actors and the 7 concluding emotions are anger, disgust, fear, happiness, sadness, surprise and neutrality. Here only 5 emotions of anger, happiness, sadness, surprise and neutrality were considered; Also, the spoken sentence consisted of 15 sentences per emotion, 2 common, 2 emotion specific and 10 generic sentences, which are phonetically balanced and different for each emotion.

4.1.4 Custom Database

This is the custom dataset that we gathered for this study. The reason for building this dataset is that usually the emotions in non-actor people's voices are not as clearly reflected as actor's voices. Therefore, we needed a database composed of non-actor and non-expert peoples’ voices trying to reflect different emotions.

In this dataset audio files were recorded from non-actor speakers. There were 4 speakers, 2 males and 2 females, with an average age of 30. Before recording the audio files from speakers, we needed to trigger a desired emotion in them. We used International Affective Picture System (IAPS) for triggering the emotions. IAPS is a database of pictures designed to provide a standardized set of pictures for studying emotion and attention(Lang et al., 1999). IAPS was developed by National Institute of Mental Health Center for Emotion and Attention at the University of Florida. IAPS is used and validated for eliciting a specific emotional response in viewers. For each emotion, our speakers were asked to look at a set of images and then to read out loud several sentences. Finally, we could collect 25 audio files for each of the 7 emotions, which led to a total of 175 audio files. In this study, we didn't aim to investigate the relationship between age and gender of the speakers with the performance measures of the voice emotion recognition system; Therefore, we followed a gender-balanced approach in gathering the data.

Gathering this dataset faced a few main challenges: first, sometimes even with looking at some chosen IAPS images, the desired emotion was not stimulated in the participant. Second, these audio files should be reviewed by some feeling experts validating the reflected emotion.

Table 4.1 shows the distribution of audio files related to each emotion in the final dataset which is the result of appending all the above four datasets:

Table 4.1 Distribution of audio files related to each emotion in the whole dataset

Emotion	Final Dataset	
	Count	Proportion
Angry	677	0.144
Happy	677	0.144
Neutral	641	0.136
Surprised	677	0.144
Sad	677	0.144
Fear	677	0.144
Disgust	677	0.144
Total	4703	1.0

4.2 Feature Sets

Two well-known and commonly used feature sets for voice emotion recognition tasks are used as our feature sets. In the following sections, we briefly introduce these features sets and the tool for extracting them from voice audio files.

We have used two feature sets provided in INTERSPEECH 2010 paralinguistic challenge and in the INTERSPEECH 2013 computational paralinguistics challenge. Both feature sets are commonly used feature sets among related works. The INTERSPEECH 2010 paralinguistic challenge contains 1582 audio features such as MFCC, PCM loudness, LPC coefficients, jitter, shimmer and so on. The 2013 COMPARE feature set provided in The INTERSPEECH 2013 computational paralinguistic challenge is also composed of 6373 features such as energy, spectral, cepstral (MFCC) and voicing related low-level descriptors (LLDs) as well as a few LLDs including logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral sharpness. We used a tool named openSMILE (Eyben, Wöllmer, & Schuller, 2010) for extracting these feature sets from our audio dataset.

4.3 Machine Learning Models

Gradient boosting: This machine learning algorithm is a powerful technique that uses some weak classifiers, for example, decision trees, to build strong classifiers. Weak classifiers are those classifiers that their prediction is at least slightly better than the random prediction. Gradient boosting leverages gradients when calculating the loss for identifying the weak classifiers. Before explaining this classifier in more detail, we should define a concept in machine learning called Boosting. The idea behind boosting is to build a model on the training dataset, then a second model is built for fixing the errors of the first model. Assuming we have N data points $\{(x_1, y_1), \dots, (x_N, y_N)\}$, where $x_N \in \mathbb{R}^D$ and D is the dimension of the feature vector and $\{+1, -1\}$ are the possible values for the output class. Initially an equal weight w_i, where $i = 1, 2, \dots, N$ is assigned to each data point. First a model called M_1 is built with a random subset of the training dataset. Since all the points have equal weights initially, there is an equal probability for each of them to be selected as the training set of M_1. After getting the predicted classes from M_1, the weights should be updated; in a way that the weight of the correctly predicted points should decrease, while the weight for wrongly predicted points should increase. Therefore, the previously misclassified points by M_1, have a better chance to be selected as the training set of the next classifier, which is M_2. This procedure is repeated until and unless the errors are minimized or reached an acceptable level and the dataset is classified correctly. Considering this, gradient boosting tries to build models sequentially and to reduce the errors in subsequent models. The weak learners are decision trees in gradient boosting. Gradient boosting can be used for both regression and classification tasks. The difference in using this technique regarding regression or classification is the loss function. The loss function should be differentiable. In both cases, the objective is to minimize the loss function by adding weak learners using gradient descent. Gradient boosting is a greedy algorithm and may overfit a training dataset quickly.

Bagging: Bootstrap Aggregation or shortly Bagging is a powerful ensemble method. Before getting more detailed about bagging or ensemble methods, we need to explain the bootstrap which lies at the heart of bagging. The Bootstrap is a statistical method for estimating a quantity based on a data sample. We can clarify this with an example. Assume we have a sample dataset of 1000 values and we want to have an estimate of the mean of the sample. We know we can directly calculate this mean:

$$mean(x) = sum(x)/1000$$

Of course, this sample is small; therefore, the mean we calculated my not properly reflect the mean of the real data. We can improve this estimation by using the bootstrap. For that we need to create many random subsamples of our current sample with replacement. Then we calculate the mean of each of these subsamples. The last step is to calculate the mean of the calculated means in the previous step. This process can also be used to reach an estimation for other descriptive statistics like standard deviation or for the quantities used in machine learning algorithms, for example learned coefficients.

After this short introduction about bootstrap we can get back to bootstrap aggregation or bagging which is a powerful ensemble technique. An ensemble method combines predictions from several separate machine learning algorithms to reach a better accuracy than each of them alone. Bootstrap Aggregation ensembles the results of several separate decision trees. Decision tree is an algorithm with high variance and sensitive to the training data. Bagging uses the Bootstrap to handle this high variance among its inner building blocks, which are decision tress. Assuming that the training data is composed of 1000 datapoints, bagging creates many random subsamples of the training data with replacement. Then a decision tree is trained by each subsample. For predicting the result on a new dataset, bagging get the prediction from each of previously trained decision trees for this new dataset. The final prediction can be the result of applying majority voting or averaging on the results of the base classifiers. Therefore, the effect of each tree overfitting the training data, is reduced. One issue to note in this process is that in bagging, each random subset of the input data set is drawn with replacement; therefore, at the end many original samples of the data set may be repeated in the training process, while some may be left out.

Random Forest: Random Forest is an ensemble learning method that builds several decision trees and does a majority voting for choosing the predicted category in a classification task.

Random forest tries to improve the accuracy of prediction compared to bagging, by building the inner decision trees less correlated to each other. For explaining this, first we should note that decision trees are greedy; that is, they choose the variable to split on by using a greedy algorithm for minimizing the error. Therefore, it's possible to end up with several decision trees that have many structural similarities. These structural similarities lead to high correlation in the predicted results. We explained previously that bagging uses these decision trees for building an ensemble classifier. Ensemble classifiers have better performance with uncorrelated or weakly correlated inner blocks. Random Forest tries to address this issue by limiting the learning procedure to a randomly selected sample of features for selecting a split point. Therefore, the split point is not the most-optimal split point and this is not the same among all the decision trees anymore. With this tweak for selecting the split point, the correlation among the prediction results of decision trees is decreased and usually better performance than bagged decision trees is achieved. The number of randomly selected features when splitting a point is a parameter for random forest which can be tunned using cross validation.

Support vector machines: This supervised machine learning technique tries to do the classification task using a hyperplane or a set of hyperplanes in a high dimensional space. Support Vector Machines or SVM is basically a binary classifier. The predictor function can be defined as $f\colon \mathbb{R}^D \rightarrow \{+1, -1\}$, where f is the predictor function, D is the dimension of the feature vector and $\{+1, -1\}$ are the possible values for the output. Now if the training dataset is composed of N data points $\{(x_1, y_1), \dots, (x_N, y_N)\}$, where $x_N \in \mathbb{R}^D$ and $y_N \in \{-1, +1\}$ and $n = 1, 2, \dots, N$, the problem is defined to find the predictor function f with the least classification error. In a binary classification problem, this function can be described as a linear model:

$$f(x,w) = w^T x + b$$

Where $w \in \mathbb{R}^D$ is the weight vector and $b \in \mathbb{R}$ is the bias. If the dataset is linearly separable in the feature space, then the objective is to find a separating hyperplane that maximizes the margin between the positive class and negative class points, that is $w^T x_n + b \geq 0$ where $y_n = +1$ and $w^T x_n + b < 0$ where $y_n = -1$. On the other hand, if the dataset is not linearly separable, a kernel function is used to map the original data points to a new space and then the SVM tries to find the separating hyperplane for the transformed data. The kernel used by SVM can be linear on non-linear.

All of this is said for binary classification. There are several methods for combining multiple binary SVMs to build a multiclass classifier. One common approach is known as one-versus-the-rest (Vapnik et al., 1998). In this approach we need several separate SVMs to the number of the output classes. Therefore, if we have K output classes, which is in our case emotion classes, the kth model is trained with the data from class C_k as the positive points in the binary example and the data from the remaining $k-1$ classes as the negative points. In another approach called one-versus-one, all possible pairs of classed are trained in several separate $K(K-1)/2$ SVMs.

SVM is commonly used in voice emotion detection. Of course, one of the challenges when using SVM is to choose the proper kernel functions.

4.4 Training and Validation

For validating the models, we followed two different approaches, each of them using parts of or the whole of our final dataset; in the first approach, the leveraged dataset contained TESS, RAVDESS and SAVEE datasets, which are all common speech emotion recognition datasets, recorded by actor speakers. In the second case, we added our custom dataset to the previous datasets. The goal of obtaining results in these two different cases is to evaluate the effect of using our custom dataset on classifiers' performance measures.

For tunning hyper parameters for each model, we have used grid search. In grid search several possible values are defined for each hyper parameter. Then models for all the possible combinations of hyper parameters are built. We have used GridSearchCV from the library Scikit-learn for implementing the grid search for each model. The tuned hyper parameters are reported for each classifier in the following parts.

After tunning hyper parameters, for evaluating the prediction results of classifiers trained with these two datasets, we have used 5-folds cross fold validation. In this technique, the whole dataset gets split into training and test sets. In 5-fold cross validation, each time, the whole dataset is divided to 5 parts, one part is considered as the test set and the other four parts together form the training set. This process repeats 5 times until each part was once considered as the test set. Here we have used a special type of cross fold validation that keeps the same distribution of classes or labels in each round of validation. For implementing this kind of validation, we used StratifiedKFold class from Sklearn library.

The predicted results of classifiers trained and tested with these two approaches are explained in the following parts.

4.4.1 Considering TESS, RAVDESS and SAVEE datasets

In this case, classifiers were trained with TESS, RAVDESS and SAVEE dataset. The total number of audio files are 3625 files. In each round of 5-fold cross validation, 20% of the whole dataset is considered as the test set, which leads to separate 2900 audio files as the training set and the remaining 725 files as the test set. The distribution of audio files related to each emotion, after splitting training and test sets, is shown in Table 4.2.

Table 4.2 Distribution of audio files related to each emotion in training and validation set in the first approach.

Emotion	Training		Testing	
	Count	Proportion	count	Proportion
Angry	522	0.144	130	0.144
Happy	522	0.144	130	0.144
Neutral	493	0.136	123	0.136
Surprised	522	0.144	130	0.144
Sad	522	0.144	130	0.144
Fear	522	0.144	130	0.144
Disgust	522	0.144	130	0.144
Total	3625	1.00	903	1.00

4.4.2 Considering TESS, RAVDESS, SAVEE and custom datasets

In the second case, the final dataset is composed of our custom dataset together with three previously mentioned datasets. In this approach, the total number of audio files reached 3765 files. Again, we validated our models using 5-fold cross validation. The distribution of audio files related to each emotion, in the training and test sets, are shown in Table 4.3.

Table 4.3 Distribution of audio files related to each emotion in training and validation set in the second approach

Emotion	Training		Testing	
	Count	Proportion	Count	Proportion
Angry	542	0.144	135	0.144
Happy	542	0.144	135	0.144
Neutral	513	0.136	128	0.136
Surprised	542	0.144	135	0.144
Sad	542	0.144	135	0.144
Fear	542	0.144	135	0.144
Disgust	542	0.144	135	0.144
Total	3765	1.00	938	1.00

4.5 Results And Discussion

In this section, several performance measure of different classifiers, trained and tested under two described approaches are reposted.

4.5.1 Considering TESS, RAVDESS and SAVEE datasets

Since there are two separate well-known and commonly used feature sets for voice emotion recognition tasks, we have repeated the process of training and validating our models, first while feeding the classifiers with INTERSPEECH 2010 feature set and the other time, using INTERSPEECH 2013 paralinguistic challenge. In the following sections, the results of each experiment are described separately.

First, we are going to reports the results that are obtained by evaluating the models that are trained by INTERSPEECH 2010 feature set. All these classifiers are implemented by using the Scikit-learn library. The tuned hyper parameters for each classifier are also obtained by GridSearchCV from Scikit-learn library:

- Support Vector Machine: {'kernel': 'poly', 'gamma': 0.001, 'C': 10}
- Random Forest: {'n_estimators': 100, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_features': 0.5, 'max_depth': 7}
- Gradient Boosting: {'subsample': 1, 'n_estimators': 70, 'min_samples_split': 2, 'min_samples_leaf': 0.2, 'max_features': None, 'max_depth': 7, 'learning_rate': 0.3}
- Bagging: {'n_estimators': 60, 'max_samples': 1.0, 'max_features': 0.5}

Table 4.4 contains the average performance measures of each classifier on the test set. The results for each performance measure are the averages over the results of that classifier over all the 5 repetitions of 5-old cross validations. As it can be seen, gradient boosting and bagging have the best average accuracies of 85.67% and 85.27% respectively. These classifiers have also the overall best performance measures among all the classifiers which shows that we can rely on ensemble methods for speech emotion prediction applications.

Table 4.4 Average performance measures of each classifier on the test set

Classifier	Feature Set	Accuracy	F1-score	Precision	Recall
SVM	INTERSPEECH 2010	0.7469	0.7466	0.7486	0.7469
Random Forest	INTERSPEECH 2010	0.7723	0.7741	0.7869	0.7723
Gradient Boosting	INTERSPEECH 2010	0.8567	0.8564	0.8577	0.8567
Bagging	INTERSPEECH 2010	0.8527	0.8525	0.8541	0.8527

Figure 4.1 shows a comparison of the accuracies of the classifiers on the test set.

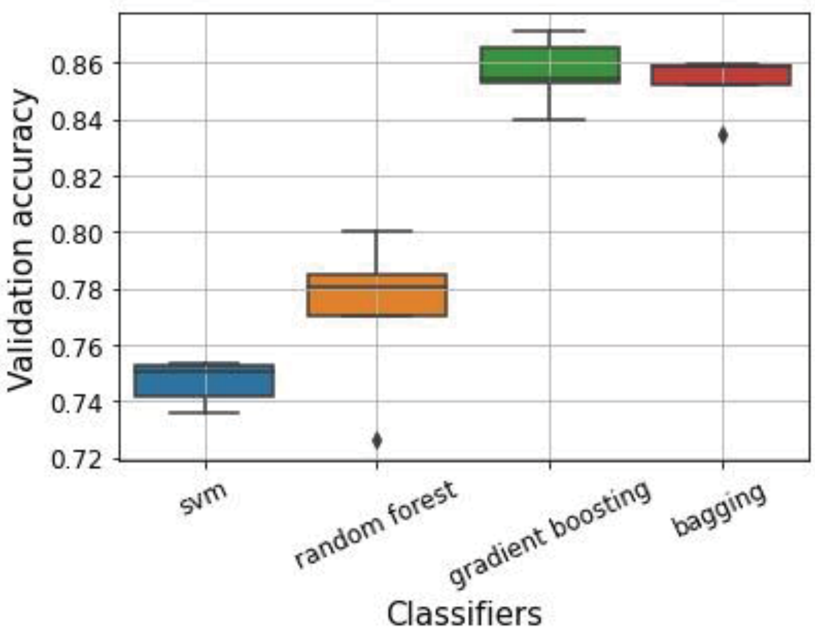
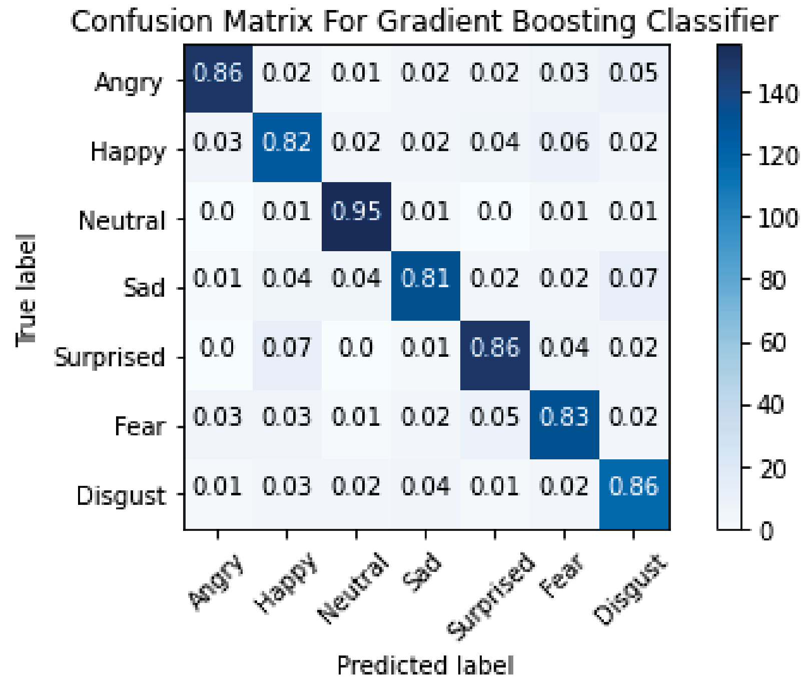


Figure 4.1 Validation accuracies of classifiers in the first approach.



	Angry	Happy	Neutral	Sad	Surprised	Fear	Disgust
Angry	0.86	0.02	0.01	0.02	0.02	0.03	0.05
Happy	0.03	0.82	0.02	0.02	0.04	0.06	0.02
Neutral	0.0	0.01	0.95	0.01	0.0	0.01	0.01
Sad	0.01	0.04	0.04	0.81	0.02	0.02	0.07
Surprised	0.0	0.07	0.0	0.01	0.86	0.04	0.02
Fear	0.03	0.03	0.01	0.02	0.05	0.83	0.02
Disgust	0.01	0.03	0.02	0.04	0.01	0.02	0.86

Figure 4.2 Confusion matrix of gradient boosting classifier

.

We also trained and tested all our classifiers, using INTERSPEECH 2013 feature set. Again, all these classifiers and the tuned hyper parameters are implemented by using the Scikit-learn library. The tuned hyper parameters by GridSearchCV for each classifier are:

- Support Vector Machine: {'kernel': 'poly', 'gamma': 0.001, 'C': 10}
- Random Forest: {'n_estimators': 100, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_features': 0.5, 'max_depth': 7}
- Gradient Boosting: {'subsample': 1, 'n_estimators': 70, 'min_samples_split': 2, 'min_samples_leaf': 0.2, 'max_features': None, 'max_depth': 7, 'learning_rate': 0.3}
- Bagging: {'n_estimators': 60, 'max_samples': 1.0, 'max_features': 0.2}

Table 4.5 contains the average performance measures of each classifier on the test set, while using INTERSPEECH 2013 as the feature set. The results for each performance measure are the averages over the results of that classifier over all the 5 repetitions of 5-fold cross validations. As it can be seen, gradient boosting, SVM and bagging have the best average accuracies of 85.62%, 85.18% and 84.17% respectively. These classifiers have also the overall best performance measures among all the classifiers which again proves that we can rely on ensemble methods for speech emotion prediction applications.

Table 4.5 Average performance measures of each classifier on the test set.

Classifier	Feature Set	Accuracy	F1-score	Precision	Recall
SVM	INTERSPEECH 2013	0.8518	0.8519	0.8543	0.8518
Random Forest	INTERSPEECH 2013	0.7845	0.7851	0.7980	0.7845
Gradient Boosting	INTERSPEECH 2013	0.8562	0.8557	0.8571	0.8562
Bagging	INTERSPEECH 2013	0.8417	0.8411	0.8427	0.8416

Figure 4.3 shows a comparison of the accuracies of the classifiers on the test set.

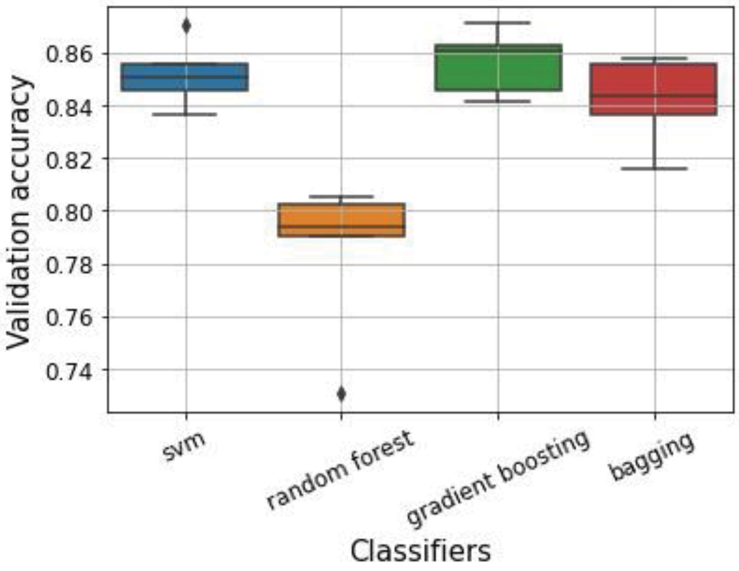


Figure 4.3 Validation accuracies of classifiers in the first approach

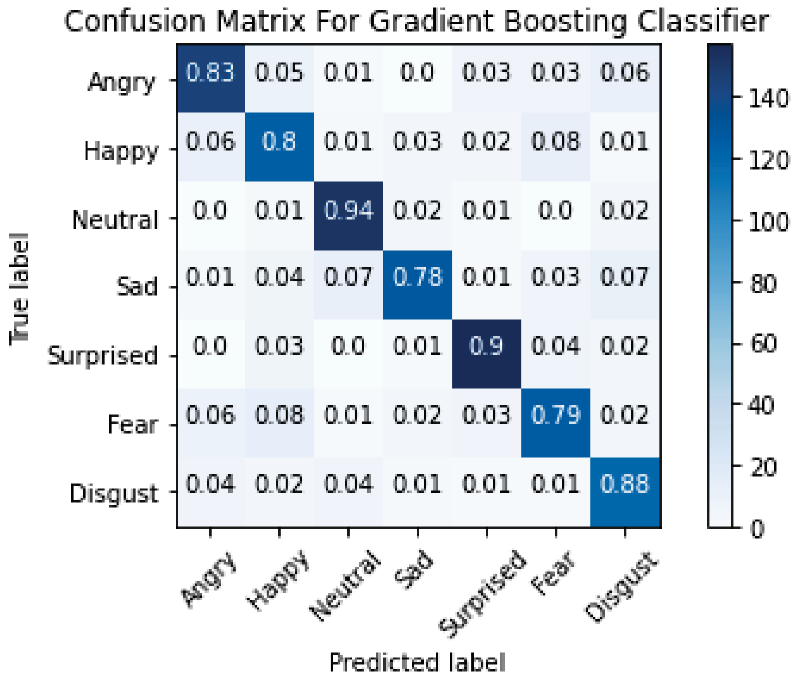


Figure 4.4 The confusion matrix of gradient booting classifier.

4.5.2 Considering TESS, RAVDESS, SAVEE and custom datasets

Since the speakers in the recorded audio files in our custom dataset were not professional actors, sometimes an audio file can be interpreted as showing more than emotion. This is the case when the speaker is asked to express happiness, but astonishment and surprise are also expressed; Or in another case, this happens for sadness and neutrality. Regarding this fact, we anticipated a decrease in the accuracy of the classifiers, which is also shown in Table 4.6. Although after adding our custom dataset, a small decrease in accuracy is observed, the models are better trained for predicting the emotions of people who do not reflect their emotions in an exaggerated way.

Similar to the previous section, first we trained the classifiers with INTERSPEECH 2010 feature set. The tuned hyper parameters by GridSearchCV for each classifier are:

- Support Vector Machine: {'kernel': 'poly', 'gamma': 0.001, 'C': 10}
- Random Forest: {'n_estimators': 100, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_features': 0.5, 'max_depth': 7}

- Gradient Boosting: {'subsample': 1, 'n_estimators': 70, 'min_samples_split': 2, 'min_samples_leaf': 0.2, 'max_features': None, 'max_depth': 7, 'learning_rate': 0.3}
- Bagging: {'n_estimators': 50, 'max_samples': 1.0, 'max_features': 0.2}

Table 4.6 depicts the performance measures of different classifiers, trained and tested on INTERSPEECH 2010 feature set in the second approach. These results show that the bagging and gradient boosting have the best prediction results with average test accuracies of 84.48% and 83.50% respectively.

Table 4.6 Average performance measures of each classifier on the test set in the second approach.

Classifier	Feature Set	Accuracy	F1-score	Precision	Recall
SVM	INTERSPEECH 2010	0.7647	0.7306	0.7345	0.7310
Random Forest	INTERSPEECH 2010	0.7595	0.7612	0.7720	0.7595
Gradient Boosting	INTERSPEECH 2010	0.8350	0.8347	0.8367	0.8350
Bagging	INTERSPEECH 2010	0.8448	0.8445	0.8448	0.8460

Figure 4.5 shows another comparison between the accuracies of the classifiers on the test set after adding our custom dataset.

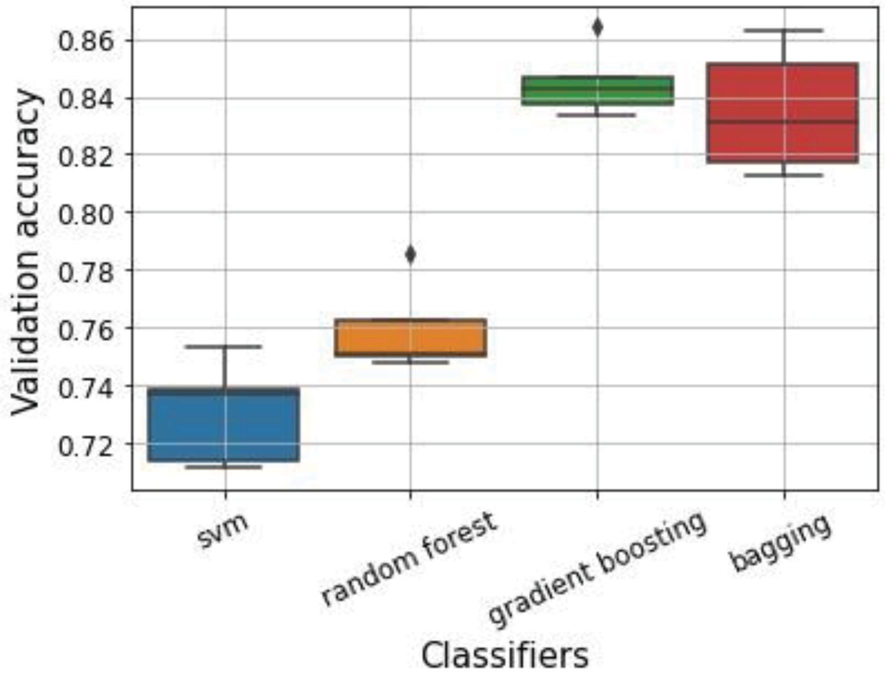


Figure 4.5 Validation accuracies of classifiers in the second approach.

Considering figure 4.6, which is the confusion matrix for gradient boosting as our best classifier in this scenario, it can be observed that predicting neutrality has the best accuracy of 91.0% among other emotions. On the other hand, predicting happiness has the lowest accuracy of 79.0%. It should also be noted that among the audio files that are predicted as happy ones, 7.0% of them are truly surprised audio files. This shows that surprise and happiness are the most confusing emotions in this case. The same scenario holds when the emotion is predicted to be neutrality, that is 7.0% of the files recognized to have neutrality as their most dominant emotion, have the correct label of sad. With the same explanation, the most confusing emotion while predicting neutrality is sadness.

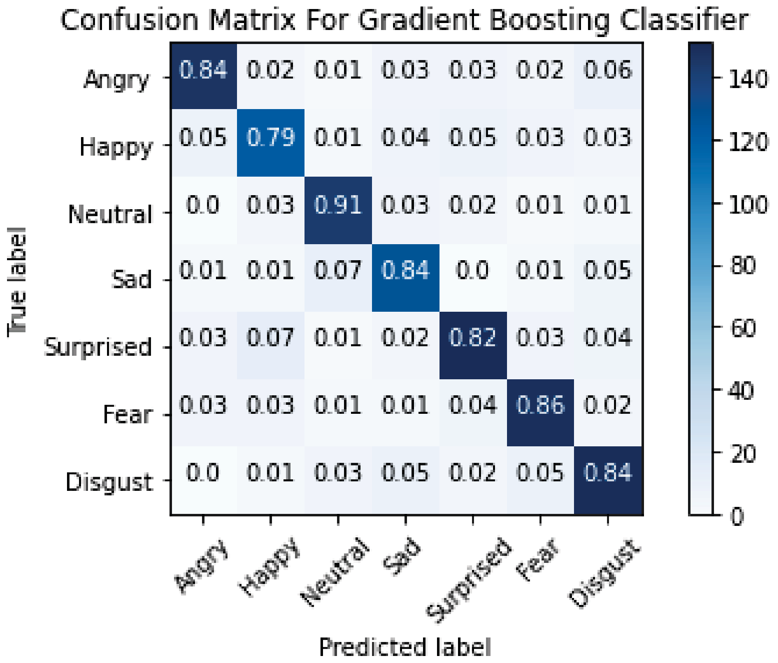


Figure 4.6 Confusion matrix of gradient boosting classifier with 5 emotions.

Table 4.7 depicts the performance measures of different classifiers, trained and tested on INTERSPEECH 2013 feature set in the second approach. The tuned hyper parameters by GridSearchCV for each classifier are:

- Support Vector Machine: {'kernel': 'poly', 'gamma': 0.001, 'C': 10}
- Random Forest: {'n_estimators': 100, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_features': 0.5, 'max_depth': 7}
- Gradient Boosting: {'subsample': 1, 'n_estimators': 70, 'min_samples_split': 2, 'min_samples_leaf': 0.2, 'max_features': None, 'max_depth': 7, 'learning_rate': 0.3}
- Bagging: {'n_estimators': 60, 'max_samples': 1.0, 'max_features': 0.2}

In this case SVM has the best accuracy of 84.41% which is a new behaviour considering the results from previous scenarios.

Table 4.7 Average performance measures of each classifier on the test set in the second approach.

Classifier	Feature Set	Accuracy	F1-score	Precision	Recall
SVM	INTERSPEECH 2013	0.8441	0.8446	0.8475	0.8442
Random Forest	INTERSPEECH 2013	0.7674	0.7675	0.7788	0.7674
Gradient Boosting	INTERSPEECH 2013	0.8286	0.8281	0.8296	0.8286
Bagging	INTERSPEECH 2013	0.8299	0.8295	0.8306	0.8299

Figure 4.7 shows another comparison between the accuracies of the classifiers on the test set after adding our custom dataset.

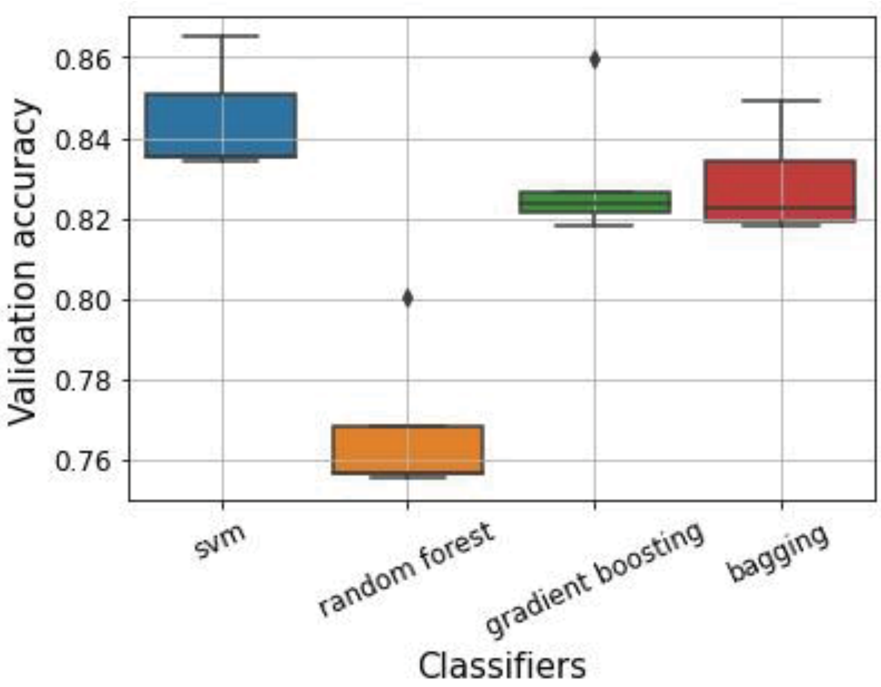


Figure 4.7 Validation accuracies of classifiers in the second approach.

Considering figure 4.8, which is the confusion matrix for SVM as our best classifier, it can be observed that again predicting neutrality has the best accuracy of 92.0% among other emotions. On the other hand, predicting anger has the lowest accuracy of 76.0%. It should also be noted that among the audio files that are predicted as angry ones, 3.0% of them are truly surprised audio files. Also 11.0% of the audio files that are predicted to be happy, are in fact angry audio files. This shows that in this case, anger and Happiness are the most confusing emotions. The same scenario holds when the emotion is predicted to be neutrality, that is 7.0% of the files recognized to have neutrality as their most dominant emotion, are in fact sad. With the same explanation, the most confusing emotion while predicting disgust is sadness.

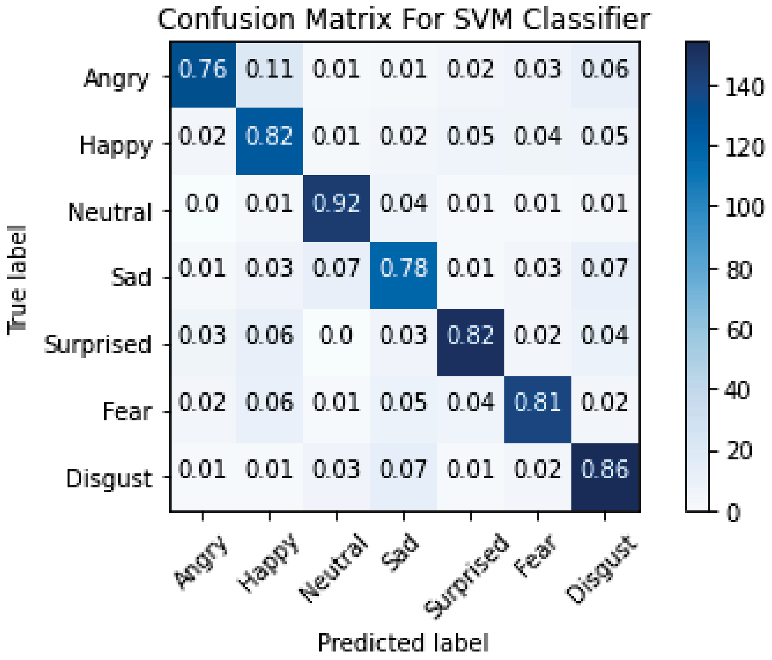


Figure 4.8 Confusion matrix of SVM classifier.

4.6 Conclusion

In this chapter we investigated recognizing emotions from speech audio files using machine learning classifiers such as SVM, Random Forest, Begging and a few others. While using the commonly perfect datasets of TESS, RAVEDESS and SAVEE, we tried two different feature sets of INTERSPEECH 2010 and INTERSPEECH 2013. In both these cases, Gradient Boosting and Bagging were more accurate in emotion recognition. This can be explained by the fact that these algorithms are more robust against overfitting the data. Although the accuracy of SVM was improved after increasing the number of features to 6373 from 1582. After adding our custom dataset to the previous described dataset, again Gradient Boosting and Bagging had better performance measures compared to SVM and Random Forest. Again, with increasing the number of features, the accuracy of SVM improved while it didn't have much effect on Random Forest. Finally, the best accuracy was for of 85.67% from gradient boosting while using INTERSPEECH IS10 as the feature set and a dataset composed of TESS, RAVDESS, SAVEE databases. In the next chapter we will explore how to leverage deep learning techniques and neural network classifiers to investigate the probability of reaching a higher accuracy in recognizing the emotions.

Chapter 5

Recognizing Emotions in Voice Using Deep Learning Classifiers

In the last chapter, we trained several machine learning classifiers such as SVM, random forest, gradient boosting and bagging. We evaluated the classifiers' accuracies in predicting the correct emotion in speech, both with and without considering our custom dataset. Now we are going to evaluate how the other powerful category of classifiers, which are deep learning classifiers, perform in detecting the emotions from speech audio files.

The outline of this chapter is like the last one, due to similar nature. In this chapter, first datasets and feature sets are described. For performing a fair evaluation, we used the same datasets and feature sets as in chapter 4. After pointing to this fact, leveraged deep learning techniques are explained briefly. Following sections explain the experiment plan and the obtained results from each scenario.

5.1 Datasets

Regarding the goal of comparing results, obtained from this approach and the previous approach mentioned in chapter 4, we used the same dataset as described in section 4.1.

5.2 Feature Sets

For the same goal, as mentioned in previous section, we used same feature sets as in the previous chapter, in section 4.2, which are the feature sets provided in INTERSPEECH 2010 paralinguistic challenge and in the INTERSPEECH 2013 computational paralinguistics challenge.

5.3 Machine Learning Models

Recurrent neural networks: Recurrent neural networks (RNNs) are powerful supervised learning algorithms that are perfect candidates for classification tasks where the input data is a sequence. Considering the sequential nature of audio files, RNNs seem to be the right classifiers;

although when audio files are long, we should face two main limitations of RNNs: gradients exploding and gradients vanishing [15]. To overcome these limitations, we get help of Long Short-Term Memory (LSTM). LSTM is a special type of RNNs that are getting popular over recent years. LSTM can mitigate the gradients exploding and vanishing problem by leveraging memory cells and gates in neurons. While training the model, gates learn how to manage the memory cells. Here we used LSTM in 3 hidden layers in addition to the input and output layers. The whole architecture of the final RNN is depicted in Figure 5.1.

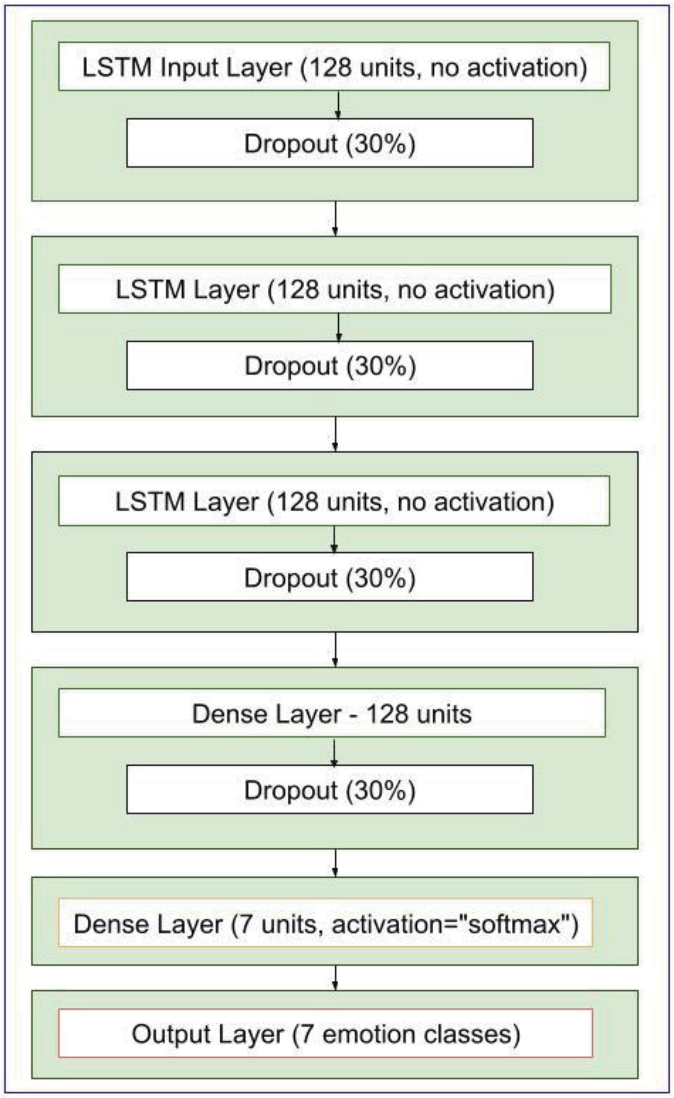

LSTM Input Layer (128 units, no activation)
Dropout (30%)
LSTM Layer (128 units, no activation)
Dropout (30%)
LSTM Layer (128 units, no activation)
Dropout (30%)
Dense Layer - 128 units
Dropout (30%)
Dense Layer (7 units, activation="softmax")
Output Layer (7 emotion classes)

Figure 5.1 The architecture of the evaluated Recurrent Neural Network.

5.4 Training and Validation

The process of training and validating deep learning classifiers is similar to the previous approach. We evaluated RNN and CNN in two different scenarios: first while considering only TESS, RAVDESS and SAVEE datasets and second while considering all four datasets of TESS, RAVDESS, SAVEE and our custom dataset. In each of these scenarios, we trained and tested classifiers using INTERSPEECH 2010 and INTERSPEECH 2013 feature sets, separately.

The predicted results of classifiers trained and tested in the described two scenarios are explained in the following parts.

5.5 Results And Discussion

In this section, several performance measure of different classifiers, trained and tested under two described approaches are reposted.

5.5.1 Considering TESS, RAVDESS and SAVEE datasets

First, we trained and tested deep learning classifiers using just the well-known datasets, without including our custom dataset.

Table 5.1 contains the average performance measures of each classifier on the test set. In this part, all the classifiers are trained using INSTERSPEECH 2010 as the feature set. The results for each performance measure are the averages over the results of that classifier over all the 5 repetitions of 5-old cross validations. As it can be seen, RNN has the noticeable highest average accuracy of 87.11%.

Table 5.1 Average performance measures of each classifier on the test set.

Classifier	Feature Set	Accuracy	F1-score	Precision	Recall
RNN	INTERSPEECH 2010	0.8711	0.8710	0.8723	0.8696
CNN	INTERSPEECH 2010	0.8394	0.8310	0.8322	0.8300

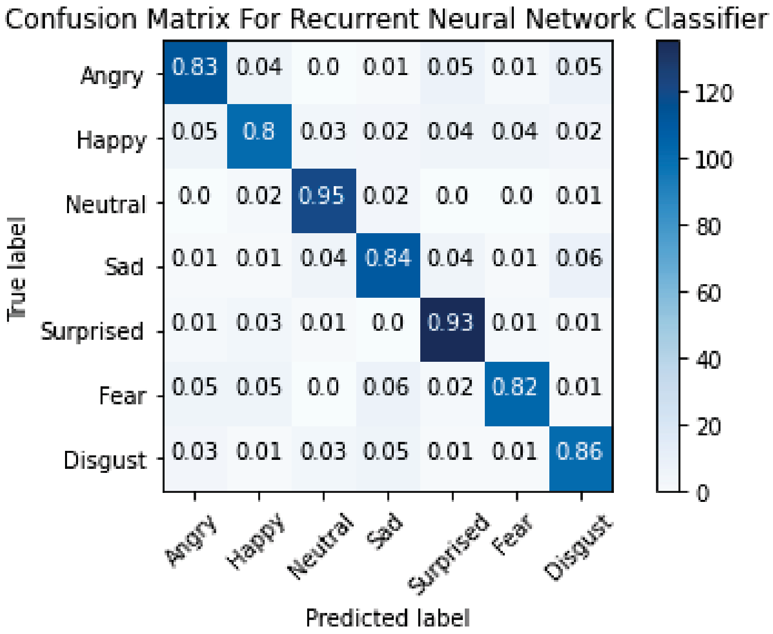


Figure 5.2 Confusion matrix of RNN.

After evaluating results of using INTERSPEECH 2010, we tried to assess our classifiers, which were trained with INTERSPEECH 2013 feature set. Table 5.2 contains the average performance measures of each classifier, trained with INTERSPEECH 2013, on the test set. The results for each performance measure are the averages over the results of that classifier over all the 5 repetitions of 5-old cross validations. As it can be seen again in this case, similar to using INTERSPEECH 2010, RNN has the best average accuracy of 89.88%. It should also be noted that this accuracy is the highest accuracy among all different approaches in this study.

Table 5.2 Average performance measures of each classifier on the test set.

Classifier	Feature Set	Accuracy	F1-score	Precision	Recall
RNN	INTERSPEECH 2013	0.8988	0.8981	0.8990	0.8973
CNN	INTERSPEECH 2013	0.8085	0.7986	0.7999	0.7974

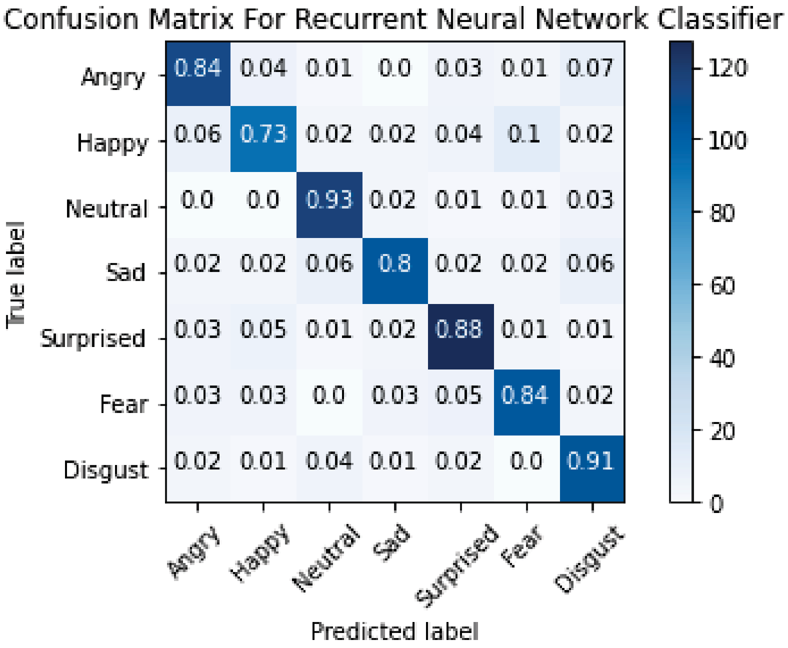


Figure 5.3 Confusion matrix of RNN.

5.5.2 Considering TESS, RAVDESS, SAVEE and custom datasets

Following a similar approach with evaluating classifiers in chapter 4, here we will report the effect of adding our custom dataset to the previous set. Again, classifiers are trained and tested using both mentioned feature sets.

Table 5.3 contains the results of evaluating RNN and CCN on a final dataset, composed of all four datasets of TESS, RAVDESS, SAVEE and custom dataset and trained with INTERSPEECH 2010 feature set. In this case, RNN is still the dominant classifier considering the accuracy of 86.71%.

Table 5.3 Average performance measures of each classifier on the test set in the second approach.

Classifier	Feature Set	Accuracy	F1-score	Precision	Recall
RNN	INTERSPEECH 2010	0.8671	0.8665	0.8672	0.8660
CNN	INTERSPEECH 2010	0.8288	0.8246	0.8260	0.8234

Figure 5.3 shows another comparison between the accuracies of the classifiers on the test set after adding our custom dataset.

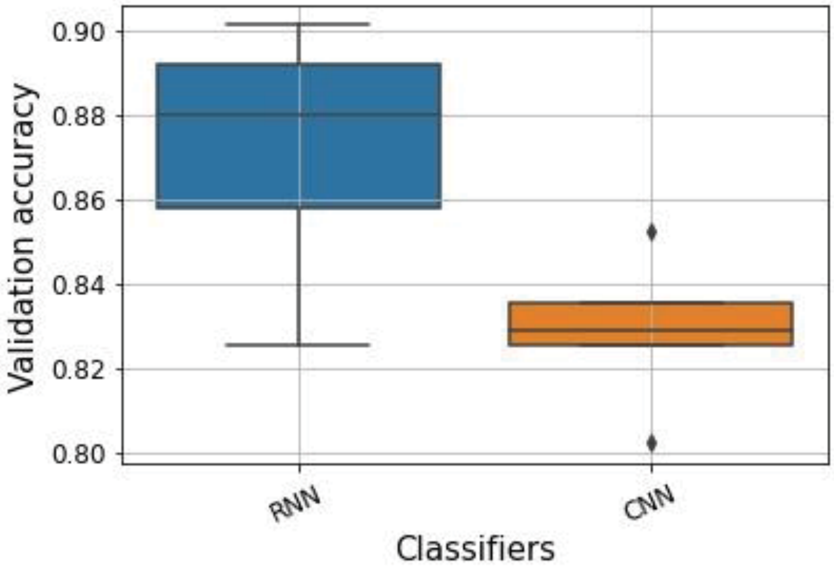


Figure 5.4 Validation accuracies of classifiers in the second approach.

Regarding figure 5.4, which is the confusion matrix for RNN as our best classifier, it can be observed that predicting anger has the best accuracy of 88.0% among other emotions. On the other hand, predicting happiness has the lowest accuracy of 78.0%.

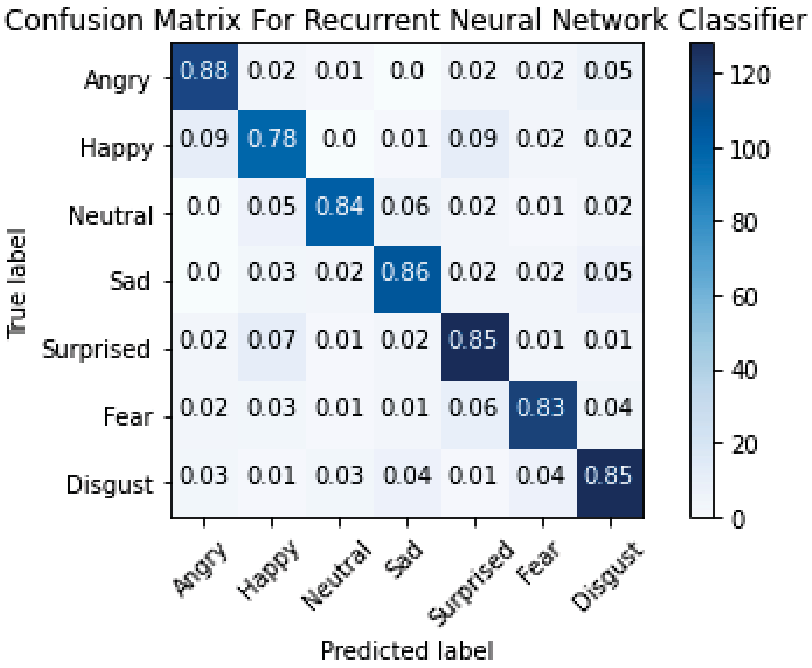


Figure 5.5 Confusion matrix RNN.

We repeated the experimenting while considering all four datasets, this time using INTERSPEECH 2013 as our feature set. Table 5.4 shows how RNN is till the dominant classifier regarding the high accuracy of 90.41%, while being trained by INTERSPEECH 2010 feature set.

Table 5.4 Average performance measures of each classifier on the test set in the second approach.

Classifier	Feature Set	Accuracy	F1-score	Precision	Recall
RNN	INTERSPEECH 2013	0.9041	0.9034	0.9041	0.9026
CNN	INTERSPEECH 2013	0.8071	0.8033	0.8040	0.8027

Figure 5.5 shows another comparison between the accuracies of the classifiers on the test set.

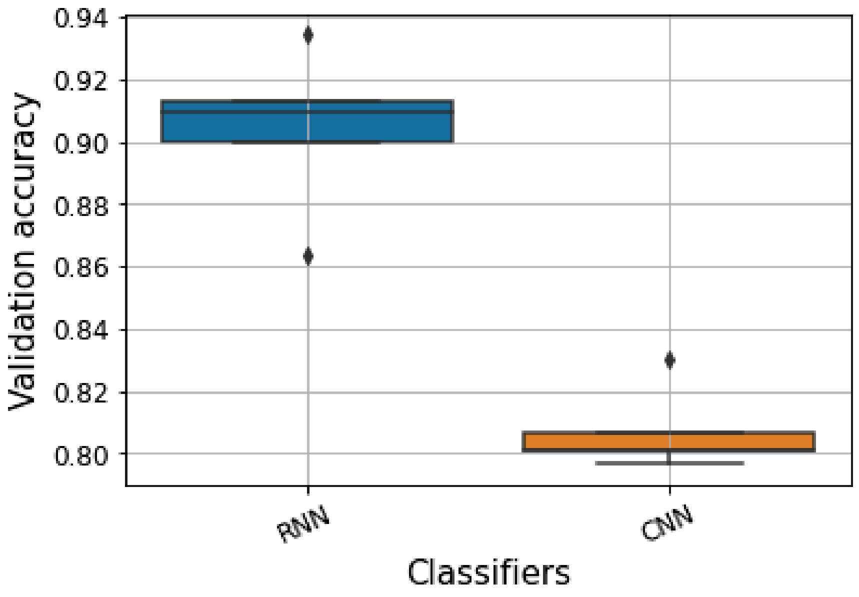


Figure 5.6 Validation accuracies of classifiers in the second approach.

Considering the confusion matrix for RNN, depicted in figure 5.6, it can be observed that this time predicting neutrality has the best accuracy of 92.0% among other emotions and fear and sadness have the lowest accuracy of 80.0%.

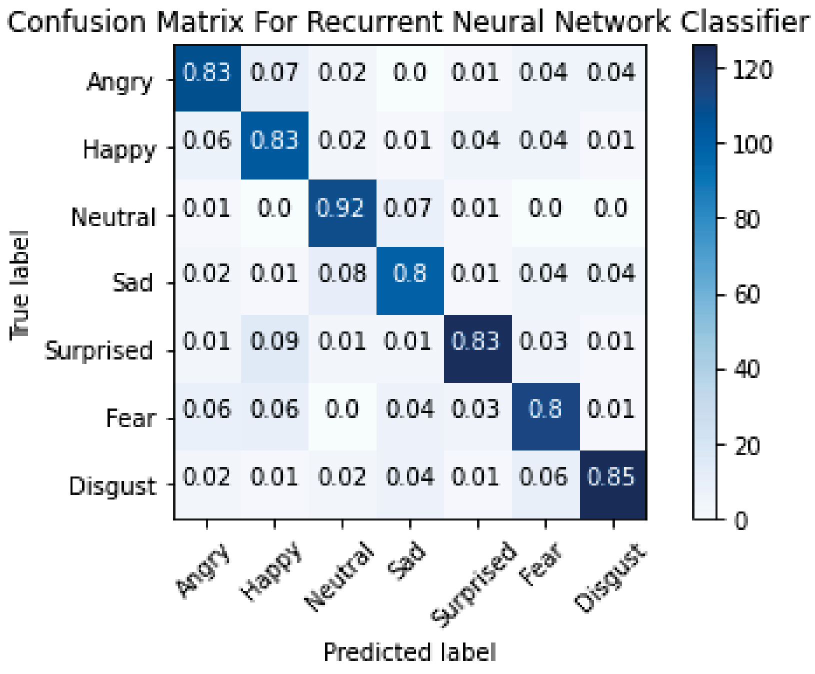


Figure 5.7 Confusion matrix of RNN.

5.6 Conclusion

In this chapter we investigated recognizing emotions from speech audio files using deep learning classifiers such as recurrent neural networks and convolutional neural networks. We trained these classifiers and got the best accuracies of 90.41% from RNN, which is a noticeable accuracy among the related studies in this field. Now considering the trained classifiers in this chapter and the best ones from the previous chapter, we have several classifiers with high accuracies in recognizing emotions from audio files. In this step one may wonder the probability of combining these classifiers to build a more accurate final classifier. This is the core idea of using ensemble classifiers. In the next chapter we will explain how we implemented this idea and what are the results of such an approach.

Chapter 6

Recognizing Emotions in Voice Using Ensemble Machine Learning Classifiers

In the last two chapters we investigated our two first hypothesis about using machine learning and deep learning classifiers for detecting the emotions from speech audio files. Result of this investigation was obtaining high accuracies of 84.48% and 90.41 from machine learning and deep learning classifiers respectively. After training and validating these classifiers, we explained the confusion matrix for each one of them. Recognizing that which classifier is more capable in accurate in detecting which emotion, was the reason behind that explanation. At this step, being inspired by several studies(Ira & Rahman, 2020; Zehra et al., 2021), which used ensemble classifiers instead of traditional classifiers for speech emotion detection tasks, we tried to evaluate this possibility using our collected dataset. The idea behind using ensemble classifiers is to combine several well-trained classifiers and form one that is more accurate than each of the forming classifiers separately.

In the following sections of this chapter, we will explain our approach in building an ensemble classifier for detecting the emotions in speech audio files. In regards of evaluating all our classifiers fairly, the same rich dataset and feature set, which were explained in previous chapters are also used in this approach. In the next sections, first we will explain the selected classifiers as building blocks of the ensemble classifier. Next section describes the process of training and testing this ensemble classifier. After that, we will report the results of validating the detected emotion classes by this approach against our validation data.

6.1 Machine Learning Models

For building an ensemble classifier, first we should decide about the inner classifiers. We have chosen several classifiers with the highest accuracies, from our previous approaches, which are explained in chapters 4 and 5. Considering results of evaluating the classifiers, trained with all mentioned datasets together with our custom dataset, the SVM classifier had the best accuracy of 85.18% among all the classifiers, while using IS2013 feature set. The most accurate classifier, while using IS10 feature set, was gradient boosting with the accuracy of 84.48%. These results were all about classifiers trained and evaluated in chapter 4. In chapter 5, while using deep learning techniques for training the classifiers, it was noticed that recurrent neural networks had the highest accuracy of 90.41% and 86.71% while being trained on IS13 and IS10 feature sets, respectively.

We are going to evaluate the performance measures of an ensemble classifier composed of these four highest accurate classifiers. In this method, after these four classifiers are trained, for detecting the emotion in each audio file in the test data, a majority voting is performed between the detected emotion by each of the four classifiers. There are two common voting approaches:

Soft Voting: For following this approach, the output of each classifier should be a set of values, each of them representing the probability that the input belongs to a specific output class. Then all the predictions from different classifiers are weighted based on the importance of the classifier. After that a sum is calculated over the results of classifiers. The final output is the greatest value.

Hard Voting: In this approach, each classifier votes for one single output class. The final output is the most voted class among all classifiers.

We have followed both approaches. It should be explained that when following the hard voting scheme, in case of a tie situation, the final vote is the vote of the recurrent neural network, trained on IS13 feature sets. The reason behind this privilege is that this RNN has the highest accuracy among all four classifiers. For doing a fair comparison, we are also going to evaluate the performance measures of ensemble classifiers which are composed of several instances of the same classifier. For example, we will see what are the performance measures of an ensemble model composed of four instances of the RNN with the same architecture as previous chapter. We will follow this procedure for RNN trained with IS10, RNN trained with IS13, SVM and gradient boosting.

6.2 Training and Validation

For training and validating the ensemble classifier, first we trained our best classifiers using all four datasets and INTERSPEECH 2013. The reason for choosing this combination is the better performance of classifiers under these conditions, as explained in the previous chapters. After training theses classifiers, we embedded them in our ensemble classifier. For validating the final classifier, again we followed the 5-fold cross validation approach.

The evaluation results of embedded classifiers and the final ensemble classifier are described in the following parts.

6.3 Results And Discussion

Table 6.1 contains the average performance measures of each classifier on the test set. The results for each performance measure are the averages over the results of that classifier over all the 5 repetitions of 5-old cross validations. As it can be seen, soft voting ensemble classifier has the noticeable highest average accuracy of 91.96%. This accuracy is higher than all the inner embedded classifiers, which their results are reported in the first 4 rows of this table. Also, we build several ensemble models which are composed of several classifiers of the same type. For example, row 5 shows the performance measures of an ensemble model which is built by doing a soft majority voting over 4 separately trained RNNs using INTERSPEECH IS10. Rows 5 and 6 show that these ensemble models are performing better than one single RNN alone. This can be explained by considering that each of the inner four separately trained RNNs had some random initial weights in the beginning. By considering four instances we have lowered the effect of initial state in recognizing the final output class. This is true for rows 7 and 8 which are ensembles of soft and hard voting over four separately trained RNNs using INTERSPEECH IS13. These ensemble classifiers have 1.4% and 0.89% improved accuracies compared to row 2 which is a single RNN using INTERSPEECH IS13. It should also be noted that rows 9 to 12 which are ensembles of four separate SVMs and Gradient Boostings, do not have much improved accuracy compared to single SVM and single Gradient Boosting, respectively.

Table 6.1 Average performance measures of each classifier on the test set in the ensemble approach.

ID	Classifier	Feature set	Accuracy	F1-score	Precision	Recall
1	RNN	INTERSPEECH IS10	0.8671	0.8665	0.8672	0.8660
2	RNN	INTERSPEECH IS13	0.9041	0.9033	0.9041	0.9026
3	SVM	INTERSPEECH IS13	0.8518	0.8530	0.8543	0.8518
4	Gradient Boosting	INTERSPEECH IS10	0.8428	0.8433	0.8437	0.8429
5	Soft Voting Ensemble of four RNNs	INTERSPEECH IS10	0.8989	0.9038	0.9087	0.8989
6	Hard Voting Ensemble of four RNNs	INTERSPEECH IS10	0.8929	0.8988	0.9047	0.8929
7	Soft Voting Ensemble of four RNNs	INTERSPEECH IS13	0.9181	0.9210	0.924	0.9181
8	Hard Voting Ensemble of four RNNs	INTERSPEECH IS13	0.9130	0.9352	0.9384	0.9130
9	Soft Voting Ensemble of four SVMs	INTERSPEECH IS13	0.8518	0.8530	0.8543	0.8518
10	Hard Voting Ensemble of four SVMs	INTERSPEECH IS13	0.8518	0.8530	0.8543	0.8518
11	Soft Voting Ensemble of four Gradient Boosting Classifiers	INTERSPEECH IS10	0.8457	0.8462	0.8467	0.8457
12	Hard Voting Ensemble of four Gradient Boosting Classifiers	INTERSPEECH IS10	0.8457	0.8462	0.8467	0.8457
13	Soft Voting Ensemble of classifiers 1, 2, 3 and 4	NA	0.9196	0.9197	0.9231	0.9196
14	Hard Voting Ensemble of classifiers 1, 2, 3 and 4	NA	0.8960	0.8958	0.9031	0.8960

Figure 6.1 shows the results obtained from RNN, SVM and gradient boosting standalone. As it was explained earlier, the results reported for soft voting ensemble and hard voting

ensemble is the results of performing soft and hard majority voting among the predicted class by RNN, SVM and gradient boosting for each test case. Table 6.1 depicts that for all the accuracies from all 5 repetitions of 5-fold cross validation, soft voting ensemble had a higher accuracy in predicting the emotion from voice.

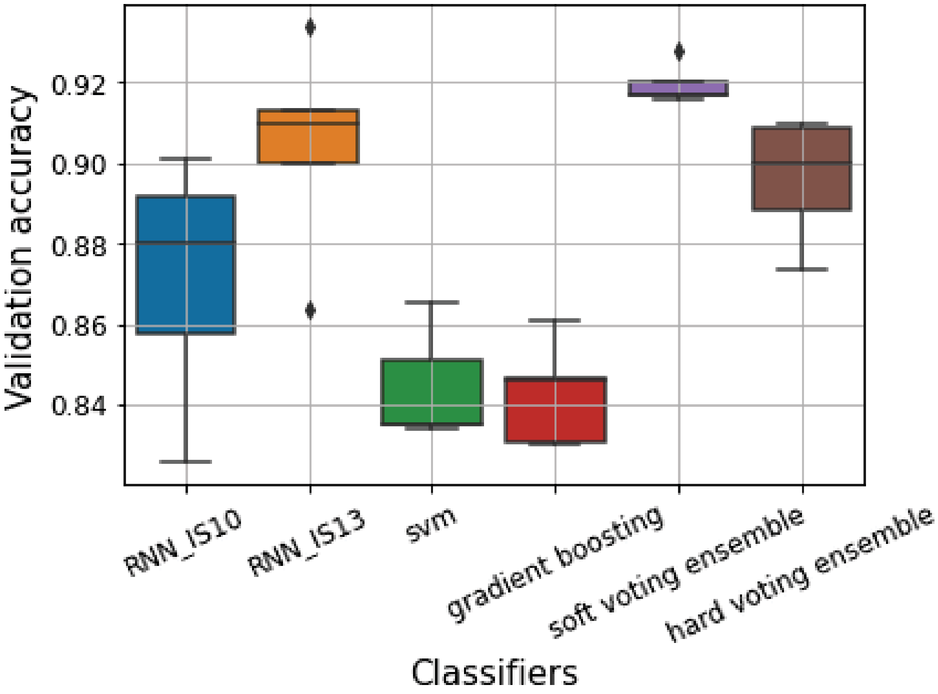


Figure 6.1 Validation accuracies of classifiers in the ensemble approach.

Considering figure 6.2 which is the confusion matrix for soft voting ensemble as our best classifier, it can be observed that like most of previous cases, predicting neutrality has the best accuracy of 91.0% among other emotions. On the other hand, predicting anger has the lowest accuracy of 8.0%. It should also be noted that among the audio files that are predicted as angry ones, 3.0% of them are truly happy audio files. This shows that in this case anger and happiness are the most confusing emotions when predicting anger. The same scenario holds when the emotion is predicted to be neutrality, that is 4.0% of the files recognized to have neutrality as their most dominant emotion, are in fact sad. With the same explanation, the most confusing emotion while predicting disgust is anger.

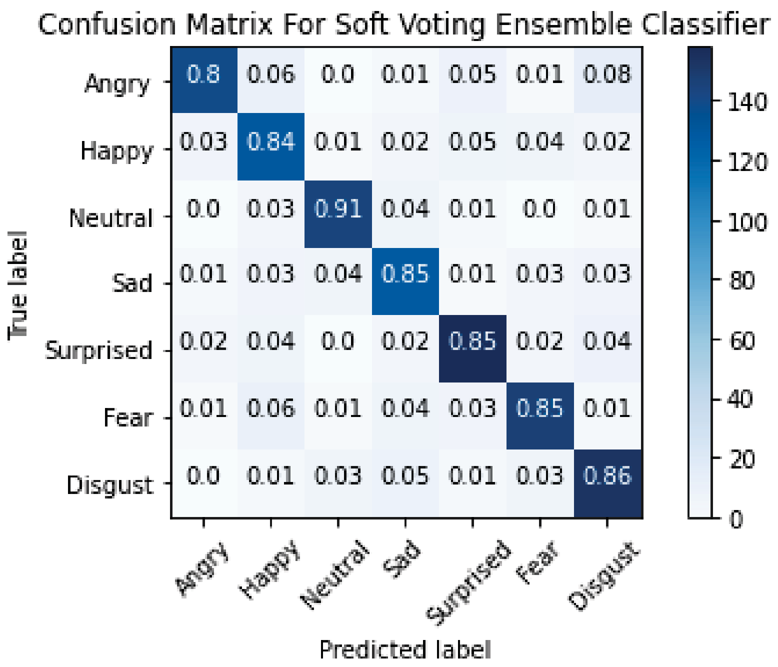


Figure 6.2 Confusion matrix for soft voting ensemble classifier.

6.4 Conclusion

In this chapter we explored how to recognize emotions from speech audio files using soft and hard majority voting ensemble classifiers. The motivation behind this approach was to combine the strengths from all our previous classifiers with high accuracies. We obtained the noticeable accuracy of 91.96% from soft voting ensemble classifier. By considering the confusion matrix of this classifier, it can be understood that most confused emotions are disgust and anger. Knowing this we are motivated to use another aspect of the speech, which is the text of speech. In the future works of this study, we are going to evaluate how leveraging text of speech and using a text emotion classifier may lead to better accuracies while recognizing the emotions from speech.

Chapter 7

Conclusion

In this study, we investigated the effectiveness of using machine learning, deep learning and ensemble learning techniques in recognizing the emotions from speech or voice. We evaluated several different approaches while leveraging different datasets, feature sets and classifiers.

One of our main goals in this study was to target detecting the emotions from people's voices who may not be as perfect and capable as actors in reflecting their emotions in their voices. To address this requirement, we gathered a custom emotion audio data set, recorded from non-actor and non-experts. We could gather 125 audio files from 4 people, 2 males and 2 females, with an average age of 30 years old. We used International Affective Picture System or IAPS for triggering the emotions in the participants. During this report, we referred to this dataset as the custom dataset.

After dealing with the challenge of addressing non-actors' emotions, we followed several approaches for obtaining high accuracies in detecting the emotions from voice. First, we reported how several machine learning classifiers such as SVM, random forest, gradient boosting and bagging classifiers perform. We trained and tested these models using different set of emotion audio datasets and different features sets, regarding the nature of voice signals. We repeated training and testing the classifiers under four scenarios:

- Considering TESS, RAVDESS and SAVEE datasets, using INTERSPEECH 2010
- Considering TESS, RAVDESS, SAVEE and custom datasets, using INTERSPEECH 2010
- Considering TESS, RAVDESS and SAVEE datasets, using INTERSPEECH 2013
- Considering TESS, RAVDESS, SAVEE and custom datasets, using INTERSPEECH 2013

After evaluating all the trained classifiers, using 5-fold cross validation, the highest accuracy of 85.67% was obtained by gradient booting, being trained with INTERSPEECH 2010 and a dataset composed of TESS, RAVDESS, SAVEE databases.

In our second approach, we evaluated the performance of deep learning models, another powerful and massively used set of classifiers in the field of voice emotion recognition. We repeated previously mentioned four scenarios, while training and testing recurrent neural networks and convolutional neural networks. The motivation behind this choice was the continuous nature of voice audio signals. Since recurrent and convolutional neural networks were excessively successful in classifying inputs with a continuous nature, such as images, they seemed as perfect candidates for detecting the emotions from another continuous input such as voice signals. This assumption led us to obtain an accuracy of 90.41%, which is noticeably higher than our first approach, while training a recurrent neural network, using INTERSPEECH 2013 feature set.

In the third approach, we tried to leverage the power of ensemble classifiers, to get a higher accuracy in detecting the emotions from voice, using our previously trained classifiers. We used several of our most accurate classifiers from the previous approaches, such as gradient booting and recurrent neural network as inner building blocks of the ensemble classifier. We evaluated the accuracies obtained from both soft and hard voting ensemble classifiers. At the end we could reach the considerably accuracy of 91.96% from our soft voting ensemble classifier. This proves the capability of ensemble learning techniques in voice emotion detection tasks.

7.1 Future Works

Since in real case scenarios, emotions may not always be reflected in the speaker's voice noticeably, getting help from spoken words may help in predicting the true emotion. Regarding this, we are going to use speech to text conversion techniques. After generating the text of speech, we will evaluate how using a text emotion classifier may help us in obtaining higher accuracies in detecting the emotions. This approach may decrease the current confusion rate between anger and happiness, since although the audio signals have a lot of common features, but the spoken words are normally very different.

We are also going to increase the size of our custom dataset which may lead to higher prediction accuracy. Currently we have 125 audio files with definitely should be increased compared to the other used dataset which is composed of 4578 audio files recorded from actors and experts.

With all of this, we are going to extend our testing procedures with using EEG experiments. We will use the results of EEG experiment to validate our classifiers accuracies in detecting the emotions from non-actor people's voices while showing them the IAPS images.

Bibliography

Aijun, L., Fang, Z., Byrne, W., Fung, P., Kamm, T., Yi, L., Zhanjiang, S., Ruhi, U., Venkataramani, V., & XiaoXia, C. (n.d.). *CASS: A PHONETICALLY TRANSCRIBED CORPUS OF MANDARIN SPONTANEOUS SPEECH*. 4.

Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, *116*, 56–76. https://doi.org/10.1016/j.specom.2019.12.001

Albornoz, E. M., Milone, D. H., & Rufiner, H. L. (2011). Spoken emotion recognition using hierarchical classifiers. *Computer Speech & Language*, *25*(3), 556–570. https://doi.org/10.1016/j.csl.2010.10.001

Batliner, A., Steidl, S., & Noeth, E. (2008). *Releasing a thoroughly annotated and processed spontaneous emotional database: The FAU Aibo Emotion Corpus*.

Bitouk, D., Verma, R., & Nenkova, A. (2010). Class-level spectral features for emotion recognition. *Speech Communication*, *52*(7), 613–625. https://doi.org/10.1016/j.specom.2010.02.010

Borchert, M., & Dusterhoft, A. (2005). Emotions in speech—Experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments. *2005 International Conference on Natural Language Processing and Knowledge Engineering*, 147–151. https://doi.org/10.1109/NLPKE.2005.1598724

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. *Interspeech 2005*, 1517–1520. https://doi.org/10.21437/Interspeech.2005-446

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

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Lee, S., Neumann, U., & Narayanan, S. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. *Proceedings of the 6th International Conference on Multimodal Interfaces*, 205–211.

Busso, C., Lee, S., & Narayanan, S. (2009). Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(4), 582–596. https://doi.org/10.1109/TASL.2008.2009578

Chen, H., Liu, Z., Kang, X., Nishide, S., & Ren, F. (2019). Investigating voice features for Speech emotion recognition based on four kinds of machine learning methods. *2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems (CCIS)*, 195–199.

Costantini, G., Iaderola, I., Paoloni, A., & Todisco, M. (2014). EMOVO Corpus: An Italian Emotional Speech Database. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 3501–3504. http://www.lrec-conf.org/proceedings/lrec2014/pdf/591_Paper.pdf

Davenport, M., & Hannahs, S. J. (2020). *Introducing Phonetics and Phonology* (4th ed.). Routledge. https://doi.org/10.4324/9781351042789

Deng, J., Xu, X., Zhang, Z., Frühholz, S., Grandjean, D., & Schuller, B. (2017). Fisher kernels on phase-based features for speech emotion recognition. In *Dialogues with social robots* (pp. 195–203). Springer.

Deng, J., Zhang, Z., Eyben, F., & Schuller, B. (2014). Autoencoder-based Unsupervised Domain Adaptation for Speech Emotion Recognition. *IEEE Signal Processing Letters*, *21*(9), 1068–1072. https://doi.org/10.1109/LSP.2014.2324759

Deng, J., Zhang, Z., Marchi, E., & Schuller, B. (2013). Sparse Autoencoder-Based Feature Transfer Learning for Speech Emotion Recognition. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 511–516. https://doi.org/10.1109/ACII.2013.90

Ekman, P., Friesen, W. V., & Ellsworth, P. (2013). *Emotion in the Human Face: Guidelines for Research and an Integration of Findings*. Elsevier.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*(3), 572–587.

Engberg, I. S., Hansen, A. V., Andersen, O., & Dalsgaard, P. (n.d.). *DESIGN, RECORDING AND VERIFICATION OF A DANISH EMOTIONAL SPEECH DATABASE*. 4.

Eyben, F., Wöllmer, M., Graves, A., Schuller, B., Douglas-Cowie, E., & Cowie, R. (2010). On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. *Journal on Multimodal User Interfaces*, *3*(1), 7–19. https://doi.org/10.1007/s12193-009-0032-6

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM International Conference on Multimedia*, 1459–1462.

Gannouni, S., Aledaily, A., Belwafi, K., & Aboalsamh, H. (2021). Emotion detection using electroencephalography signals and a zero-time windowing-based epoch estimation and relevant electrode identification. *Scientific Reports*, *11*(1), 7071. https://doi.org/10.1038/s41598-021-86345-5

Grimm, M., Kroschel, K., Mower, E., & Narayanan, S. (2007). Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, *49*(10), 787–800. https://doi.org/10.1016/j.specom.2007.01.010

Grimm, M., Kroschel, K., & Narayanan, S. (2008). The Vera am Mittag German audio-visual emotional speech database. *2008 IEEE International Conference on Multimedia and Expo*, 865–868. https://doi.org/10.1109/ICME.2008.4607572

Han, K., Yu, D., & Tashev, I. (2014, September 1). *Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine*. Interspeech 2014. https://www.microsoft.com/en-us/research/publication/speech-emotion-recognition-using-deep-neural-network-and-extreme-learning-machine/

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

Haq, S., & Jackson, P. J. B. (2010). *Machine Audition: Principles, Algorithms and Systems* (W. Wang, Ed.; pp. 398–423). IGI Global.

Hu, H., Xu, M.-X., & Wu, W. (2007). GMM Supervector Based SVM with Spectral Features for Speech Emotion Recognition. *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, *4*, IV-413-IV–416. https://doi.org/10.1109/ICASSP.2007.366937

Ira, N. T., & Rahman, M. O. (2020). An Efficient Speech Emotion Recognition Using Ensemble Method of Supervised Classifiers. *2020 Emerging Technology in Computing, Communication and Electronics (ETCCE)*, 1–5. https://doi.org/10.1109/ETCCE51779.2020.9350913

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Kossaifi, J., Tzimiropoulos, G., Todorovic, S., & Pantic, M. (2017). AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, *65*, 23–36. https://doi.org/10.1016/j.imavis.2017.02.001

Kwon, O.-W., Chan, K., Hao, J., & Lee, T.-W. (2003). *Emotion Recognition by Speech Signals*. 4.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1999). International affective picture system (IAPS): Instruction manual and affective ratings. *The Center for Research in Psychophysiology, University of Florida*.

Lee, C. M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., & Narayanan, S. (2004). Emotion recognition based on phoneme classes. *Eighth International Conference on Spoken Language Processing*.

Lee, C.-C., Mower, E., Busso, C., Lee, S., & Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, *53*(9), 1162–1171. https://doi.org/10.1016/j.specom.2011.06.004

Lee, S., Yildirim, S., Kazemzadeh, A., & Narayanan, S. (2005). *An Articulatory Study of Emotional Speech Production*. 4.

Li, Y., Tao, J., Chao, L., Bao, W., & Liu, Y. (2017). CHEAVD: A Chinese natural emotional audio–visual database. *Journal of Ambient Intelligence and Humanized Computing*, *8*(6), 913–924. https://doi.org/10.1007/s12652-016-0406-z

Liberman, Mark, Davis, Kelly, Grossman, Murray, Martey, Nii, & Bell, John. (2002). *Emotional Prosody Speech and Transcripts* (p. 1124092 KB) [Data set]. Linguistic Data Consortium. https://doi.org/10.35111/37FF-A902

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS One*, *13*(5), e0196391.

Luengo, I., Navas, E., Hernáez, I., & Sánchez, J. (2005). Automatic Emotion Recognition using Prosodic Parameters. *In Proc. of INTERSPEECH*, 493–496.

Mao, Q., Dong, M., Huang, Z., & Zhan, Y. (2014). Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks. *IEEE Transactions on Multimedia*, *16*(8), 2203–2213. https://doi.org/10.1109/TMM.2014.2360798

Mao, X., Chen, L., & Fu, L. (2009). Multi-level Speech Emotion Recognition Based on HMM and ANN. *2009 WRI World Congress on Computer Science and Information Engineering*, *7*, 225–229. https://doi.org/10.1109/CSIE.2009.113

Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006). The eNTERFACE' 05 Audio-Visual Emotion Database. *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, 8–8. https://doi.org/10.1109/ICDEW.2006.145

McFee, B., Raffel, C., Liang, D., Ellis, D., McVicar, M., Battenberg, E., & Nieto, O. (2015). *librosa: Audio and Music Signal Analysis in Python*. 18–24. https://doi.org/10.25080/Majora-7b98e3ed-003

McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schroder, M. (2012). The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. *IEEE Transactions on Affective Computing*, *3*(1), 5–17. https://doi.org/10.1109/T-AFFC.2011.20

Mirsamadi, S., Barsoum, E., & Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2227–2231.

Mori, S., Moriyama, T., & Ozawa, S. (2006). Emotional Speech Synthesis using Subspace Constraints in Prosody. *2006 IEEE International Conference on Multimedia and Expo*, 1093–1096. https://doi.org/10.1109/ICME.2006.262725

Nakatsu, R., Solomides, A., & Tosa, N. (1999). Emotion recognition and its application to computer agents with spontaneous interactive capabilities. *Proceedings IEEE International Conference on*

Multimedia Computing and Systems, *2*, 804–808 vol.2. https://doi.org/10.1109/MMCS.1999.778589

Nogueiras, A., Moreno, A., Bonafonte, A., & Mariño, J. B. (2001). *Speech Emotion Recognition Using Hidden Markov Models*. 4.

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

Oflazoglu, C., & Yildirim, S. (2013). Recognizing emotion from Turkish speech using acoustic features. *EURASIP Journal on Audio, Speech, and Music Processing*, *2013*(1), 26. https://doi.org/10.1186/1687-4722-2013-26

Petrushin, V. (1999). Emotion in speech: Recognition and application to call centers. *Proceedings of Artificial Neural Networks in Engineering*, *710*, 22.

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

Pichora-Fuller, M. K., & Dupuis, K. (2020). *Toronto emotional speech set (TESS)* [Data set]. Scholars Portal Dataverse. https://doi.org/10.5683/SP2/E8H2MF

Plutchik, R. (2001). The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, *89*(4), 344–350.

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

Ringeval, F., Sonderegger, A., Sauer, J., & Lalanne, D. (2013). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. *2013 10th IEEE International Conference and*

Workshops on Automatic Face and Gesture Recognition (FG), 1–8. https://doi.org/10.1109/FG.2013.6553805

Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, *11*(3), 273–294. https://doi.org/10.1016/0092-6566(77)90037-X

Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Höthker, A., & Konosu, H. (2009). Being bored? Recognising natural interest by extensive audiovisual integration for real-life application. *Image and Vision Computing*, *27*(12), 1760–1774. https://doi.org/10.1016/j.imavis.2009.02.013

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. S. (2010). The INTERSPEECH 2010 paralinguistic challenge. *Eleventh Annual Conference of the International Speech Communication Association*.

Shaqra, F. A., Duwairi, R., & Al-Ayyoub, M. (2019). Recognizing emotion from speech based on age and gender using hierarchical models. *Procedia Computer Science*, *151*, 37–44.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Surrey Audio-Visual Expressed Emotion (SAVEE) Database. (n.d.). Retrieved January 5, 2021, from http://kahlan.eps.surrey.ac.uk/savee/

Tao, F., Liu, G., & Zhao, Q. (2018). An ensemble framework of voice-based emotion recognition system for films and TV programs. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6209–6213.

Tao, J., Liu, F., Zhang, M., & Jia, H. (2008). *Design of Speech Corpus for Mandarin Text to Speech*. https://www.semanticscholar.org/paper/Design-of-Speech-Corpus-for-Mandarin-Text-to-Speech-Tao-Liu/32a8b963725fe935f3bc3fc0bfeb74b4333cf137

Tao, J., & Tan, T. (2005). *Affective Computing: A Review* (p. 995). https://doi.org/10.1007/11573548_125

Tian, L., Moore, J., & Lai, C. (2016). Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features. *2016 IEEE Spoken Language Technology Workshop (SLT)*, 565–572. https://doi.org/10.1109/SLT.2016.7846319

Toledano, D. T., Ramos, D., Gonzalez-Dominguez, J., & González-Rodríguez, J. (2009). Speech Analysis. In S. Z. Li & A. Jain (Eds.), *Encyclopedia of Biometrics* (pp. 1284–1289). Springer US. https://doi.org/10.1007/978-0-387-73003-5_200

Vapnik, V. N., VAPNIK, V. A., & Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley.

Wang, K. (2018). *A Database of elderly emotional speech*.

Yacoub, S., Simske, S., Lin, X., & Burns, J. (2003). Recognition of emotions in interactive voice response systems. *Eighth European Conference on Speech Communication and Technology*.

Zehra, W., Javed, A. R., Jalil, Z., Khan, H. U., & Gadekallu, T. R. (2021). Cross corpus multi-lingual speech emotion recognition using ensemble learning. *Complex & Intelligent Systems*, *7*(4), 1845–1854. https://doi.org/10.1007/s40747-020-00250-4

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(1), 39–58. https://doi.org/10.1109/TPAMI.2008.52

Zhalehpour, S., Onder, O., Akhtar, Z., & Erdem, C. E. (2017). BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States. *IEEE Transactions on Affective Computing*, *8*(3), 300–313. https://doi.org/10.1109/TAFFC.2016.2553038

Annexes

9.1 Voice Emotion Recognition in Real Time Applications.

During this work, we could publish one paper about recognizing the emotions from voice. In this publication we explained the advantages of using emotion by computers for serving humans' needs better. The paper was published in ITS 2021, The 17th Intelligent Tutoring Systems International Conference, Athens, Greece, June 7-11, 2021, Springer Verlag Lecture Notes in Computer Science.

The online version can be found in: https://link.springer.com/chapter/10.1007/978-3-030-80421-3_53

The contribution of the writers is as follows:

- Mahsa Aghajani: Gathering custom dataset, doing the experiments and writing the paper
- Hamdi Ben Abdessalem: contributed with his advice in this project and revised the paper.
- Claude Frasson: revised the paper and financed the project.

This complete version of the publication is as follows:

Voice Emotion Recognition in Real Time Applications

Mahsa Aghajani, Hamdi Ben Abdessalem and Claude Frasson

Département d'Informatique et de Recherche Opérationnelle
Université de Montréal, Montréal, Canada H3C 3J7
{ mahsa.aghajani,hamdi.ben.abdessalem}@umontreal.ca,
frasson @iro.umontreal.ca

Abstract. This paper reports the results of voice emotion recognition in real time using machine learning models. The models are trained with some commonly used and well-known audio emotion datasets together with a custom dataset. This custom dataset was recorded from non-actor and non-expert people who were trying to imagine themselves in scenarios leading to arise of the related emotion. The reason for considering this important dataset is to make the model proficient in recognizing emotions in people who are not perfect in reflecting their emotions in their voices. The results from several machine learning classifiers while recognizing five emotions like anger, happiness, sadness, neutrality and surprise are compared. Models were evaluated with and without considering the custom data set to show the effect of employing an imperfect dataset. Our experiments showed that without using our custom dataset, the ensemble machine learning models such as gradient boosting, begging and random forest reach validation accuracies 89.82%, 88.58% and 84.83% respectively, which are higher than other evaluated models. After considering our custom dataset, again these ensemble methods obtained better accuracies of 87.34%, 86.71% and 82.98% respectively. This shows that although considering our custom dataset lowers the overall accuracy but empowers the model for predicting the emotions in everyday scenarios.

Keywords: Voice Emotion Recognition, Machine Learning, Brain Computer Interface, Affective Computing

1 Introduction

In today's applications, recognizing emotions in users' voices leads to a better user experience and more user-friendly applications. In applications where the user's satisfaction is an important factor or the competitive advantage of that application, detecting the emotion of the user through interacting with the application becomes of great value. Examples of these applications are customer support applications, artificially intelligent virtual assistants, etc. [1, 2]. Along with these mentioned applications, many other applications can also benefit from voice emotion detection hugely but have some specific requirements too. One of these specific requirements for this set of applications is that the emotion detection should be done in real time. Instances of such applications are educational applications, video games, driving assistant applications, etc. The need for real time emotion detection in these applications originates from enabling the application to behave differently based on the current emotion of the user. For example, if the student is stressed, angry or sad, the hardness level of following materials can be reduced; On the other hand, getting positive feedback from the student's emotion, can allow us to increase the hardness of the upcoming material. For Alzheimer patients detecting negative emotions would be important to apply for instance relaxing techniques.

For detecting the emotion from voice, we have used several machine learning models. For training these models, there are a few rich datasets. In these datasets, usually the speakers are actors, who are most of the time experts in showing and reflecting their emotions in their voices. Although this characteristic of these commonly used datasets, makes them ideal for training and validating machine learning models, but debilitates the models in predicting the emotions in non-actor and non-expert people's voices in everyday scenarios. Therefore, in addition to these available datasets, we need other datasets, recorded from non-actor people for training and validating the models.

Detecting the emotion using the speaker's voice in real time applications has sever-al important challenges; First the nature of this detection is a completed task, even for humans. In different scenarios, based on the person's ability to reflect his emotions in his voice, this recognition task can get even harder. For example, when the speaker suffers from the Alzheimer disease or when the speaker has autism. Even with people who do not have these conditions, detecting several emotions from each other in their voice, such as happiness and surprise, or neutrality and sadness requires high proficiency. In all these cases, the predictor model should already be trained with similar data; Second, the feature set for machine learning models should be selected carefully. The third challenge is related to the real time emotion detection requirement of the target applications. Considering this specific need, the emotion detection should be done in an acceptable amount of time.

The outline of this paper is as follows: section 2 reviews related work in voice emotion recognition. Section 3 introduces the datasets we have used. In section 4 we de-scribe our methodology and machine learning models. Section 5 is about explaining the experiments and reporting their results. Section 6 concludes our work and discusses our future works that may lead to improvements.

2 Related Work

Voice emotion recognition has been around for decades[3, 4, 5]. In most of these works, the data collection process was done in a studio environment(F. Tao et al., 2018), leading to using clear audio files for training the models. In (F. Tao et al., 2018) it is aimed to use a corpus that has background noise and is close to the real world. The authors used the Multimodal Emotion Challenge (MEC) 2017 corpus. This corpus contains clips from films and TV programs and the speakers are professional actors; therefore, the problem of having models that are trained with emotion reflective voice is still remaining.

H. Chen et al in (Chen et al., 2019) used CASIA Chinese Emotional Speech Corpus in training several models. The authors report SVM as the best model with the highest accuracy of 81.11%. A perfect dataset, from reflecting the emotional point of view, is also considered in this work.

S. Yacoub et. al in (Yacoub et al., 2003) have focused on extracting features from short utterances which are commonly used in Interactive Voice Response (IVR) applications. Authors have recognized the anger and neutral emotions from each other with an accuracy of 90% with models trained over the Linguistic Data Consortium at University of Pennsylvania. In our work, we obtained the accuracy of 90% and 87% (without and with our custom dataset respectively) while predicting five emotions at the same time.

In (Shaqra et al., 2019) the relation between age and gender and the emotion recognition accuracy is studied. The authors have obtained an accuracy of 74% by using a hierarchal model, trained and validated over the RAVDESS (Livingstone & Russo, 2018) dataset, which contains speakers of different ages.

Our work's novelty is in gathering and using non-expert and non-actor speaker voices, trying to involve them in scenarios leading to emerging the related emotion. We also evaluated several machine learning models, trying to predict 5 emotions at the same time. Due to the covid-19 restrictions, while conducting our experiments, we were not able to extend our validation techniques with EEG experiments able to directly provide emotions assessments. After returning to the normal situation, we will resume our validation procedures with EEG experiments.

3 Datasets

In all machine learning applications, selecting the proper dataset is extremely important. There are many different datasets for voice emotion recognition (El Ayadi et al., 2011). In our work, Toronto Emotional Speech Set (TESS) (Pichora-Fuller & Dupuis, 2020), Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Surrey Audio-Visual Expressed Emotion (SAVEE) Database (*Surrey Audio-Visual Expressed Emotion (SAVEE) Database*, n.d.) and a Custom Database these datasets are used for training and validating the models. In our custom dataset, audio files were recorded from non-actor speakers whom were asked to say different sentences while trying to imagine themselves in a situation causing them to have the related emotion. Table 1 shows the distribution of audio files related to each emotion in the final dataset which is the result of appending all the above datasets:

Table 9.1. Distribution of audio files related to each emotion in the whole dataset

Emotion	Final Dataset	
	Count	Proportion
Angry	677	0.202
Happy	677	0.202
Neutral	641	0.191
Surprised	677	0.202
Sad	677	0.202

4 Methodology

After gathering a rich dataset from mentioned emotional databases, we extracted appropriate features from the audio files. The extracted feature set was fed into several machine learning classifiers separately. For validating the performance of each classifier, the input data set was split into training and validation sets under two different approaches. In the first approach, 25% of the input data set was randomly chosen as the validation set. In the second approach, 25% of the input data set was selected as the validation set in an actor-based approach; that is there was no common actor between the training and validation sets. The goal of this approach was to validate the performance of the classifiers dealing not only with unseen data but also when the input voice is completely new for the classifier. For feature extraction, we have used the feature set provided in INTERSPEECH 2010 paralinguistic challenge (Schuller et al., 2010), which is a common selected feature set among related works. These features are fed into several classifiers such as SVM, random forests, bagging, gradient boosting and RNN.

5 Results and Discussion

For validating the models, we considered two datasets; The first dataset contained TESS, RAVDESS and SAVEE datasets, which are all common speech emotion recognition datasets, recorded by actor speakers. In the second case, we added our custom dataset to the previous datasets.

For validating the prediction results of classifiers trained and tested with these two datasets, we have used 5-folds cross fold validation technique with the same distribution of classes in each round of validation, using StratifiedKFold class from Sklearn library.

The predicted results of classifiers trained and tested with these two datasets are explained in the following parts.

5.1 Considering TESS, RAVDESS and SAVEE datasets

In this case, classifiers were trained with TESS, RAVDESS and SAVEE dataset. Table 2 contains the average performance measures of each classifier on the test set. The results for each performance measure are the averages over the results of that classifier over all the 5 repetitions of 5-old cross validations. As it can be seen, gradient boosting, bagging and random forest have the best average accuracies of 89.82%, 88.58% and 84.83% respectively. These classifiers have also the overall best performance measures among all the classifiers which shows that we can rely on ensemble methods for speech emotion prediction applications.

Table 2. Average performance measures of each classifier on the test set

Classifier	Accuracy	F1-score	Precision	Recall
SVM	0.7795	0.7794	0.7811	0.7795
Random Forest	0.8483	0.8482	0.8506	0.8483
Bagging	0.8858	0.8855	0.8862	0.8858
Gradient Boosting	0.8982	0.8981	0.8995	0.8982
RNN	0.7444	0.7454	0.7744	0.7193

5.2 Considering TESS, RAVDESS, SAVEE and custom datasets

In the second case, the dataset included our custom dataset together with three previously mentioned datasets. In this approach, the total number of audio files reached 3349 files. Again, we validated our models using 5-fold cross validation.

Since the speakers in the recorded audio files in our custom dataset were not professional actors, we anticipated a decrease in the accuracy of the classifiers, which is also shown in Table 3.

Table 3. Average performance measures of each classifier on the test set in the second approach

Classifier	Accuracy	F1-score	Precision	Recall
SVM	0.7647	0.7639	0.7669	0.7647
Random Forest	0.8298	0.8296	0.8328	0.8298
Bagging	0.8671	0.8669	0.8681	0.8671
Gradient Boosting	0.8734	0.8732	0.8748	0.8734
RNN	0.7082	0.7077	0.7093	0.7061

Figure 1 shows another comparison between the accuracies of the classifiers on the test set after adding our custom dataset. In this case, the gradient boosting and bagging classifiers had the best prediction results with average test accuracies of 87.34% and 86.71% respectively.

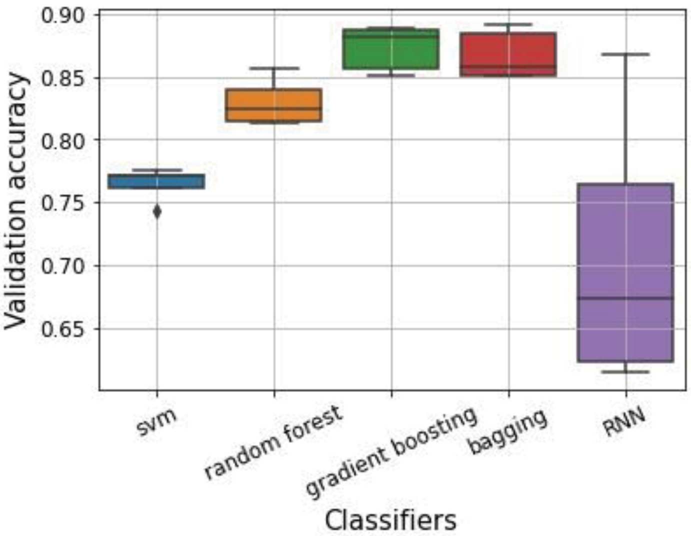


Fig. 1. Validation accuracies of classifiers in the second approach

Considering figure 2, which is the confusion matrix for gradient boosting as our best classifier, it can be observed that predicting neutrality has the best accuracy of 93.0% among other emotions. On the other hand, predicting happiness has the lowest accuracy of 84.0%. It should also be noted that among the audio files that are predicted as angry ones, 4.7% of them are truly happy audio files. This shows that anger and happiness are the most confusing emotions when predicting angriness. The same scenario holds when the emotion is predicted to be neutrality, that is 4.1% of the files recognized to have neutrality as their most dominant emotion, are in fact sad. With the same explanation, the most confusing emotion while predicting surprise is happiness.

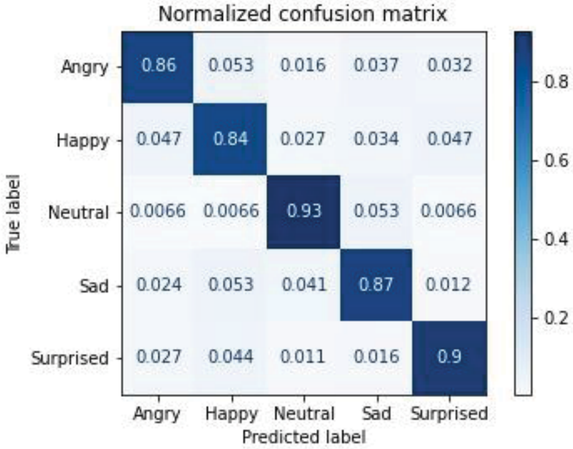


Fig. 2. Confusion matrix of gradient boosting classifier with 5 emotions

6 Conclusion and Future Work

In this study, we reported that by using audio features provided in INTERSPEECH 2010 paralinguistic challenge, we could get acceptable prediction accuracies from several ensemble classifiers. We also explained why considering datasets recorded from non-actor speakers is necessary while training the machine learning models. We tried to tackle this challenge by gathering a custom data set. We also discussed the most confusing emotion pairs in our experiments.

Since in real case scenarios, emotions may not always be reflected in the speaker's voice noticeably, getting help from spoken words may help in predicting the true emotion. Regarding this, we are going to use speech to text techniques and add the recognized words to our feature set. We are also going to increase the size of our custom dataset to evaluate its effect in real case emotion detection scenarios. In this work we recognized 5 emotions, considering the limitations in gathering the custom dataset; We are going to expand our considered emotion set to 7 emotions of angry, happy, sad, neutral, disgust, fear and surprised. With all of this, we are going to extend our testing procedures using EEG experiments able to measure emotions and compare them with our current and modified classifiers.

Acknowledgment. We acknowledge NSERC-CRD (National Science and Engineering Research Council Cooperative Research Development), Prompt, and BMU (Beam Me Up) for funding this work.

References

1. V. Petrushin, “Emotion in speech: Recognition and application to call centers,” in Proceedings of artificial neural networks in engineering, 1999, vol. 710, p. 22.

100

2. V. A. Petrushin, “Emotion recognition in speech signal: experimental study, development, and application,” 2000.

3. C. Busso et al., “Analysis of emotion recognition using facial expressions, speech and multimodal information,” in Proceedings of the 6th international conference on Multimodal interfaces, 2004, pp. 205–211.

4. C. M. Lee et al., “Emotion recognition based on phoneme classes,” 2004.

5. J. Deng, X. Xu, Z. Zhang, S. Frühholz, D. Grandjean, and B. Schuller, “Fisher kernels on phase-based features for speech emotion recognition,” in Dialogues with social robots, Springer, 2017, pp. 195–203.

6. F. Tao, G. Liu, and Q. Zhao, “An ensemble framework of voice-based emotion recognition system for films and TV programs,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 6209–6213.

7. H. Chen, Z. Liu, X. Kang, S. Nishide, and F. Ren, “Investigating voice features for Speech emotion recognition based on four kinds of machine learning methods,” in 2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems (CCIS), 2019, pp. 195–199.

8. S. Yacoub, S. Simske, X. Lin, and J. Burns, “Recognition of emotions in interactive voice response systems,” 2003.

9. F. A. Shaqra, R. Duwairi, and M. Al-Ayyoub, “Recognizing emotion from speech based on age and gender using hierarchical models,” Procedia Comput. Sci., vol. 151, pp. 37–44, 2019.

10. S. R. Livingstone and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English,” PloS One, vol. 13, no. 5, p. e0196391, 2018.

11. M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” Pattern Recognit., vol. 44, no. 3, pp. 572–587, 2011.

12. M. K. Pichora-Fuller and K. Dupuis, “Toronto emotional speech set (TESS).” Scholars Portal Dataverse, 2020,

13. “Surrey Audio-Visual Expressed Emotion (SAVEE) Database.” http://kahlan.eps.surrey.ac.uk/savee/ (accessed Jan. 05, 2021).

14. B. Schuller et al., “The INTERSPEECH 2010 paralinguistic challenge,” 2010.

15. F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 1459–1462.

16. R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” in international conference on machine learning, 2013, pp. 1310–1318.