

DATA MINING FOR EMOTION RECOGNITION IN SPEECH



GAMZE AKKURT

JULY 2019

DATA MINING FOR EMOTION RECOGNITION IN SPEECH



A THESIS SUBMITTED TO
THE GRADUATE SCHOOL
OF
IZMIR UNIVERSITY OF ECONOMICS

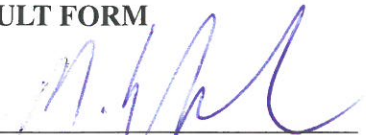
BY
GAMZE AKKURT

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE
IN THE GRADUATE SCHOOL

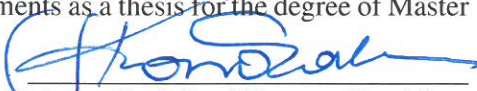
JULY 2019

M.S. Thesis EXAMINATION RESULT FORM

Approval of the Graduate School


Prof. Dr. M. Efe Biresselioğlu
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.


Assoc. Prof. Dr. Süleyman Kondakçı
Head of Department

We have read the dissertation entitled “**DATA MINING FOR EMOTION RECOGNITION IN SPEECH**” completed by **GAMZE AKKURT** under supervision of **Assoc. Prof. Devrim ÜNAY** and **Asst. Prof. Umut AVCI** and we certify that in our opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Master of Science.


Assoc. Prof. Devrim ÜNAY
Supervisor

Examining Committee Members

Date: 08.07.2019

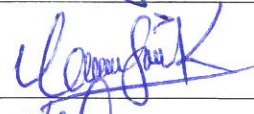
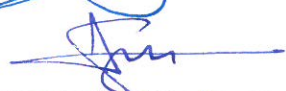
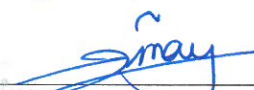

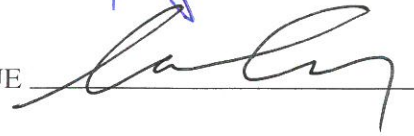
Assoc. Prof. Devrim ÜNAY
Dept. of Biomedical Engineering, IUE

Asst. Prof. Umut AVCI
Dept. of Software Engineering, Yaşar U.

Prof. Dr. Cem EVRENDİLEK
Dept. of Computer Engineering, IUE

Asst. Prof. Dr. Korhan KARABULUT
Dept. of Software Engineering, Yaşar U.

Asst. Prof. Dr. Mehmet TÜRKAN
Dept. of Electrical and Electronics Engineering, IUE

ABSTRACT

DATA MINING FOR EMOTION RECOGNITION IN SPEECH

GAMZE AKKURT

M.S. in Computer Engineering

Graduate School

Supervisor: Assoc. Prof. Devrim ÜNAY Co-Supervisor: Asst. Prof. Umut AVCI

July 2019

The popular features used in speech signal for emotion classification are fundamental frequency, voice quality, energy, spectral, and MFCC. While most of the work focuses on these acoustic features in speech emotion recognition, we handle the problem of emotion recognition using features that are obtained from emotional patterns. In our approach, we transform the speech signal to discretized signal and extract distinctive patterns that can distinguish between different emotions. Then, a set of feature vectors is created using extracted patterns in order to feed a classifier. Experimental results indicate that the proposed approach learns the emotional state of speech efficiently from both pattern-based features and acoustic features that are supported by pattern features. Pattern-based features have resulted in 35 % improvement in accuracy using two classifiers compared to state of the art acoustic features. Moreover, when all acoustic features are combined with pattern-based features, classification accuracy enhances over 80 % in emotion recognition.

Keywords: Emotion recognition, Speech Processing, Pattern Mining, Feature Extraction.

ÖZ

SESTE DUYGU TANIMA İÇİN VERİ MADENCİLİĞİ

GAMZE AKKURT

Bilgisayar Mühendisliği, Yüksek Lisans

Lisansüstü Eğitim Enstitüsü

Tez Danışmanı: Doç. Dr. Devrim Ünay İkinci Tez Danışmanı: Dr.Öğr.Üyesi Umur

AVCI

Temmuz 2019

Konuşma sinyalinde duygu sınıflandırması için kullanılan popüler özellikler temel frekans, ses kalitesi, enerji, spektral ve MFCC'dir. Çalışmaların çoğu konuşmadaki duyguların tanınmasında bu akustik özelliklere odaklanırken, bu tezde biz; duygusal kalıplardan elde edilen özellikleri kullanarak duygu tanıma sorunu ele alınmıştır. Yaklaşımımızda, konuşma sinyalini ayrıklaştırılmış sinyale dönüştürür ve farklı duygular arasında ayırım yapabilen ayırt edici kalıplar çıkartılmaktadır. Ardından, sınıflandırıcıyı güçlendirmek için; çıkartılan kalıplarla bir dizi vektör özelliği oluşturulur. Deneysel sonuçlar, önerilen yaklaşımın, hem desene dayalı özelliklerden hem de desene ait özelliklerle desteklenen akustik özelliklerden duygusal konuşma durumunu etkili bir şekilde öğrendiğini göstermektedir. Desen bazlı özellikler, son teknoloji akustik özelliklere kıyasla iki sınıflandırıcı teknik kullanılarak doğrulukta % 35 'lik artış ile sonuçlanmaktadır. Ayrıca, bütün akustik özellikler, desen bazlı özellikler ile desteklendiğinde % 80 'nin üzerinde artış göstermektedir.

Anahtar Kelimeler: Duygu Tanıma, Ses İşleme, Desen Madenciliği, Özellik Çıkarma.

ACKNOWLEDGEMENT

To begin with, I would like to express my sincere gratitude to my professors Umut Avcı and Devrim Ünay, who expertly guided me through my thesis and shared my excitement. I'm grateful to work with them on my research.

Also, I extend my appreciation to my family, my father Harun, my mother Nevin and my brother Gökhan. Thanks for their endless support and love.

I would like to thank Levent Tolga Eren for always being encouraging and believing in me.

TABLE OF CONTENTS

Front Matter	i
Abstract	iii
Öz	iv
Acknowledgement	v
Table of Contents	vii
1 Introduction	3
1.1 Related Work	4
2 Methodology	8
2.1 Database	9
2.1.1 The RAVDESS Dataset	12
2.2 Preprocessing	12
2.3 Dimension Reduction	14
2.4 Discretization	14
2.5 Pattern Mining	16
2.6 Pattern Features	17
2.7 Acoustic Features	18
2.8 Classification Schemas	21
2.8.1 Maximum Voting Algorithm	21
2.8.2 SVM	22
3 Results and Analysis	24
4 Conclusion	40
4.1 Future Work	41

A Tools	51
A.1 OpenSMILE	51
A.2 LibSVM	53
A.3 Matlab	53



LIST OF TABLES

2.1	Characteristic of emotional speech databases	11
3.1	The Computational Duration	27
3.2	Experimental set 1	28
3.3	Experimental set 2	28
3.4	Experimental set 3	29
3.5	Experimental set 4	29
3.6	Experimental set 5	29
3.7	Experimental set 6	30
3.8	Experimental set 7	30
3.9	Experimental set 8	31
3.10	The classification results of acoustic features	31
3.11	The classification results of pattern-based features	32
3.12	The classification results of combining acoustic and pattern based features (12 statistics were applied to acoustic features)	33
3.13	The classification results of combining acoustic and pattern based features (All statistics were applied to acoustic features)	33
3.14	Classification accuracy for different alphabet size (%)	34
3.15	Classification accuracy for different frame size (%)	34
3.16	Classification accuracy for 6 Frame-5 Alphabet (%)	35
3.17	McNemar Test Results One Versus All & One Versus One Pattern-Pased Features	36
3.18	McNemar Test Results One Versus All & One Versus One Pattern-Pased and Acoustic Features (12 Statistics)	37
3.19	McNemar Test Results One Versus All & One Versus One Pattern-Pased and Acoustic Features (All Statistics)	37

3.20 Confusion matrix of experimental set 3	38
3.21 Confusion matrix of Energy+100 Pattern One versus All (S1, S2 separate)	38



LIST OF FIGURES

2.1	Basic system diagram of speech emotion recognition.	9
2.2	(a) A speech signal visualization: x-axis coordinate represents speech samples, and y-axis coordinate represents speech amplitude, (b) Represents speech signals after removing silent parts	13
	a Original Speech Signal	13
	b After Removing Silence Part	13
2.3	A speech signal is normalized with z-score normalization	14
2.4	Breakpoints of Gaussian Distribution (Source: [1])	15
2.5	(a) t number of a speech signal is reduced t/w. In this example, a speech signal of length 88 is divided by 8 window size and it is reduced to 11 dimensions, (b) signal is discretized depending on Gaussian distribution table	15
	a Dimension Reduction	15
	b Discretization	15
3.1	Maximum voting results	25
3.2	DAGSVM classification results	26

List of Abbreviations

- ADABOOST** Adaptive Boosting Algorithm. 6
- ANN** Artificial Neural Network. 6
- ComParE** Interspeech 2013 ComputationalParalinguistics Evaluation. 20
- ConsgapMiner** Contrast Sequences with Gap Mine. 16
- DAGSVM** Directed Acyclic Graph Support Vector Machines. 4, 21, 23
- DSP** Distinguish Sequential Pattern Mining. 16
- GepDSP** Gene Expression Programming Distinguishing Sequential Pattern. 16
- HCI** Human Computer Interaction. 3
- HMM** Hidden Markow Models. 6
- iDSP-Miner** item set Distinguish Sequential Pattern Mining. 16
- KDSP-Miner** Top-k Distinguishing Sequential Patterns with gap constraint. 16
- LDC** Linear Discriminative Classifier. 6
- LLDs** Low-Level Descriptors. 18
- LPCC** Linear Prediction Cepstral Coefficient. 6
- MEDC** Mel Energy Spectrum Dynamic Coefficients. 6

MFCC Mel Frequency cepstral coefficients. 4, 6, 19

openSMILE Munich open-Source Media Interpretation by Large feature-space Extraction. 51

RAVDESS Ryerson Audio-Visual Database of Emotional Speech and Song. 11, 12

S2ST Speech to Speech Translation. 4

SVC Support Vector Classification. 53

SVM Support Vector Machine. 6, 22, 53

Chapter 1

Introduction

Speech is one of the most effective and fastest ways to provide communication between humans. It has inspired researchers to find efficient methods for human-computer interaction (HCI). HCI approaches employ speech recognition, gesture recognition, etc. to understand human intention and recognize speech from a human voice. Despite substantial improvements in speech recognition research, allowing 'natural interaction' between user and machine, i.e. recognizing and understanding the emotional state of the speaker is a crucial but difficult task. Therefore, speech emotion recognition aims to extract useful information from speech signals and improve emotion recognition performance [2].

Speech emotion recognition is widely used in several application areas of HCI. In the automotive industry, it may help to obtain information about the mental state of the driver to avoid accidents and provide safety for a driver via in-car board systems [3]. In call center applications, speech emotion recognition may be used to analyze customer behavior and provide emotional feedback of the customer to the call center operator to improve the quality of the call center service [4],[5].The medical industry also uses speech emotion recognition to detect the mental state of patients so that they can use the data of patients as a diagnostic tool in mental health problems like depression, suicide cases or lie detection [6]. It may be used in noisy environments to reach better system performance, particularly in aircraft cockpits, where the speakers are having a vital role in communication. As the pilots face a high level of emotional stress while performing hard tasks, speech emotion recognition is trained by the stress in the

speech and improves overall system performance [7]. Also, robots have an important role and increased popularity in our society. Enhancements in the humanoid robot technologies make our lives easier than ever. Today, most people use those smart robots in their daily tasks. Amazon Alexa and Google Home Assistant are examples of those robots. People often use smart assistant devices to perform tasks such as playing music, searching for things on the Internet or asking for a question, etc. In order to use those devices, people use their voices as a command. The robots interpret the voice command and perform relevant tasks. Apart from that, another application area, which is speech to speech translation system (S2ST), is used in emotion recognition applications. S2ST is a process that a spoken speech in a language is used to generate a spoken output in another language. In this process, both emotion recognition and synthesis are used. The emotional state of a speaker is recognized by the system, and that emotion will be converted into another language with the same emotional state. This is what precisely S2ST process does [8].

In this work, we propose a novel approach to extract a new set of features for emotion recognition from speech. Our approach consists of two phases. First, we convert a speech signal to a discretized representation and extract the most related patterns that define emotion from this representation using a data mining algorithm. Later on, we generated a feature vector for each pattern of each emotion by counting the frequency of the pattern that exists on the discretized signal. Secondly, we use existing techniques in the literature to extract acoustic features from the speech signal, such as mel frequency cepstral coefficient (MFCC), voice quality, energy and spectral. Finally, we combine our proposed pattern features with these acoustic features to improve classification accuracy. Specifically, in classification, we used two different classification techniques, namely maximum voting and Directed Acyclic Graph Support Vector Machines (DAGSVM).

1.1 Related Work

There are many studies for recognizing emotions in the literature. These studies are divided into two different domains that can include either unimodal (speech-only or visual-only) or multimodal (audiovisual, i.e., speech +facial and body expressions)

data. Most of the studies show that speech and facial expressions are related to the influence of emotions. Facial expression recognition uses video or image sequences where facial expressions only represent the object without any speech. Facial features are extracted from the nose, lips, eyebrows, and eyelids. Essa et al. [9] used optical flow method identifying a computer vision system to observe facial motion. There are six universal statements (angry, sad, happy, surprise, disgust, fear) in facial expression that are introduced by Ekman [10]. Most of the studies ([11],[12],[13],[14]) used these statements in order to classify the emotions. It is difficult to collect information from authentic facial expression because they are uncommon and filled with subtle context-based changes, which make it difficult to detect emotions without affecting the outcomes [15].

Speech emotion recognition belongs to unimodal data. It uses only audio information to distinguish emotions. Acoustic parameters are used for features, i.e., energy, pitch duration, and spectral. Yixiong [16] explained the acoustic features in the human emotion by using Berlin database. Several statistical pattern recognition techniques were noted by Dellaert et al.[17] to classify emotional speech according to emotional content. Ververidis et al. [18] classified anger, happiness, neutral, sadness, and surprise emotions in Danish database. The paper proposed using virtual reality to measure the effect of the emotional content of speech. In order to detect the stress in the speech, Kwon et al. [19] used acoustic features such as pitch, MFCC log energy formants, and Mel frequency bands in SUSAS database. Chavhan et al. [20] used spectral features to recognize seven different emotions in Berlin database.

On the other hand, the modularity of multimodal data is highly correlated with speech and other expressions such as facial and body expressions. Castellano et al. [21] used face, body gesture, and speech for eight emotions (anger, despair, interest, pleasure, sadness, irritation, joy, and pride) with a Bayesian classifier in order to analyze three modularities on recognition of emotion. Busso et al. [22] used audio and video information for recognizing four different emotions. These emotions are sadness, happiness, anger, and neutral. Also, the paper examined the system limitation depending on facial expression or acoustic information.

Researchers have been using various feature extraction and classification methods to

identify emotions from speech. Feature extraction is a crucial process in speech emotion recognition as different types of features capture different information present in speech such as speaker's emotion and language. The most extracted features in the literature are those capturing prosodic information from speech such as energy, format, fundamental frequency, MFCC, speaking rate, and shimmer. Various statistical measures like mean, median, skewness and kurtosis are then calculated from these extracted features. In addition, after feature extraction, a wide variety of classification methods is used to recognize speech emotion. These classification methods include Hidden Markov Models (HMM), Artificial Neural Network (ANN), SVM, boosting and ADABOOST learnings. Kopal et al. used prosodic features and functionals that have mean, median, kurtosis and skewness to identify five emotions namely happiness, sadness, anger, fear and neutral. They reported classification accuracies as 69.41% using SVM and 61.97% using random forest [23]. Pan et al. analyzed the discriminative power of energy, pitch, LPCC (linear prediction cepstral coefficient), MFCC and MEDC (Mel energy spectrum dynamic coefficients) for emotion classification. They used two different datasets and SVM for classification. The classification accuracy for three emotions were reported as 95,1% and 91,3% respectively [16]. Lee et al. analyzed a database of real world recordings from call center conversations using fundamental frequency, energy, duration and formants. The results of the research revealed that combining all the information (acoustic, lexical, and discourse) enhanced emotion classification by 40,7% for males and 36.4% for females. (Linear discriminant classifier (LDC) is utilized for acoustic features [24]. Nicholson et al.[25] used phonetic and prosodic features together with neural networks to analyze a database of radio actors. The accuracy of classification was obtained about 50%.

Also, there are several studies available in the literature that use pattern mining in feature extraction. However, these studies differ in the domain of application, such as EEG, face, and music. Cabroda et al. [26] proposed another approach to identify music features that impress emotion. In that paper, they identified patterns in psychophysiological data utilizing a motif discovery algorithm and assessed the music elements. Shan et al.[27] introduced LBP for facial expression recognition, and they proposed that the face images could be thought as a combination of micro-patterns which may be well described by LBP. Also, Tweri et al.[28] focused on motif face features for

EEG data to recognize sleep state and the effect of anesthesia. Zao et al. [29] presented dynamic or temporal textures that are textures with motion in video sequences. They modeled textures by using volume local binary patterns. To our knowledge, this is the first study to suggest using pattern mining for emotion recognition from speech.

The rest of this thesis is organized as follows. In chapter 2, we respectively describe the database, the methodology, the related details to extract our novel pattern features, acoustic features, and the classification schemes. In chapter 3, we present the experimental results and analysis. Finally, the concluding notes and the future work explained in chapter 4.

Chapter 2

Methodology

Speech is an important form of emotional expression. The voice in speech not only contains a grammatical message but also information about the speaker's emotional status. Neutral, calm, happy, sad, angry, and fearful are those examples of emotions. The fundamental issues which need to be taken into consideration for successful speech emotion recognition are feature extraction and classification respectively [30]. It is considered that a correct selection of features remarkably affects the classification performance[2].

Feature extraction is the principal issue in speech emotion recognition. Feature extraction is a particular form of data-set, and it ends up the extraction of specific features about the speech. By extracting those features, we obtain the defining characteristic of speech signals. Those signals may contain information about a speaker, vocabulary, language, and emotion. These parameters may affect the accuracy in speech emotion recognition. In the literature, there are many feature extraction algorithms available. These are MFCC, RASTA filtering, Linear Prediction Cepstral Coefficient (LPCC), and Linear Prediction Coefficients (LPC), etc. Features can be extracted by using one of those algorithms. We applied this technique to result in the extraction of specific acoustic and pattern features that we deal with.

Emotions have an impact on speech's acoustic characteristics. These characteristics are detected by prosodic and spectral features using feature extraction technique. It is possible to find these acoustic features in most of emotion recognition studies. In

our research, we proposed an approach to identify specific acoustic patterns for different emotions. With this purpose, speech signals were converted into strings by discretization. Discrete representations of the signal were utilized to extract patterns of emotion in a discriminative way. Then, recognition of emotion was fulfilled with comprehensive features from the patterns. Also, these features were combined with existing prosodic and spectral features in the literature. The approach is outlined in Figure 2.1. At first, we will describe databases that exist in the literature and we will give information about our databases. Then, we will explain the approaches in the following sections in detail.

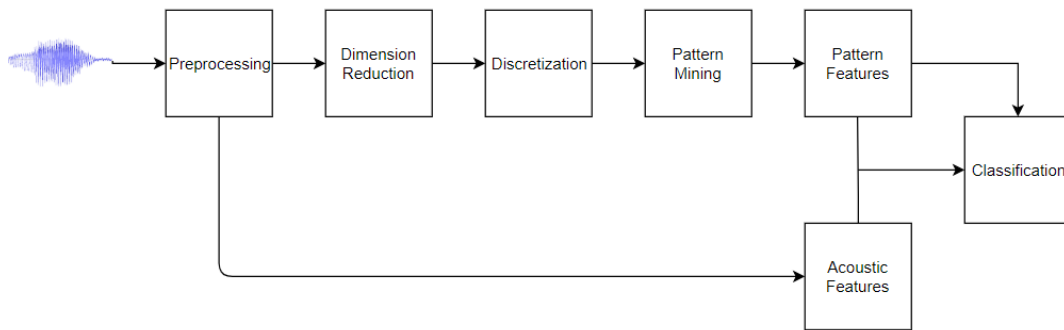


Figure 2.1 Basic system diagram of speech emotion recognition.

2.1 Database

Databases have a significant role in terms of recognition and appraisal of emotional speech. The degree of naturalness of the database affects emotion recognition performance. When a database with a low degree of naturalness is utilized, imprecise results may be constructed [2]. On the other hand, a high degree of naturalness provides more realistic and reliable results in emotion recognition. For this reason, the design of the databases is organized for the classification task. In the classification task, e.g., emotions can be classified by adult-directed or infant-directed. In terms of infant-directed emotions ([31],[32]), utterances are collected by parents who talk with their infants. Each utterance is classified as approval, attention bids, or prohibition. In

adult-directed emotions [33], utterances are recorded from males and females who express their emotions such as happy, angry, sad, and so on. Except for these emotions, other databases detect stress in the speech for the classification task [34]. Also, the number and type of emotions are determined by the classification task according to the content of the databases. In the literature, there are three types of databases available. These are actor (simulated) based, elicited, and natural emotional speech databases.

Actor based emotional speech databases are collected from real life situations such as radio broadcast or theater. Actors express natural sentences with different emotions. This is the most reliable and commonly used technique to collect data [35]. Also, these databases are available in a wide variety of languages and include all emotions.

Elicited emotional speech databases are collected from sound laboratories. In these laboratories, speakers talk about different situations which are created by the presenter. Different emotions are elicited without knowledge of the speakers. One of the drawbacks of these databases is that the speakers may exaggeratedly express emotions when they realize being recorded. Therefore, emotions are not the same as real emotions, and they are less realistic than actual emotions [35].

Natural emotional speech databases are collected from real-world conversations. They are composed from call center conversations, patient and doctor dialogues, cockpit, and so on. Natural emotions are slightly expressed in real-world conversations. Thus, it cannot be easy to recognize all emotions in real conversations. Also, there may be privacy and copyright issues to access them. Therefore, it can be hard to deal with legal issues to utilize these databases for research purposes [35].

As described before, a wide variety of designed databases including various information are available for research purpose. Hence, databases may provide data such as language, the number of emotions, naturalness, source, actors, linguistic, accessibility, etc. Table 2.1 shows the most frequently used speech emotion databases and their characteristics. As shown, most of the databases use adult-directed emotions except for BabyEars and KISMET, and also they share the same emotions that are joy, anger, fear, disgust, and sadness. Both professional and non-professional actors perform in all databases. Furthermore, different languages are available in the table. Moreover, linguistic column shows the content of the speech as well as the words or sentences to be pronounced. In Hedrew database, the subjects refer to the speech of an incident that

was to remember an emotional event in the past, pronouncing as if feeling the same experience as they felt at the first experiment. However, most of the databases are not directly accessible for public use due to copyright and privacy issues.

In our research, we used the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [36]. It's publicly available, and it can be used for research purposes. It contains a number of emotions and both female and male actors are available. There are lots of sample speech data, so this helped us to retrieve and recognize speech emotions from these samples. These are the main reasons that we used this database in our research. We will describe the details in the following section.

Table 2.1 Characteristic of emotional speech databases

Corpus	Actors	Naturalness	Language	Emotions	Linguistic	Source	Accessible
Berlin [37]	5 males, 5 females (10 Actors)	Simulated	German	Neutral, anger, fear, joy, sadness, disgust, freedom	10 sentences	Professional actors	Public
Danish [33]	2 males, 2 females (4 Actors)	Actor based	Danish	Neutral, surprise happiness,sadness anger	9 sentence 2 words	Professional actors	Public
SUSAS [34]	19 males, 13 females (32 Actors)	Natural	English	Depression, fear anxiety, anger	interview	Natural speech	Private
BabyEars [32]	6 mothers, 6 fathers	Natural	English	Approval, attention prohibition	single sentence or phase	Natural speech	Private
KISMET [31]	3 females	Elicited	American English	Approval, attention prohibition, soothing neutral	1002 utterances	Nonprofessional actors	Private
INTERFACE [38]	1 male, 1 female (2 actors)	Elicited	English, Spanish, French, Slovenian	Anger,fear,sadness joy, disgust, surprise	8928 sentence English 6080 sentence Slovenian 5600 sentence French 5520 sentence Spanish	Actors	Private
RUSLANA [39]	12 males, 49 females (61 actors)	Elicited	Russian	Surprise, happiness anger, sadness, fear	3660 utterances	Unprofessional actors	Private
Hedrew [40]	19 males, 21 females	Elicited	Hedrew	Anger, fear, joy, sadness, disgust	subject	Nonprofessional actors	Private
Yu et al. [41]	4 students	Elicited	Chinese	Anger, happiness, neutral, sadness	2000 utterances	Nonprofessional actors	Private
Petrushin et al. [4]	30 actors	Natural	English	Happiness, anger, sadness,fear	700 utterances	Professional actors	Private
RAVDESS [36]	12 males, 12 females	Elicited	English	Happy, anger,neutral sad,fear,neutral disgust,surprise	2 sentences	Professional actors	Public

2.1.1 The RAVDESS Dataset

RAVDESS dataset is a multi-modal database which has emotional speech and song samples. This dataset is utilized by researchers for examining the similarities and differences between acoustic and visual signals of emotional speaking and singing. The visual signals refer to the content of a video which sounds are derived from. RAVDESS includes audio and video recordings of 24 actors (12 female and 12 male) speaking and singing. In the recordings, there are two types of sentences which are 'Kids are talking by the door' and 'Dogs are sitting by the door'. Each actor expresses these sentences with different emotions in a North American accent. The speech recordings include eight different adult-directed emotions. These emotions are neutral, calm, happy, sad, angry, fearful, disgust, and surprise. Except for disgust and surprise emotions, the song recordings consist of six emotions. All emotions are performed at normal and strong intensities, and each emotion with two repetitions. Also, only neutral emotion does not have strong intensity. In the dataset, each actor performs 60 spoken and 44 sung specific vocalization. The total number of specific vocalization is 2452 (24 actors * 60 utterances + 23 actors * 44 utterances) available in all three modularities that are audio, video, and audio-video.

In our research, we only used speech signals with normal intensity from RAVDESS dataset. The research that we use as a criteria [42] focuses on a domain-independent recognition of emotions such as song and speech. Only common emotions are paid attention to both domains. We excluded disgust and surprise emotions from the speech recording in order to be coordinated with this work. Total number of audio utterances is 576 in (24 actors * 2 sentences * 2 repetitions * 6 emotions).

2.2 Preprocessing

Before starting our approach, we did some preprocessing to speech signals. These include silence removal and normalization. First of all, we removed silence parts from the speech signal. Silence part is unnecessary information for speech processing and it does not include any information about the speaker and spoken subject. The idea of the silence removal is that speech signal is divided into frames and then each frame

value is compared with the specified threshold value that can be defined by the end user. The threshold is selected value for removing silent parts of sound signals from samples. If the frame value is lower than the threshold value, this frame will not be included resulting output speech signal. In this research, we used default threshold settings to remove silent parts. The default threshold value is equal to 0.03. After this process, a new speech signal is constructed without silence parts. We focused on only spoken part of speech in order to decrease the computational time and unnecessary information. Also, seen in Figure 2.2, while Figure 2.2a represents the original speech samples, Figure 2.2b represents removed silence part from the speech signal.

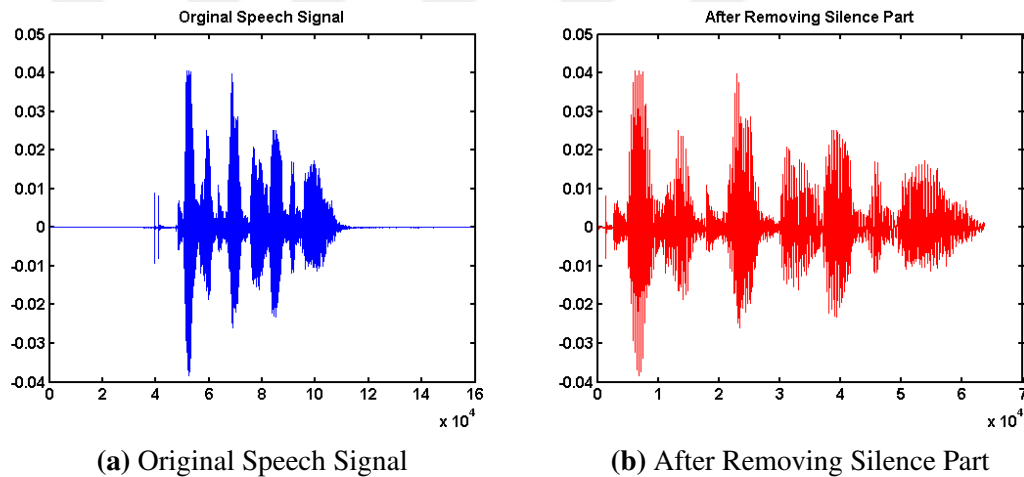


Figure 2.2 (a) A speech signal visualization: x-axis coordinate represents speech samples, and y-axis coordinate represents speech amplitude, (b) Represents speech signals after removing silent parts

The second preprocessing technique the most frequently used in speech emotion recognition is normalization of the speech signal. The method of z normalization modifies the volume of the speech to a standard level. Thus, it provides a comparable volume level to each speech signal. We normalized each speech signal with z-score technique in order to improve computational efficiency. We applied z-score to each speech sample so as to be the mean value equals to zero and the variance equals to one. Also, when you look at the figure, when the z-score is applied, each speech sample shows a distribution between -1 and 1. Thus, each signal sample has a standard scale. The figure 2.3 shows normalized speech signal.

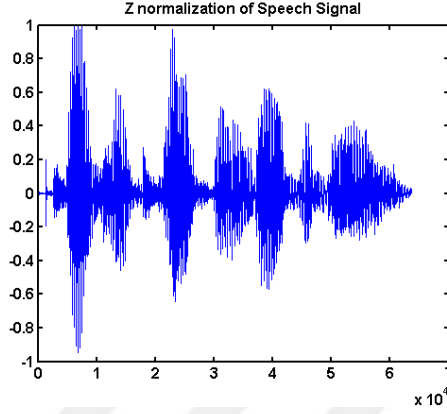


Figure 2.3 A speech signal is normalized with z-score normalization

2.3 Dimension Reduction

A speech signal of dimension t , $D = d_1, \dots, d_t$ can be reduced the dimension of the original speech signal to t/w into a vector $\bar{D} = \bar{d}_1, \dots, \bar{d}_i$. The i th element is calculated by the below equation:

$$\bar{d}_i = \frac{1}{w} \sum_{j=w \times (i-1)+1}^{w \times i} d_j \quad (2.1)$$

where w represents window size of arbitrary length $i = 1, \dots, t/w$. Briefly stated, the speech signal that has t number of data points is divided into non-overlapping windows (partitions) of equal length, w window size. The average of data falling into w^{th} window is calculated and each window is represented by the mean of its data points. Thus, the speech signal reduces to \bar{d}_i dimensions. This representation can be demonstrated as the part of speech with a combination of each window in Figure 2.5a. In our study, we determined window size w to 8.

2.4 Discretization

In the discretization technique, we transform a vector \bar{D} to a string character \bar{S} . The purpose of this is that we represent each unit of \bar{D} , \bar{d}_i with each unit of \bar{S} , \bar{s}_i . For this representation, we will utilize a Gaussian distribution table in order to obtain the coefficient of each \bar{d}_i . Distributions have equal size areas n under the Gaussian curve.

These areas are represented by breakpoints $B = B_1, \dots, B_{n-1}$. These breakpoints can be obtained by table in Figure 2.4. This figure shows the breakpoints for values of a from 3 to 10. For example, when a equals to 3, breakpoints have 2 equal size areas and they

$a \backslash \beta_i$	3	4	5	6	7	8	9	10
β_1	-0.43	-0.67	-0.84	-0.97	-1.07	-1.15	-1.22	-1.28
β_2	0.43	0	-0.25	-0.43	-0.57	-0.67	-0.76	-0.84
β_3		0.67	0.25	0	-0.18	-0.32	-0.43	-0.52
β_4			0.84	0.43	0.18	0	-0.14	-0.25
β_5				0.97	0.57	0.32	0.14	0
β_6					1.07	0.67	0.43	0.25
β_7						1.15	0.76	0.52
β_8							1.22	0.84
β_9								1.28

Figure 2.4 Breakpoints of Gaussian Distribution (Source: [1])

are tagged with a character from English alphabet. According to the table, a character is assigned depending on the breakpoints. If \bar{d}_i is less than -0.43, \bar{d}_i obtains a character to convert to \bar{s}_i . If \bar{d}_i is between -0.43 and 0.43, \bar{d}_i obtains b character to convert to \bar{s}_i and if \bar{d}_i is greater -0.43, \bar{d}_i obtains c character to convert to \bar{s}_i . In our research, we selected a as 5. Figure 2.5b shows an example of discretization. We transform \bar{D} to corresponding \bar{S} as a word representation of characteristic **aabcceeeebc**.

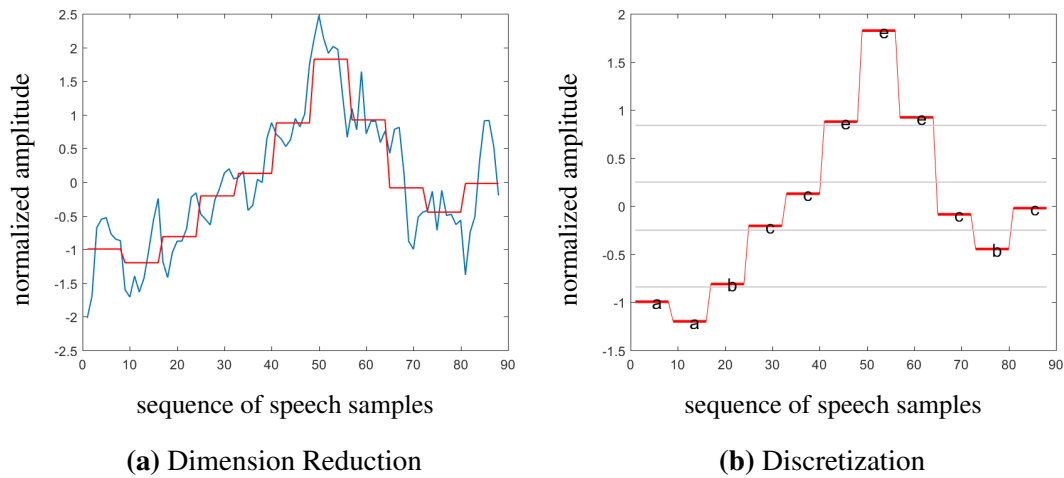


Figure 2.5 (a) t number of a speech signal is reduced t/w . In this example, a speech signal of length 88 is divided by 8 window size and it is reduced to 11 dimensions, (b) signal is discretized depending on Gaussian distribution table

2.5 Pattern Mining

Contrast sequential mining is a crucial subarea of data mining. The purpose of the contrast sequential mining is to discover diverse patterns that have differences in data sets. The particular type of contrast mining is distinguish sequential pattern mining (DSP). DSP mining discovers diverse sequential patterns and it extracts these patterns with the discriminative way from different classes of sequences. Particularly, DSP mining focuses on extracting patterns which are found frequently in sequences of one class but not frequently in sequences of another class. In other words, it is possible to compare sequential patterns which distinguish one sequence from other sequences with DSP mining.

DSP mining has been used in several application areas such as protein (DNA) [43], information retrieval [43], analysing purchase behaviors. For example, in biology, protein sequences may be related to protein family or genes. DSP mining contributes protein sequences to characterize and distinguish protein family in DNA. Web pages and books include a set of word sequences and differences can be extracted between sentence collection to detect distinctive phrases. They may be useful for indexing. DSP can be used for changes in customer behaviors in online shopping in order to purchase transaction. It can propose shopping preferences to costumes depending on the shopping habit [43].

Constraints have an important part in data mining. They provide scalability of entire process and define the quality of the results during pattern extraction. Data mining constraints may be time stamp, removing redundancy, length of pattern and gap constraints. The specific restriction in DSP mining is gap constraint. 'Gap' constraint set boundaries between two consecutive values in a sequence. In DSP mining, there are several algorithms that use gap constraint. These algorithms are GepDSP (Gene Expression Programming) [44], KDSP-Miner (top-k distinguishing sequential patterns with gap constraint) [45], ConsgapMiner (Contrast Sequences with Gap Miner) [46] and iDSP-Miner (item set DSP) [47]. Except for K-DSP algorithm, minimum support (minsup) and minimum confidence (minconf) thresholds are predetermined by user. Minsup is the percentage of transactions in the sequences of one class. Also, minconf is the conditional probability of these transactions. These two parameters are utilized

to exclude the rules in the result that have support or confidence less than minimum support and minimum confidence. The support of sequential patterns with thresholds never exceeds the support of its subset. When the user does not determine the threshold correctly, the problem may occur due to missing important contrast patterns.

In our research, we used KDSP-Miner algorithm and extracted top-1000 patterns with gap constraint using both one-versus-all and one-versus-one strategies. Data is divided into two classes in this algorithm. These classes are positive and negative respectively. In one-versus-all strategy, while all discrete representation of an emotion is in a positive class, all discrete representation of other emotions is in the negative class. The algorithm extracts m patterns for each emotion. In one-versus-one strategy, while all discrete representations of one emotion are in the positive class, all discrete representations of another emotion must be in the negative class. This technique is applied in the same way for all other emotions respectively. The number of patterns is $(j - 1) * n$ for each emotion where j is the number of emotions and n is the number of patterns that are obtained from each pair emotion. Gap constraint is set to zero for each emotion. As each pattern can become visible in both positive and negative classes simultaneously, the algorithm gives two measurements that a pattern appears in each class as positive support (PosSup) and negative support (NegSup). When PosSup is divided to NegSup, a measure is obtained called the C-ratio. We have created set of patterns from each emotions and these patterns are selected by the highest C-Ratio.

2.6 Pattern Features

Patterns that obtained the highest C-Ratio are used for acoustic descriptors. Assume that there are i number of pattern in j^{th} emotion. Patterns are described by p_i^j where $i=1, \dots, n$ and $j=1, \dots, m$. The feature vector f_t for t^{th} size is $n \times m$. When a pattern specific to emotion p_i^j is given to a discretized speech signal, we count the frequency of patterns that is exactly matched on discretized signal in order to compute the feature vector f_t . While applying exact matching, we don't permit any gaps between matches of the pattern. The purpose of this is to preserve the motif that is represented by the pattern. For example, **dcaecbaeddc** is a discrete representation of a speech signal. This representation has 2 different patterns which are **dc** and **aec**. Also, patterns are

detected in two different parts of representation. The first pattern is seen twice in the discrete speech signal which makes its corresponding in the feature vector 2. However, the second pattern is matched once exactly and twice with a gap constraint of two. **dd** represents gap in the sub string of **aeddc**. For our approach, we focus on exact match. Therefore, the value of feature vector is set to 1 for aec pattern.

2.7 Acoustic Features

Speech emotion recognition is essentially conducted without linguistic information via sound processing. In a sense of acoustics, speech processing techniques propose extremely important information based generally on prosodic and spectral features [48]. The feature vector of acoustic is calculated for each speech signal. Most of the features are extracted from speech-based information, i.e. pitch, energy and duration, as well as spectral parameters such as formants and MFCC. These features are used for recognizing emotions from speech. Depending on the temporal structure, the feature vector of acoustic is grouped into two categories that are namely short (segmental) and long (suprasegmental). Firstly, segmental features are computed once for each small time frame that is the duration of between 25 and 50 milliseconds using window technique in order to analyze their temporal progression. On the contrary, the suprasegmental features are computed over the whole duration of the speech. In a linguistic context, the suprasegmental feature is described as an attribute containing discrete phonetic and linguistic unit boundaries that can be identified in order to extract the analysis of the speaker's behavioral feature [49].

The feature vector of acoustic is classified into two unique classes. These classes are Low-Level Descriptors (LLDs) and functionals. The first class is LLDs that consist of prosodic, spectral and voice quality features. In terms of prosody, most of human use the prosodic signal to identify their emotions in daily conversation. Prosody gives information about the person's speaking. In the literature, prosody is known as prosodic features ([50],[48]). Prosodic features are the part of speech that goes outside of phonemes (words, phrase, and sentence) and address sound's auditory qualities. These features are fundamental frequency (F0) and energy. Prosodic features are

segmental features and extracted from the time domain. Spectral features are used successfully in most of the studies to improve the accuracy of emotion recognition. It is well known that spectral features are extracted in frequency domain analysis [51]. The features depend on short time power spectrum. These features can be seen from the spectrum such as formants, bandwidth, and spectral energy and they also are segmental features. The most popular spectral features are MFCC, Linear Prediction Cepstral Coefficient (LPCC) and Log Filter PowerCoefficient (LFPC). Contrary to spectral and prosodic features, there is no acoustic property in the quality of voice. For this reason, it is not measured or analyzed from a speech signal. It consists of many parts of speech processing. Generally, qualitative terms characterize voice quality such as harsh, tense and breathy. Voice quality is suprasegmental features. The most popular features in voicing quality are namely jitter, shimmer and Harmonics-to-Noise Ratio [50].

The second class of feature vector is functionals that include statistic features. Statistic features are obtained from LLDs so functionals are suprasegmental. The statistical features are such as mean, median, standard deviation, kurtosis, maximum, minimum and etc.

If we have a look at the description of speech features and temporal structure are:

- MFCC describes major phonetic attributes in speech. It also is strongly related to the spoken content. The basic idea for extracting MFCC feature is that time domain speech signal transforms to a frequency domain based on critical bands using Fourier transform. The critical band is a bandpass filter for center frequency adjustment. The resulting power spectrum is filtered using a filter bank with critical bands. The filter bank is chosen with a logarithmic mel scale [52].
- The sound of speech begins in chords of voice. The vocal chords vibrate with a fundamental frequency that is high and low in speech sound. This vibration of the cords is called fundamental frequency (F0). Fundamental frequency can be calculated from the speech signal and it gives information about speaker and type of sound. Every person has a different fundamental frequency. For this reason, it is a delicate feature in the meaning of listening. In the literature, there are 3 types of algorithms to extract this feature and detect the emotions. The algorithms are applied to 3 different domains which are frequency, time and frequency-time domain [53].

- Both jitter and shimmer features are a segment of voiced speech. Period-to-period fluctuations of the fundamental frequency are available in speech. Jitter is a metric for these fluctuations in fundamental frequency. Shimmer is a metric of amplitude value for period to period variability. Most of the studies use the glottal flow function for extracting both voice quality features. The glottal flow function can be computed with inverse filtering algorithm. An inverse filtering algorithm predicts the vocal tract filter and performs the inverse of filter estimation to speech signal to present a glottal flow estimation [54].
- Energy is the most popular feature in speech emotion recognition. It plays an important role to detect emotion from the speech signal. Energy is a measure of how much signal occurs at a specific time. The energy of a signal is computed by shifting short time window technique, squaring the samples and taking this mean. The usage of square root is called as root-mean-square (RMS). The RMS is a common approach for estimating signal's energy [53].
- The spectral centroid is a method of measuring in order to define a spectrum used in speech processing. It specifies the location of the center of gravity for spectrum. It represents the brightness and sharpness of a sound [55].
- The energy of speech signal that contains emotion information is observed to be within a specific range of frequency. Spectral roll-off specifies the content of frequency below which specific percentage (cutoff) of the total amount of energy remains. It is possible to use roll-off frequency in order to differentiate between harmonic and noisy sounds [50].
- The spectral flux refers to the change of an emotional signal in the local spectrum. It shows how rapidly the planned signal power spectrum changes within frames [50].

In our research, we used 65 provided LLDs features. These features are defined in Interspeech 2013 Computational Paralinguistics Evaluation (ComParE) [56] feature set. The feature set contains 41 spectral, 14 MFCC, 4 energy and voicing related LLDs features. We used a tool for extracting LLDs features. ComParE feature set includes 6373 statistic features in OpenSMILE [57]. OpenSMILE is feature extraction tool for audio signals. You can find more information in appendix A.1. We only applied

12 statistics to speech signals. The statistics are mean, standard deviation, maximum, minimum, range, interquartile ranges (1-2,2-3,1-3), the position of mean, kurtosis and skewness. We used both part of LLDs and delta LLDs to acoustic features. The total number of acoustic feature is 1313 LLDs features that are used in our research.

2.8 Classification Schemas

We used two different algorithms for patterns. In these approaches, the maximum voting algorithm and classification is performed for features. In this study, the maximum voting algorithm is presented without learning a task and then DAGSVM (directed acyclic graph support vector machine) is used for classification. In the next subsection, we will describe the basic idea of the maximum voting algorithm and we will explain the operation logic of DAGSVM.

2.8.1 Maximum Voting Algorithm

Before applying a classification technique for patterns, we utilize a simple voting algorithm independent of learning in order to detect the effect of patterns on emotions. First of all, a speech signal is transformed into the \bar{S} discretization representation described in section 2.6. We take the pattern p_i^j that belongs to a specific emotion j (p_i^j is i^{th} pattern for j^{th} emotion) and count the frequency of this pattern that is exactly matched on \bar{S} . (A pattern can be seen more than one on \bar{S} or not). The matching process is iterated for each pattern i of an emotion. At the end of the process, the matching frequency of each pattern is summed up in order to obtain the total number of matches for emotion j . This process is repeated the same for all emotions j . At last, the emotion that has highest the total result is determined as speech's emotional label \bar{S} . In this algorithm, there may be a situation that the total results are the same for a different emotion. In this situation, we select random the label among conflicting emotions and then we assign one of the conflicting emotions as the label of the speech.

2.8.2 SVM

We used SVM (Support Vector Machine) for pattern classification in our research. SVM is normally designed for binary classification that is introduced by Vapnik [58]. There are two types of classifications that are used in SVM. These are namely single and multi-class classification. We focused on the multi-class classification of SVM in order to detect different emotion recognition from speech. In the literature, multi-class classification is examined in five types of techniques that are one versus one, one versus all, directed acyclic graph, tree-based and error correcting output codes [59].

In order to compare with the results in [42], we adapted DAG to SVM with multi-class emotion classification. In the training phase, DAGSVM is similarly associated with the one-versus-one classification. For this reason, we will introduce the one-versus-one technique in the next section and then we will explain how DAGSVM works.

2.8.2.1 One-Versus-One

One of the implementation techniques of multi-class classification in SVM is one-versus-one method. This method builds $k(k-1)/2$ classifiers where k is number of classes. Each classifier is trained by two classes which exist in data set. The first class is tagged as positive examples and second class is tagged as negative examples. In training data for i th and j th classes, [60] solves the classification problem with :

$$\begin{aligned}
 \min_{w,b,\xi} \quad & \frac{1}{2}(w^{ij})^T w^{ij} + C \sum_t \xi_i (w^{ij})^T \\
 \text{s.t.} \quad & (w^{ij})^T \phi(x_t) + b^{ij} \geq 1 - \xi_i, \quad \text{if } y_t = i \\
 & (w^{ij})^T \phi(x_t) + b^{ij} \leq -1 + \xi_i \quad \text{if } y_t = j \\
 & \xi_t^{ij} \geq 0
 \end{aligned} \tag{2.2}$$

When solving this problem, suppose that we have training samples $(X_1, Y_1), \dots, (X_N, Y_N)$ where X_N are training samples and Y_N are class labels. The training samples X_N are mapped by a function Φ to higher dimensional feature space and C is the training error parameter to use for providing the balance margin. W is a coefficient vector for hyperplane in feature space. ξ represents the number of pattern variables and b is a

constant variable. The Max Wins algorithm is adapted to associate with classifiers. The algorithm works with a voting strategy. The voting strategy is that if the decision boundary indicates the sample of X in i^{th} class, a score of one is added to i^{th} class otherwise the j^{th} class is enhanced by one.

SVM is a learning algorithm that is based on the kernel function. Kernel function defines the nonlinearity level for one-versus-one SVM characteristics. The kernel is used to prevent obvious calculation of interval product in high dimensional space. The function can be define with $K(X_i, Y_j) = \Phi(X_i) \cdot \Phi(Y_j)$. It is possible to express this function with the dual form of SVM classification problem which changes the kernel function with Mercer's theorem. $K(X_i, Y_j)$ is the specific kernel function. There are four types of kernel function available to use in SVM. These are linear, Gaussian, polynomial and sigmoid that are discussed in [61]. In our research, we used Gaussian kernel (radial basis function, RBF) for training a model.

2.8.2.2 DAGSVM

Directed acyclic graph (DAG) is a graph whose edges are balanced and has no cycles. DAG includes root and leaf nodes. A root node of DAG is an only unique node which does not have arcs into it. Also, it has nodes that have separation to either 0 or 2 arcs to use class of function in the classification task. DAGSVM is associated with one versus one (1vs1) SVM in the training phase. The 1vs1 method is that it constructs a binary classifier for only two classes. For multi-class problem, the number of m classes are trained by $\frac{m(m-1)}{2}$ classifiers. In the testing phase, the input variable starts from the root node and moves to left or right node depending on the output classifier. If the classifier is zero, the node is the left edge or if the classifier is one, the node is right edge. This process goes on until the leaf node which indicates the decision node of the predicted class. Also, $m-1$ decision nodes are trained for m classes. Although DAGSVM has the same classification performance with SVM, the advantage of DAGSVM is that testing time is less than 1vs1 SVM [16]. In our research, we used the LibSVM library for the implementation of DAGSVM. Firstly, we adapted SVM testing implementation with the DAG approach. LibSVM is a free library for support vector machine. Detailed information is available in the appendix [A.2](#).

Chapter 3

Results and Analysis

In this section, the results of the proposed method were assessed and compared with [42] results. We created separate sets with 5, 10, 15, 50, 100, 150, 250, 500 and 1000 patterns for each emotion according to the C-ratio criteria in order to analyze the influence of pattern count on both the classification and maximum voting algorithm. The size of the feature vector was determined by selecting the pattern count as seen Section 2.6. For example, the feature vector size of 10 patterns for each emotion is 60. Thus, a feature vector enables a speech signal to be represented as independently of the type of emotion. In terms of acoustic, we extracted MFCC, voice quality, spectral, energy, and RASTA features, as described in Section 2.7. The feature vector for acoustic features was selected 12 different statistics, which were mean, standard deviation, maximum, minimum, range, inter-quartile ranges (1-2, 1-3, 2-3), skewness, and kurtosis. The size of the feature vector was different from each other, and the total number of the feature was 1312 (80 energy, 280 MFCC, 520 RASTA, 300 spectral and 132 voicing quality). Before using in classification schemes, the feature vectors were normalized by z-score normalization. In the normalization technique, we applied performer dependent z normalization to features. Each feature of each performer had mean value 0 and standard deviation 1.

In all experiments that we modeled a classifier, leave-one-performer-and-sentence-out cross validation technique was used. One sentence from one performer was analyzed in each round of testing data. The training data was consisting of other performance and other sentences. Model parameters for optimization were carried out through 5-fold

internal cross-validation on the training data folds, and the parameters were selected by grid search. The grid search defines the optimal hyper-parameters of SVM such as C and λ where C is misclassification error rate for training and λ is kernel function to check the bandwidth of the radial basis function (RBF). In our experiments, we used the popular RBF kernel and grid search in order to model the training data. ($C = 2^{-5}, 2^{-3}, \dots, 2^{-15}$ and $\lambda = 2^{-15}, 2^{-13}, \dots, 2^3$).

In the pattern extraction technique, we applied four different experimental approaches to our data. A quick reminder; we have two predefined statements which are (S1): kids are talking by the door, (S2): dogs are sitting by the door. As we described earlier, we discretized each statement of all emotions, and we converted each speech to the word representation. Then, we extracted patterns from the discretized speech signals by using two scenarios. The first scenario is that S1 and S2 are together; the second scenario is that S1 and S2 are separated. When S1 and S2 are separated, patterns of each statement represents itself. To say that, S1 patterns represent S1 statement, and S2 patterns represent S2 statement. For each statement, we extract an equal number of 1000 pattern. When two statements are together, extracted patterns represent the combination of S1 and S2 statements. We extract 1000 pattern for this scenario too. Also, we have two different strategies in mining patterns; one-versus-one and one-versus-all described earlier in Section 2.5. Briefly, all experimental results were produced by these pattern extraction scenarios and strategies.

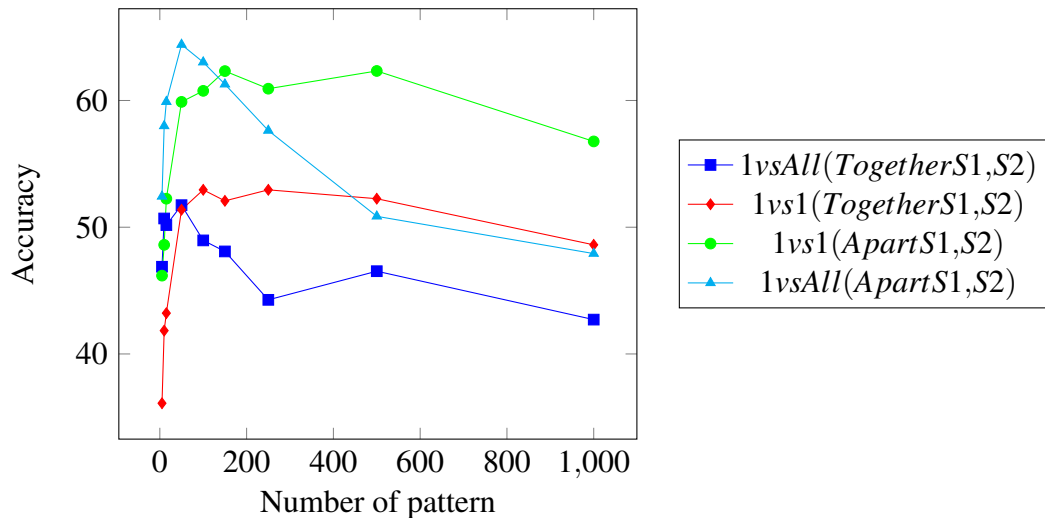


Figure 3.1 Maximum voting results

Figure 3.1 shows voting-based results for extracted patterns as accuracy. From this figure, we can see that there are four different approaches as we mentioned above. Voting-based results provide comparable results between pattern approaches. It's obvious that accuracies decrease when the number of pattern count increases all approaches. When we sort patterns according to the C-ratio and choose the pattern set from the list, the patterns in the top of the list will have larger discriminative influence than next following patterns. When the number of patterns increases, discriminative influence decreases. As a result, an increase in patterns will result in a decrease in accuracy. When statements are separated, the result will be more accurate than two statements together. For this reason, the patterns are more descriptive characteristics in discretized speech signals when they are not together. In terms of strategies, the one-versus-one method gives better results in accuracy, because emotion patterns are discriminated from one to another easily.

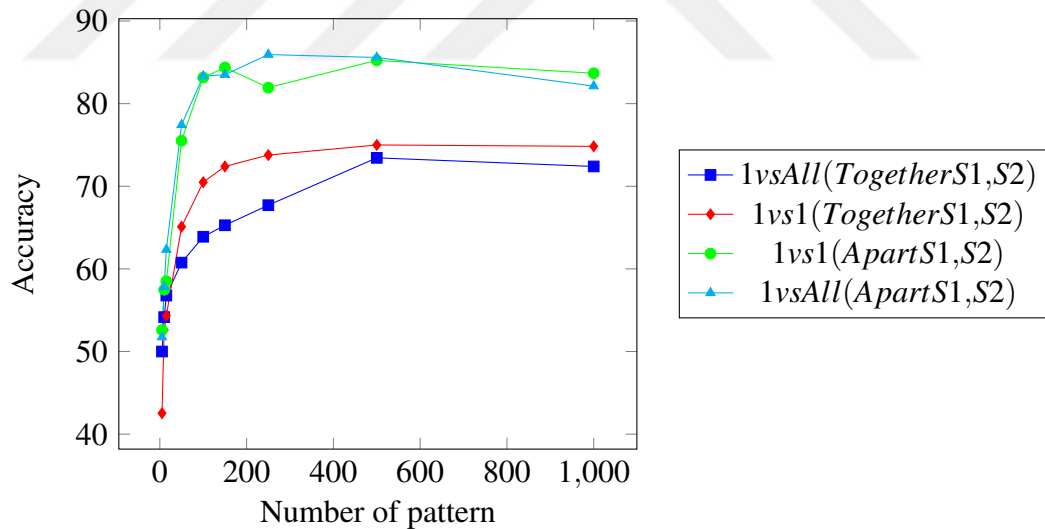


Figure 3.2 DAGSVM classification results

Figure 3.2 shows DAGSVM classification results for all pattern extraction approaches as accuracies. In nearly all cases, classification results surpass from the voting based results. Also, classification results enhance the voting results by up to 34 percent for larger patterns. When the number of patterns increases, classification results increase, unlike voting based results. The feature vector of each pattern contains more information than the pattern used, so it affects the rate of classification. To obtain the classification results, we used single classifier when two statements are together,

and we used two classifiers when statements are separated. Only in the two classifiers, each statement is classified separately, then the average of classification results is taken. In Figure 3.2, it is seen that S1 and S2 patterns (apart S1-S2) achieve an increase of 85 percent for recognizing the emotion compared to patterns of all statement (together S1+S2). As each statement is supported by their patterns, the classification accuracy increases. That's why patterns of the statements are declared separately, and this situation increases the classification accuracy and provides a better perception of the statement.

Table 3.1 The Computational Duration

	DAGSVM		
	Voting-Based	Training Time	Test Time
5 Pattern	1.5 min	30 min	12 min
10 Pattern	3 min	46 min	13 min
15 Pattern	4.5 min	53 min	15 min
50 Pattern	14 min	3 hrs	24 min
100 Pattern	30 min	4.30 hrs	40 min
150 Pattern	47 min	6.30 hrs	58 min
250 Pattern	1.30 hrs	11 hrs	1.41 hrs
500 Pattern	3.15 hrs	20 hrs	4 hrs
1000 Pattern	6.30 hrs	32 hrs	7 hrs

The computational duration of each pattern for voting based and DAGSVM classification is shown in the Table 3.1. Testing time includes the following steps: preprocessing, normalization, discretization, feature extraction, cross-validation and testing for classification. As seen, when the number of patterns increases, training and voting time increase. In terms of classification, when the number of patterns in the experimental set increases, the model of training takes too much time. This also affects the duration of the classification, and the results can be obtained after a long period of time. Therefore, we selected 100 pattern for classification because we observed that increasing the number of pattern beyond 100 pattern will not affect the accuracy noticeably. The computational time is relatively low when 100 pattern used. Therefore, we selected 100 pattern in the rest of the experiments.

In the following paragraphs, we'll explain our strategies and experimental results by use of DAGSVM classification in detail.

Table 3.2 Experimental set 1

	One Versus All
	S1, S2 Together
100 Pattern	63,89

Table 3.2 shows our first experimental result. This table presents the classification results of the extracted 100 pattern from S1 and S2 statements which are together. We applied one-versus-all strategy in this experiment and used a single classifier for 100 pattern. The resulting classification accuracy is 63,89.

Table 3.3 Experimental set 2

	One Versus One
	S1, S2 Together
100 Pattern	70,49

As seen in Table 3.3, we have another experimental set which has the same scenario, but a different strategy than the previous one. In this experiment, we used one-versus-one strategy for pattern extraction, and again, we used a single classifier. In this experiment, the classification accuracy is equal to 70,49.

Both in Table 3.2 and 3.3, we see the classification results of different strategies for the same scenarios. When we compare the results, one-versus-one strategy has a higher classification accuracy than one-versus-all strategy even if they have the same scenario. One-versus-one strategy gives better results in accuracy because emotion patterns are more discriminative from one to another. In one-versus-all strategy, emotion patterns are less discriminative one to all others. The pattern distribution between all emotions might not be equal, and this may affect the classification. Hence, one-versus-one provided us better results in classification accuracy.

Table 3.4 Experimental set 3

	One Versus All		
	S1	S2	Average
100 Pattern	81,60	85,07	83,33

Table 3.4 shows our third experimental set results. In this experiment, we applied one-versus-all strategy for pattern extraction, and we used two classifiers for each statement S1 and S2. This is the case when S1 and S2 statements are separate. After that classification, we calculated the average classification accuracy, and this is also illustrated in this table. For this experiment, results are 83.68 for S1 statement, 82.64 for S2 and the average 83.33 respectively.

Table 3.5 Experimental set 4

	One Versus One		
	S1	S2	Average
100 Pattern	83,68	82,64	83,16

Table 3.5 shows another experimental set having the same scenario but distinct strategy according to Table 3.4. In this experiment, we applied one-versus-one strategy and two classifiers for S1 and S2. Also, we calculated average classification accuracy for this scenario too. These are all depicted in this table. Accuracy values are 83.68 for S1, 82.64 for S2 and the average value 83.16 respectively. When we compare Table 3.4 and 3.5, we observe that there is a slight change between those strategies.

In the following part, we'll explain new experiments which are elaborated according to the previous ones.

Table 3.6 Experimental set 5

	One Versus All		
	S1, S2 Together (2 classifiers)		
	S1	S2	Average
100 Pattern	57,29	60,79	59,03

Table 3.6 shows the classification results of the extracted 100 pattern from S1 and S2 statements, which are together. However, S1 and S2 are classified separately in this experiment. This table is the specialized version of Table 3.2. In Table 3.2, we used one classifier, but in here we used two classifiers. Aim of this experiment is to observe the effects of patterns on each statement S1 and S2. As you can see in Table 3.6, patterns are extracted balanced for each S1 and S2 statements. When we compare Table 3.6 with Table 3.2, we observe that there is a small decrease in accuracy by 5%. The main reason for the difference is that when classifying each statement (S1, S2), we use patterns that are extracted from S1 and S2 together. Hence, supporting patterns for S1 may decrease the classification accuracy of S2 and vice versa.

Table 3.7 Experimental set 6

	One Versus One		
	S1, S2 Together (2 classifiers)		
	S1	S2	Average
	100 Pattern	62,85	61,81

Table 3.7 shows a specialized version of the experiment in Table 3.3. They share the same scenario and strategy, but as we did in the previous experiment, each statement (S1 and S2) is classified separately. The results of classification are 62,85 for S1, 61,81 for S2 and 62,32 for average respectively. The reason for the result difference is also familiar with the previous experiment, which is supporting patterns for S1 may decrease the classification accuracy of S2 and vice versa.

Table 3.8 Experimental set 7

	One Versus All
	S1, S2 Together (Single classifier)
	100 Pattern

Another experimental set is depicted in Table 3.8. In this experiment, we applied one-versus-all strategy for pattern extraction, and we used a single classifier for statements S1 and S2 together. The classification accuracy is equal to 72,05. This experiment is a specialized version of Table 3.4. Our aim is to show and evaluate that the patterns,

which are extracted from each statement (S1, S2), are whether discriminative or not. When we compare the classification results with Table 3.4, there is a significant reduce in accuracy by 11%. Patterns are extracted from each statement (S1 and S2). Patterns of a statement (S1) may intersect with the patterns of another statement (S2). In this experiment, we combine those patterns and perform the classification task. Therefore, a combination of these two may lead to a decrease in classification accuracy. This is the exact reason for the decline in accuracy.

Table 3.9 Experimental set 8

	One Versus One
	S1, S2 Together (Single Classifier)
100 Pattern	71,7

In Table 3.9, we performed another experimental set which is a specialized version of Table 3.5. We applied one-versus-one strategy for pattern extraction, and we used a single classifier when S1 and S2 patterns separate. As we did in the previous experiment, we combined those patterns and performed the classification task. The resulting value of the classification is 71,7, which is quite less than in Table 3.5. The difference is about 12%. The reason also is the same as the previous experiment.

After these experiments, we compare the classification results of acoustic and pattern-based features of our approach in the following paragraphs.

Table 3.10 The classification results of acoustic features

	12 Statistics	# of Features	12 Statistics	All Statistics	# of Feature	All Statistics
Energy	34,78 [42]	80		74,13	400	
Voicing	38,41 [42]	132		55,73	473	
MFCC	48,73 [42]	280		68,23	1400	
Spectral	48,01 [42]	300		76,39	1500	
RASTA	30,43 [42]	520		63,19	2600	

Table 3.10 shows the classification results of the reference paper, and each acoustic feature having all statistics. In this table, the reference paper uses 12 statistics to each acoustic feature. We described these statistics in section 2.7 before. These statistics are

mean, max, min, standard deviation, skewness, range, kurtosis, inter-quartile ranges, mean, and the position of the mean. Also, we applied all statistics to acoustic features without selecting any special statistics. These statistics also are available in ComParE feature set [56].

Table 3.11 The classification results of pattern-based features

	Classification Accuracy (%)	# of Features
1vsAll-100 Pattern (S1,S2 Together)Proposed	63,89	600
1vs1-100 Pattern (S1,S2 Together)Proposed	70,49	600
1vsAll-100 Pattern (S1,S2 Apart)Proposed	83,33	600
1vs1-100 Pattern (S1,S2 Apart) Proposed	83,16	600

Table 3.11 shows the classification results of our pattern-based features. For comparison of Table 3.11 and 3.10, we used two proposed pattern-based classification results of 1vsAll-100 Pattern (S1, S2 Together) and 1vs1-100 Pattern (S1, S2 Together). The purpose of using only these two proposed results is that the two statements (S1 and S2) are classified together in acoustic features. We do not compare the results of 1vsAll-100 Pattern (S1, S2 Apart) and 1vs1-100 Pattern (S1, S2 Apart), because they are classified separately and comparing them is not a fair approach. Therefore, we only used 1vsAll-100 Pattern (S1, S2 Together) and 1vs1-100 Pattern (S1, S2 Together) classification results for comparing the two tables. When we compare our proposed approaches with reference paper results in Table 3.10, we can say that our approaches produced the best results in all acoustic features. The size of the feature vector in the pattern-based approach is a little more than the number of RASTA features. Our approaches give better results even in the worst strategy 1vsAll-100 Pattern (S1, S2 Together) as compared to the case of using RASTA features. When we compare our proposed approaches with acoustic features having all statistics, we can say that although the number of the feature of pattern-based approaches is less than the number of the feature of MFCC and RASTA, our approaches give better results in both strategies. However, our pattern-based approaches have low accuracy outcomes according to energy and spectral features.

Table 3.12, and Table 3.13 show the classification results of combining acoustic and our pattern-based features. In Table 3.12, we applied 12 statistics to acoustic features. In Table 3.13, we applied all statistics to acoustic features in ComPaRe feature set. In

each table, when pattern-based features are combined with acoustics, they contributed to increase the classification accuracy. The classification accuracies of pattern-based features for one-versus-all and one-versus-one strategies when S1-S2 apart are 83.33 and 83.16 respectively. When RASTA and MFCC features are combined with the pattern-based features, the classification accuracy decreases. This is the negative effect of RASTA and MFCC on pattern-based features. Each statement is not accurately predicted by RASTA and MFCC. When RASTA and MFCC features are combined with the pattern-based features, they generate redundant information for the machine and this makes training difficult. For this reason, there is a reduction in the classification accuracy caused by two features. However, better results are still obtained than the existing techniques. In both tables, when pattern-based features are combined with all acoustic features, the classification accuracy is over eighty percent. Also, the accuracy result of each strategy is close to one another.

Table 3.12 The classification results of combining acoustic and pattern based features (12 statistics were applied to acoustic features)

	One Versus All	One Versus One	One Versus All	One Versus One
	S1,S2 Together	S1,S2 Together	S1,S2 Apart	S1-S2 Apart
Energy +100 Pattern	80,73	82,12	92,53	89,76
Voicing +100 Pattern	73,61	75,52	85,94	83,51
MFCC+ 100 Pattern	70,83	76,22	82,29	82,99
Spectral +100 Pattern	80,90	84,38	91,15	88,89
RASTA+100 Pattern	67,88	73,78	82,64	77,43
AllAcoustic+100 Pattern	87,85	89,23	90,10	88,89

Table 3.13 The classification results of combining acoustic and pattern based features (All statistics were applied to acoustic features)

	One Versus All	One Versus One	One Versus All	One Versus One
	S1,S2 Together	S1,S2 Together	S1,S2 Apart	S1-S2 Apart
Energy +100 Pattern	85,42	86,11	92,19	90,97
Voicing +100 Pattern	73,78	76,73	82,12	83,33
MFCC+ 100 Pattern	75	78,81	79,86	77,95
Spectral +100 Pattern	84,55	85,76	86,98	85,24
RASTA+100 Pattern	74,83	75,69	75,70	74,31
AllAcoustic+100 Pattern	90,27	89,06	87,15	86,45

In the following paragraphs, we'll explain experimental results with different frame and alphabet size using our best-proposed approach.

Table 3.14 Classification accuracy for different alphabet size (%)

	One Versus All (100 Pattern)		
	S1	S2	Average
8 Frame-3Alphabet	76,04	74,65	75,35
8 Frame-4Alphabet	78,82	75,35	77,08
8 Frame-5Alphabet	81,6	85,07	83,33
8 Frame-6Alphabet	75	81,25	78,13
8 Frame-7Alphabet	74,99	80,21	77,6

Table 3.14 shows classification accuracy for different alphabet size when the frame size is the same. In this table, we applied one-versus-all strategy for pattern extraction and we used two classifiers for each statement S1 and S2. After the classification, we calculated the average classification accuracy. When the different alphabet size is applied, classification accuracies are close to each other. Also, we can say that 5 alphabet has more optimal results compared to other alphabet sizes.

Table 3.15 Classification accuracy for different frame size (%)

	One Versus All (100 Pattern)		
	S1	S2	Average
6 Frame-5Alphabet	69,09	74,65	71,87
8 Frame-5Alphabet	81,6	85,07	83,33
16 Frame-5Alphabet	60,07	57,29	58,68
32 Frame-5Alphabet	55,9	54,51	55,21
64 Frame-5Alphabet	48,96	47,22	48,09

Table 3.15 shows classification accuracy for different frame size when the alphabet size is the same. This table presents the classification results of the extracted 100 pattern from S1 and S2 statements which are apart from each other. Each statement is classified separately and the average classification accuracy is calculated. In this table, when the frame size increases, classification accuracies decrease except 6 frames. 6 frames is an

exceptional case for this situation. For this reason, we examined the 6 frames for the different pattern set in Table 3.16. This table indicates that when the pattern number increases, classification accuracies are not over eighty percent. Generally, we can say that in table 3.15, 8 frame size is more optimal compared to other frames.

Table 3.16 Classification accuracy for 6 Frame-5 Alphabet (%)

	One Versus All		
	S1	S2	Average
5 Pattern	44,44	51,04	47,74
10 Pattern	52,43	51,73	52,08
15 Pattern	55,21	52,77	53,99
50 Pattern	66,66	69,79	68,23
100 Pattern	68,50	74,29	71,40
150 Pattern	74,30	77,43	75,86
250 Pattern	75,35	78,47	76,91
500 Pattern	74,65	75	74,83
1000 Pattern	74,99	74,65	74,83

In the following paragraphs, statistical analysis for pattern-based and acoustic features will be explained.

The McNemar is a statistical test that is used for testing paired nominal data. This test uses non-parametric variables for performing the task. In the McNemar test, we suppose that there are two systems available on the same data. For example, two systems are A and B. These systems are trained and estimated on the corresponding test set. Null hypothesis indicates that A and B systems include the same error rate, Alternative hypothesis indicates that these systems have a different error rate. McNemar statistical significance score can be calculated by the following equation:

$$McNemarValue = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}} \quad (3.1)$$

In this equation, n_{01} represents to the number of examples misclassified by system A but not system B, and n_{10} represents to the number of examples misclassified by system B but not system A. We have the critical value in this test that represents the

significance level. The critical value is 3.8415 for 5% significance level. If the value is larger than the critical value, the null hypothesis is rejected and the alternative hypothesis is accepted. In our project, system A refers to the one-versus-all strategy in pattern extraction and system B refers to the one-versus-one strategy in pattern extraction. Our purpose is to examine whether the predicted values after the classification are different in two strategies or not. In our project, we used this test for pattern-based and acoustic features to analyze the one-versus-all and one-versus-one strategies. Table 3.17, Table 3.18 and Table 3.19 show our McNemar test results. These tables show the results of classification accuracy for two different strategies and different scenarios. The strategies are one-versus-one (1vs1) and one-versus-all (1vsAll). The scenarios are S1, S2 together and S1, S2 Apart respectively. We applied the McNemar test for the same pattern number using different pattern extraction strategies such as 1vs1 and 1vsAll. We used the equation 3.1 to apply classification results, and we obtained McNemar value from each pair of two strategies. You may see star signs(*) in the following tables (Table 3.17, Table 3.18, Table 3.19). This star sign refers to the situation when there is a difference of more than 5% between test results. This means that one-versus-one and one-versus-all strategies have a different error rate.

Table 3.17 McNemar Test Results One Versus All & One Versus One Pattern-Pased Features

	S1,S2 Apart					
	S1, S2 Together		S1		S2	
	1vsAll	1vs1	1vsAll	1vs1	1vsAll	1vsOne
5 Pattern	50*	42,53	55,90*	52,08	47,57	53,13
10 Pattern	54,16	52,60	62,85	59,03	52,78	55,90
15 Pattern	56,77	54,34	69,44	59,38	55,21	57,64
50 Pattern	60,76	65,10	78,82	75,69	76,04	75,35
100 Pattern	63,85	70,48*	81,60	83,68	85,07	82,64
150 Pattern	65,28	72,39*	82,64	85,42	84,38	83,33
250 Pattern	67,71	73,78*	82,99	84,03	88,8*	79,86
500 Pattern	73,44	75	85,76	85,42	85,42	85,07
1000 Pattern	72,40	74,83	81,60	82,64	82,64	84,72

In Table 3.17, Table 3.18, Table 3.19, we applied McNemar test with the same statements for two strategies (one versus one and one versus all) when S1 and S2 are apart.

Table 3.18 McNemar Test Results One Versus All & One Versus One Pattern-Pased and Acoustic Features (12 Statistics)

	S1,S2 Apart					
	S1,S2 Together		S1		S2	
	1vsAll	1vs1	1vsAll	1vs1	1vsAll	1vs1
Energy+100 Pattern	80,73	82,12	93,05	90,97	92,01	88,54
Voicing+100 Pattern	73,61	75,52	86,81	83,68	85,06	83,33
MFCC +100 Pattern	70,83	76,21*	77,78	82,64	86,80*	83,33
Spectral+100 Pattern	80,90	84,37*	89,93	91,32	92,33*	86,45
RASTA+100 Pattern	67,88	73,88*	84,03	80,90	81,25*	73,95

Table 3.19 McNemar Test Results One Versus All & One Versus One Pattern-Pased and Acoustic Features (All Statistics)

	S1,S2 Apart					
	S1,S2 Together		S1		S2	
	1vsAll	1vs1	1vsAll	1vs1	1vsAll	1vs1
Energy+100 Pattern	85.42	86.11	90.97	90.97	93.40	90.97
Voicing+100 Pattern	73.78	76.73	83.68	85.06	80.55	81.59
MFCC +100 Pattern	75	78.81	76.73	77.77	82.98*	78.12
Spectral+100 Pattern	84.55	85.76	85.06	84.38	88.88	86.11
RASTA+100 Pattern	74.83	75.69	77.08	75.69	74.31	72.91

Amongst all experimental sets, we selected two experimental sets resulting in higher classification accuracies. One of them has the highest accuracy in the pattern-based approach and the second one achieves the highest accuracy in combining acoustic and pattern-based features. We calculated the performance metric for each set. Table 3.20 shows the confusion matrix of experimental set 3 in Table 3.4. As described earlier, we have six different emotions (angry, calm, fearful, happy, neutral, and sad).

Table 3.20 Confusion matrix of experimental set 3

		Predicted Labels					Precision	
		Angry	Calm	Fearful	Happy	Neutral		Sad
Actual Labels	Angry	87	1	1	4	1	2	90.63
	Calm	0	74	2	1	6	13	76.29
	Fearful	1	6	85	3	0	1	84.16
	Happy	4	1	4	81	2	4	88.04
	Neutral	2	6	3	2	79	4	85.87
	Sad	2	9	6	1	4	74	75.51
Recall		90.63	77.08	88.54	84.38	82.29	77.08	
F1 Score		90.63	77.08	86.29	86.17	84.04	76.28	

As seen, we have 90.63 precision for angry, 76.29 precision for calm, 84.16 precision for fearful, 88.04 precision for happy, 85.87 precision for neutral and 75.51 precision for sad emotions correspondingly. Results show that we have good precision values on emotions. It can be seen that angry, fearful, and happy emotions have higher classification accuracies than calm, sad and neutral emotions.

F1 Score refers to the measure of tests accuracy. In Table 3.20, we have a good result in an angry emotion.

Table 3.21 Confusion matrix of Energy+100 Pattern One versus All (S1, S2 separate)

		Predicted Labels					Precision	
		Angry	Calm	Fearful	Happy	Neutral		Sad
Actual Labels	Angry	88	3	1	4	0	0	93.62
	Calm	1	91	1	0	3	0	95.79
	Fearful	3	5	85	2	0	1	88.54
	Happy	2	0	4	85	5	0	91.40
	Neutral	0	4	2	2	88	0	91.67
	Sad	0	0	0	0	0	96	98.97
Recall		91.66	94.79	88.54	88.54	91.66	1	
F1 Score		92.63	92.85	89.94	89.94	91.67	99.48	

On the other hand, Table 3.21 shows another confusion matrix of combined acoustic and pattern based feature set in Table 3.12. We have pretty good results in precisions. The values are 93.62 for angry, 95.76 for calm, 88.54 for fearful, 91.40 happy, 91.67 neutral, and 98.97 for sad precision values accordingly. This shows that we have about 90 percent precision in this feature set. It can be seen that sad and calm emotions have remarkable classification results than other emotions. In Table 3.21, you can see that we have almost perfect F1 score for sad emotion with 99.48 value.



Chapter 4

Conclusion

In this research, we proposed a novel method for feature extraction in the recognition of emotion from speech. Before applying our approach, we applied some preprocessing technique to speech signals such as normalization and silence removal. We converted speech signals to discretized representations. Also, we extracted pattern features using top-k contrast mining technique from discretized speech signal representations. Extracted patterns are selected according to the highest C-Ratio. Moreover, we extracted acoustic features that are commonly used in literature, i.e. MFCC, RASTA, voice quality, energy and spectral. We analyzed classification performance with both pattern-based and acoustic features. In order to evaluate the performance, we used two different classification approaches i.e voting-based and DAGSVM. Voting results indicate that patterns alone are fairly good predictors of emotion. Experimental results show that a set of pattern features having different strategies outperformed all features in literature used. When the number of pattern feature set is increased, the classification enhances in term of accuracy. Despite the increase in complexity by the reason of the number of patterns and training, improvement of accuracy achieved up to 35 percent compared to state of art features i.e MFCC. When all acoustic features are supported by pattern-based features, emotion is predicted more accurately in classification. It gives results over 80 percent in different scenarios and techniques. In addition to this, we applied our approach in different frame sizes and alphabets to test and compare our accuracy results. Results show that 8 frame and 5 alphabet has the optimal results when

processing speech signals. We applied 100 pattern however there may be different results in accuracy when different pattern size is used. We applied McNemar statistical tests for examining the results of significant difference of two separated strategies. The results of McNemar tests show that both strategies one-versus-one and one-versus-all have quite similar error rates.

4.1 Future Work

There are lots of supportive and descriptive ideas that we can advance this research. First of all, our pattern-based approach can be applied to any data which are similar to the speech signal such as time-series, economy, EEG, EKG, music, and so on. In the field of medicine, the pattern-based approach can be used to detect a different emotion such as anxiety, depression, autism, and stress. The pattern-based approach may be combined with other information sources such as linguistic or video to improve the recognition performance. Common dimensionality reduction techniques e.g principal component analysis, information gain can be applied to our project to improve the classification accuracy. By changing the data set, emotions can be detected in different languages. Particularly, pattern extraction process takes too much time when the number of alphabet increases and frame size decreases. For this reason, this process requires both computational resources and time. Also, there are a certain number of implementations available online that have 'max-gap' parameter in contrast mining. As the algorithms using 'max-gap' parameter varies, we will be able to use different contrast mining implementations and benchmark the processing times against our current algorithm. In addition, HMMs and GMMs are very popular classification technique to recognize emotion in speech. We will be able to classify our presented method with these classification techniques in order to get a comparative result for improving accuracy. Therefore, we will have a chance to develop our methodology. Lastly, we are planning to add both disgust and surprise emotions to our approach so that we can detect more emotions in the speech.

BIBLIOGRAPHY

- [1] J. Lonardi and P. Patel, "Finding motifs in time series," in *Proc. of the 2nd Workshop on Temporal Data Mining*, 2002, pp. 53–68.
- [2] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.
- [3] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–577.
- [4] V. Petrushin, "Emotion in speech: Recognition and application to call centers," in *Proceedings of artificial neural networks in engineering*, vol. 710, 1999, p. 22.
- [5] C. M. Lee, S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal," in *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*. IEEE, 2001, pp. 240–243.
- [6] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.
- [7] J. H. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech communication*, vol. 20, no. 1-2, pp. 151–173, 1996.

- [8] M. Akagi, X. Han, R. Elbarougy, Y. Hamada, and J. Li, "Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. IEEE, 2014, pp. 1–10.
- [9] I. A. Essa and A. P. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 757–763, 1997.
- [10] P. Ekman, "Strong evidence for universals in facial expressions: a reply to russell's mistaken critique." 1994.
- [11] K. Mase, "Recognition of facial expression from optical flow," *IEICE Transactions on Information and Systems*, vol. 74, no. 10, pp. 3474–3483, 1991.
- [12] L. C. De Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information," in *Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat., vol. 1*. IEEE, 1997, pp. 397–401.
- [13] T. Otsuka and J. Ohya, "Recognizing multiple persons' facial expressions using hmm based on automatic extraction of significant frames from image sequences," in *Proceedings of International Conference on Image Processing*, vol. 2. IEEE, 1997, pp. 546–549.
- [14] M. Rosenblum, Y. Yacoob, and L. S. Davis, "Human expression recognition from motion using a radial basis function network architecture," *IEEE transactions on neural networks*, vol. 7, no. 5, pp. 1121–1138, 1996.
- [15] Z. Zeng, Y. Fu, G. I. Roisman, Z. Wen, Y. Hu, and T. S. Huang, "Spontaneous emotional facial expression detection." *Journal of multimedia*, vol. 1, no. 5, pp. 1–8, 2006.
- [16] Y. Pan, P. Shen, and L. Shen, "Speech emotion recognition using support vector machine," *International Journal of Smart Home*, vol. 6, no. 2, pp. 101–108, 2012.

- [17] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proceeding of Fourth International Conference on Spoken Language Processing. IC-SLP'96*, vol. 3. IEEE, 1996, pp. 1970–1973.
- [18] D. Ververidis, C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. I–593.
- [19] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," in *Eighth European Conference on Speech Communication and Technology*, 2003.
- [20] Y. Chavhan, M. Dhore, and P. Yesaware, "Speech emotion recognition using support vector machine," *International Journal of Computer Applications*, vol. 1, no. 20, pp. 6–9, 2010.
- [21] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: face, body gesture, speech," in *Affect and emotion in human-computer interaction*. Springer, 2008, pp. 92–103.
- [22] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 2004, pp. 205–211.
- [23] B. Dropuljić, S. Skansi, and R. Kopal, "Analyzing affective states using acoustic and linguistic features," in *Central European Conference on Information and Intelligent Systems (CECIIS)*, 2016.
- [24] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE transactions on speech and audio processing*, vol. 13, no. 2, pp. 293–303, 2005.
- [25] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Mariño, "Speech emotion recognition using hidden markov models," in *Seventh European Conference on Speech Communication and Technology*, 2001.

- [26] R. Cabredo, R. S. Legaspi, and M. Numao, "Identifying emotion segments in music by discovering motifs in physiological data." in *ISMIR*, 2011, pp. 753–758.
- [27] C. Shan, S. Gong, and P. W. McOwan, "Robust facial expression recognition using local binary patterns," in *IEEE International Conference on Image Processing 2005*, vol. 2. IEEE, 2005, pp. II–370.
- [28] A. Tiwari and T. H. Falk, "Fusion of motif-and spectrum-related features for improved eeg-based emotion recognition," *Computational intelligence and neuroscience*, vol. 2019, 2019.
- [29] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 915–928, 2007.
- [30] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raouf, M. A. Mahjoub, and C. Cleder, "Automatic speech emotion recognition using machine learning," in *Social Media and Machine Learning*. IntechOpen, 2019.
- [31] C. Breazeal and L. Aryananda, "Recognition of affective communicative intent in robot-directed speech," *Autonomous robots*, vol. 12, no. 1, pp. 83–104, 2002.
- [32] M. Slaney and G. McRoberts, "Baby ears: a recognition system for affective vocalizations," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, vol. 2. IEEE, 1998, pp. 985–988.
- [33] I. S. Engberg and A. V. Hansen, "Documentation of the danish emotional speech database des," *Internal AAU report, Center for Person Kommunikation, Denmark*, p. 22, 1996.
- [34] J. H. Hansen and S. E. Bou-Ghazale, "Getting started with susas: A speech under simulated and actual stress database," in *Fifth European Conference on Speech Communication and Technology*, 1997.
- [35] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International journal of speech technology*, vol. 15, no. 2, pp. 99–117, 2012.

- [36] 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.
- [37] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of german emotional speech,” in *Ninth European Conference on Speech Communication and Technology*, 2005.
- [38] V. Hozjan, Z. Kacic, A. Moreno, A. Bonafonte, and A. Nogueiras, “Interface databases: Design and collection of a multilingual emotional speech database.” in *LREC*, 2002.
- [39] V. Makarova and V. A. Petrushin, “Ruslana: A database of russian emotional utterances,” in *Seventh international conference on spoken language processing*, 2002.
- [40] N. Amir, S. Ron, and N. Laor, “Analysis of an emotional speech corpus in hebrew based on objective criteria,” in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.
- [41] F. Yu, E. Chang, Y.-Q. Xu, and H.-Y. Shum, “Emotion detection from speech to enrich multimedia content,” in *Pacific-Rim Conference on Multimedia*. Springer, 2001, pp. 550–557.
- [42] B. Zhang, G. Essl, and E. M. Provost, “Recognizing emotion from singing and speaking using shared models,” in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 139–145.
- [43] G. Dong and J. Bailey, *Contrast data mining: concepts, algorithms, and applications*. CRC Press, 2012.
- [44] C. Gao, L. Duan, G. Dong, H. Zhang, H. Yang, and C. Tang, “Mining top-k distinguishing sequential patterns with flexible gap constraints,” in *International Conference on Web-Age Information Management*. Springer, 2016, pp. 82–94.
- [45] H. Yang, L. Duan, B. Hu, S. Deng, W. Wang, and Qin, “Mining top-k distinguishing sequential patterns with gap constraint,” vol. 26, no. 11, pp. 2994–3009, 2015.

- [46] X. Ji, J. Bailey, and G. Dong, “Mining minimal distinguishing subsequence patterns with gap constraints,” *Knowledge and Information Systems*, vol. 11, no. 3, pp. 259–286, 2007.
- [47] H. Yang, L. Duan, G. Dong, J. Nummenmaa, C. Tang, and X. Li, “Mining itemset-based distinguishing sequential patterns with gap constraint,” in *International Conference on Database Systems for Advanced Applications*. Springer, 2015, pp. 39–54.
- [48] E. Väyrynen, “Emotion recognition from speech using prosodic features,” *University of Oulu, Oulu*, 2014.
- [49] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, “Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011,” *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, 2015.
- [50] H. Palo, P. Kumar, and N. Mohanty, “Emotional speech recognition using optimized features,” *International Journal of Research in Electronics and Computer Engineering*, vol. VOL. 5, pp. 4–9, 12 2017.
- [51] S. Wu, T. H. Falk, and W.-Y. Chan, “Automatic speech emotion recognition using modulation spectral features,” *Speech communication*, vol. 53, no. 5, pp. 768–785, 2011.
- [52] J. Lyons, “Mel frequency cepstral coefficient (mfcc) tutorial,” *Practical Cryptography*, 2015.
- [53] A. Firoz Shah, “Study and analysis of speech emotion recognition,” 2016.
- [54] Y.-R. Chien, D. D. Mehta, J. Gunason, M. Zañartu, and T. F. Quatieri, “Evaluation of glottal inverse filtering algorithms using a physiologically based articulatory speech synthesizer,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1718–1730, 2017.
- [55] U. Nam, “Spectral centroid,” 2001.
- [56] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, “The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion,

- autism,” in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, 2013*.
- [57] 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*. ACM, 2010, pp. 1459–1462.
- [58] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [59] G. Madzarov, D. Gjorgjevikj, and I. Chorbev, “A multi-class svm classifier utilizing binary decision tree,” *Informatica*, vol. 33, no. 2, 2009.
- [60] C.-W. Hsu and C.-J. Lin, “A comparison of methods for multiclass support vector machines,” *IEEE transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.
- [61] R. Sangeetha and B. Kalpana, “Identifying efficient kernel function in multi-class support vector machines,” *International Journal of Computer Applications*, vol. 28, no. 8, 2011.
- [62] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The weka data mining software: an update,” *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.
- [63] S. Ramakrishnan and I. M. El Emary, “Speech emotion recognition approaches in human computer interaction,” *Telecommunication Systems*, vol. 52, no. 3, pp. 1467–1478, 2013.
- [64] C.-W. Hsu, C.-C. Chang, C.-J. Lin *et al.*, “A practical guide to support vector classification,” 2003.
- [65] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” in *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 1992, pp. 144–152.
- [66] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, “Large margin dags for multiclass classification,” in *Advances in neural information processing systems*, 2000, pp. 547–553.

- [67] G. P. Kumari and M. U. Rani, "A study of adaboost and bagging approaches on student dataset."
- [68] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.
- [69] M. Sidorov, C. Brester, W. Minker, and E. Semenkin, "Speech-based emotion recognition: Feature selection by self-adaptive multi-criteria genetic algorithm."
- [70] S. Haq, P. J. Jackson, and J. Edge, "Audio-visual feature selection and reduction for emotion classification," in *Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP08), Tangalooma, Australia, 2008*.
- [71] A. Shirani and A. R. N. Nilchi, "Speech emotion recognition based on svm as both feature selector and classifier." *International Journal of Image, Graphics & Signal Processing*, vol. 8, no. 4, 2016.
- [72] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 2. IEEE, 2003, pp. II-1.
- [73] M. Song, J. Bu, C. Chen, and N. Li, "Audio-visual based emotion recognition a new approach," in *null*. IEEE, 2004, pp. 1020-1025.
- [74] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Tenth Annual Conference of the International Speech Communication Association, 2009*.
- [75] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Eleventh Annual Conference of the International Speech Communication Association, 2010*.
- [76] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The interspeech 2011 speaker state challenge," in *Twelfth Annual Conference of the International Speech Communication Association, 2011*.
- [77] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. v. Son, F. Weninger, F. Eyben, T. Bocklet *et al.*, "The interspeech 2012 speaker trait

challenge,” in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

- [78] F. Eyben, F. Weninger, N. Lehment, G. Rigoll, and B. Schuller, “Violent scenes detection with large, brute-forced acoustic and visual feature sets,” in *Proceedings MediaEval 2012 Workshop*, 2012.
- [79] C.-C. Chang and C.-J. Lin, “Libsvm: A library for support vector machines,” *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, p. 27, 2011.
- [80] M. MathWorks, “the mathworks,” *Inc., Natick, MA*, 1992.

Appendix A

Tools

In our research, we took advantage of some special toolkits and software to perform the research's implementation. In this chapter, we'll explain those toolkits and software in detail.

A.1 OpenSMILE

The Munich open-Source Media Interpretation by Large feature-space Extraction (openSMILE) [57] is a toolkit for signal processing and machine learning application. It is a modular and easy to use feature extractor tool. The design of openSMILE has a cross-platform environment for running all operating systems, e.g., Windows, Linux, and MacOS. Researchers who deal with the field of speech recognition, music information retrieval, and affective computing can use OpenSmile. The main aim of the toolkit is to focus on audio signal features. Alternatively, it may be utilized to examine other modularities, such as physiological and visual signals.

There are a couple of categories presented in OpenSMILE toolkit for end-users. These are signal processing, general data processing, low-level audio features, and other capabilities such as data input, functionals, classifiers and data output. Also, many functionalities are presented for signal processing or pre-processing to feature extraction,

e.g., windowing functions, Re-sampling, FFT, scaling spectral axis, and so on. OpenSMILE carries out a variety of operations for feature normalization, modification, differentiation such as mean-variance normalization, range normalization, etc. Low-level descriptors are related to audio signals. Both low-level descriptors and video features can be computed by openSMILE. Low-level descriptors include frame energy/ intensity/loudness, voice quality, critical harmonic ratios. Video features consist of LBP histogram or optical flow. Also, many statistical functionalities can be applied to these features. These statistics are the mean, regression, centroid segments, peaks, zero-crossings, etc.

OpenSMILE provides some default feature sets which are common in speech processing fields. These are standard feature sets :

- The INTERSPEECH 2009 Emotion Challenge feature set [74]
- The INTERSPEECH 2010 Paralinguistic Challenge feature set [75]
- The INTERSPEECH 2011 Speaker State Challenge feature set [76]
- The INTERSPEECH 2012 Speaker Trait Challenge feature set [77]
- The INTERSPEECH 2013 ComParE feature set [56]
- The MediaEval 2012 TUM feature set for violent scenes detection. [78]

Default feature sets are based on most comprehensive conference papers from spoken language processing. The papers are related to speech communication and science and support speech applications. Also, default feature sets are organized in accordance with the contents of the papers, such as the number of LLDs and applied statistics.

In our research, we used the default feature set of the Interspeech 2013 Computational Paralinguistics Evaluation (ComParE) [56] and extracted LLDs features. The feature set comprises of 4 energy, 41 spectral, 14 cepstral (MFCC), 6 voicing-related LLDs. The total number of features is 6373. (400 energy, 4100 spectral, 1400 MFCC, 473 voicing-related).

A.2 LibSVM

LibSVM [79] is a free library for support vector machines(SVM). It is developed by National Taiwan University. The aim of this library is to provide a convenient way for users to apply SVM to their applications. There is a built-in implementation of SVM for classification, regression, and distribution estimation. LibSVM also supports learning tasks. These tasks are support vector classification (SVC for multi-class and two-class), support vector regression (SVR) and one-class SVM. The usage of LibSVM includes two phases: first, a set of data is trained to obtain a model, and then this model is used for predicting information by processing a set of testing data.

Furthermore, LibSVM has a friendly user interface and principal features including efficient multi-class classification, probability estimates, cross-validation for model selection. This library also has many options for selecting desired parameters such as kernel type (linear, radial, polynomial), type of SVM (one versus one or one versus all methods), gamma and cost. When users need to determine some parameters to train SVM problems, LibSVM analyzes grid parameters. For each parameter, libSVM gets cross-validation (CV) accuracy. It returns the parameters with the highest CV accuracy.

In our research, we used SVC for multi-class with grid search method.

A.3 Matlab

Matlab is a computing platform which is designed for programming, is best suitable for analyzing data and designing computational algorithms. It is designed and developed by Mathworks. It is commonly used by researchers, scientists in multiple areas such as computation, visualization, data modeling and analysis and so on. It provides a user-friendly interface to perform tasks. It is highly effective when it comes to large data and tasks. It also comes with built-in libraries for visualizing the data on graphs. There are some additional benefits of Matlab: [80]

- External toolboxes are available in Matlab. These toolboxes can be used for signal processing, neural networks, simulation, etc.

- You can interact with programs that are written other programming languages like C or Fortran.
- A built-in mathematical function library is initially available and there is a vast collection of computational functions like sum, sine, cosine and some complex matrix operations functions.

In our research, we took advantage of Matlab and we performed the following tasks:

- Dimension reduction
- Discretization
- Create classes for positive and negative emotions
- Signal representation and plotting
- Cross-validation

We used the version of MATLAB 2017a to design and implement our researching algorithms.