

UNIVERSITY OF KASDI MERBAH OUARGLA
FACULTY of new Information Technologies and Communication
DEPARTMENT of computer science and Information Technology



ACADEMIC MASTER Thesis

Domain: Computer science and Information Technology

Faculty: Computer Science

Specialty: Industrial

Presented by:

□ Ms.Djeridi Dalal

□ Ms.Kedidi Rayhana

Topic

**Emotion recognition in
Arabic speech signal**

Submit Date: September 2020

before the jury:

□ President UKM OUARGLA

□ Mr.Belhadj Mourad Supervisor UKM OUARGLA

□ Examiner UKM OUARGLA

University year: 2019/2020

Abstract

Recognizing emotions has become an area of great interest to researchers in the past few years. Emotion recognition is a multidisciplinary area, among which is the recognition of emotions from speech. Recognizing speech emotion is a significant endeavor in human speech processing and developing human-computer interaction.

This work presents the performance of machine learning approaches for the recognition of emotions from an Arabic speech signal. Initially, we used the Lebanese audio database Arabic-Natural-Audio-Dataset (ANAD), which contains 384 records with 505 happy, 137 surprises, and 741 angry units. Next, we use the OpenSMILE toolkit to extract the necessary speech features with two methods, Low-Level Descriptors (LLDs) with 988 features, and Mel-frequency cepstral coefficient (MFCC) with 39 features. Also, we applied features selection on LLDs and MFCC using Learner Based Feature Selection. We suggested Rough set theory for select features in order to improve results. Then, for classifying the emotions into different classes, Multilayer Perceptron (MLP), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Logistic Regression (LR) are employed. Results showed that MLP outperformed other models when applied on LLDs and MFCC features with accuracy 87%, 83% respectively.

Keywords— Arabic speech signal, speech emotion recognition, low-level descriptors(LLDs), Mel-frequency cepstral coefficient (MFCC), Machine learning.

ملخص

أصبح التعرف على المشاعر مجال اهتمام كبير للباحثين في السنوات القليلة الماضية. التعرف على المشاعر هو مجال متعدد التخصصات ، من بينها التعرف على المشاعر من الكلام. التعرف على عاطفة الكلام هو مسعى مهم في معالجة الكلام البشري وتطوير التفاعل بين الإنسان والحاسوب.

يقدم هذا العمل أداء مناهج التعلم الآلي للتعرف على المشاعر من إشارة الكلام العربي. في البداية ، استخدمنا قاعدة البيانات الصوتية اللبنانية (ANAD) Arabic-Natural-Audio-Dataset ، والتي تحتوي على 384 تسجيلًا مع 505 وحدة سعيدة و 137 مفاجأة و 741 وحدة غاضبة. بعد ذلك ، نستخدم مجموعة أدوات OpenSMILE لاستخراج ميزات الكلام الضرورية بطريقتين ، الواصفات منخفضة المستوى (LLDs) مع 988 ميزة ، ومعامل تردد الميل (MFCC) مع 39 ميزة. أيضًا ، قمنا بتطبيق اختيار الميزات على LLDs و MFCC باستخدام طريقة اختيار الميزات على أساس المتعلم . اقترحنا نظرية Rough set لتحديد الميزات من أجل تحسين النتائج ، ثم لتصنيف المشاعر إلى فئات مختلفة تم استخدام متعدد الطبقات (MLP) ، آلة متجه الدعم (SVM) ، و الجار الاقرب (KNN) ، والانحدار اللوجستي (LR) . أظهرت النتائج تفوق MLP على النماذج الأخرى عند تطبيقها على LLDs و MFCC بدقة 87% ، 83% على التوالي.

الكلمات المفتاحية-- إشارة الكلام باللغة العربية، التعرف على عاطفة الكلام، الواصفات منخفضة المستوى (LLDs) ،معامل تردد الميل (MFCC) ، التعلم الآلي.

Résumé

La reconnaissance des émotions est devenue un domaine de grand intérêt pour les chercheurs au cours des dernières années. La reconnaissance des émotions est un domaine multidisciplinaire, parmi lesquels la reconnaissance des émotions à partir de la parole. Reconnaître l'émotion de la parole est une entreprise importante dans le traitement de la parole humaine et le développement de l'interaction homme-machine.

Ce travail présente les performances d'approches d'apprentissage automatique pour la reconnaissance des émotions à partir d'un signal de parole arabe. Au départ, nous avons utilisé la base de données audio libanaise Base-de-Donnée-Audio-Naturel-Arabe (BANA), qui contient 384 enregistrements avec 505 joyeux, 137 surprises et 741 unités en colère. Ensuite, nous utilisons la boîte à outils OpenSMILE pour extraire les caractéristiques vocales nécessaires avec deux méthodes, des Descripteurs de Bas Niveau (DBN) avec 988 caractéristiques et le Coefficient Cepstral de Fréquence Mel (CCFM) avec 39 caractéristiques. En outre, nous avons appliqué la sélection des fonctionnalités sur les DBN et CCFM à l'aide de Sélection de fonctionnalités basée sur l'apprenant. Nous avons suggéré la théorie des ensembles approximatifs pour certaines caractéristiques afin d'améliorer les résultats. Ensuite, pour classer les émotions en différentes classes, le Perceptron Multicouche (PMC), la Machine à Vecteur de Support (MVS), les K-Voisins les plus Proches(KVP) et la Régression Logistique (RL) sont utilisés. Les résultats ont montré que PMC surpassait les autres modèles lorsqu'il était appliqué sur les fonctionnalités DBN et CCFM avec une précision de 87%, 83% respectivement.

Mots clé— signal de parole Arabe, reconnaissance des émotions vocales, descripteurs de bas niveau (DBN), Coefficient Cepstral de Fréquence Mel (CCFM), Apprentissage automatique



ACKNOWLEDGMENTS

For this THESIS we thank **ALLAH**
who helped us and provided us with patience and courage during these years of study.

Thank you to **Mr. Mourad Belhadj**
for his patience and efforts to achieve good results in our work.

We would like to express our sincere thanks to the people who helped us and who contributed to the development of this message as well as the success of this great academic year.





DEDICATIONS

After conciliation from God Almighty, I dedicate this humble work to:

My dear parents, who took care of raising me, and helped me in my life.

Dear brothers and sisters.

All my relatives, friends and loved ones.

Honorable Mr. Belhadj Mourad, who shone the way for us.

Everyone who taught me letters in my school life.

Djeridi Dalal





DEDICATIONS

With the grace of God first and foremost I dedicate this humble work:

To my Mother

You have given me life, the tenderness and the courage to succeed. Everything that I can offer you will not be able to express the love and the recognition that I carry. As a testimony, I offer you this modest job to thank you for your sacrifices and for the affection with which you have always surrounded me.

To my Father

The solid shoulder, the attentive, understanding eye and the person most worthy of my esteem and respect. No dedication can express our feelings, may God preserve you and give you health and long life with health.

To our Esteemed Professor

Who was the first reason after God Almighty, that we reached the end of this work and whenever we asked for part of his precious time he provided us despite all the circumstances Dr. Belhadj Mourad

To my Brothers and my Sisters

To my fiance and my friends

To all my Relatives, Loved ones, and Colleagues

Kedidi Rayhana



Contents

Abstract	i
Acknowledgements	iv
Dedication	v
General Introduction	1
General Introduction	2
Motivation	2
Aims and Objectives	2
Description of the work	3
Related Work	3
Arabic studies	3
External studies	4
1 Background Information	7
1.1 Introduction	8
1.2 Speech	8
1.2.1 Speech Signal	8
1.2.2 Speech Signal processing	9
1.3 Emotions	9
1.3.1 Taxonomy of Emotions	10
1.3.2 Emotion Recognition	10
1.4 Speech Emotion Recognition System	11
1.5 Arabic Language	11
1.5.1 Language and express emotion	13

1.6	Conclusion	13
2	Materials and Methods	14
2.1	Introduction	15
2.2	Emotional Speech Dataset	15
2.2.1	Types of Emotional Speech Dataset	15
2.2.2	Arabic Emotional Speech Datasets	16
2.3	Feature Extraction	18
2.3.1	OpenSMILE toolkit	18
2.3.2	Low Level Descriptors (LLDs)	19
2.3.3	Mel Frequency Cepstral Coefficient (MFCC)	19
2.4	Feature Selection	21
2.4.1	Learner Based Feature Selection	22
2.4.2	Rough set for Features Selection	23
2.5	What is MultiLayer Perceptron (MLP)?	27
2.6	What is Support Vector Machine (SVM)?	27
2.6.1	SVM Kernels	28
2.7	What is K-Nearest Neighbors(KNN)?	29
2.7.1	How do you decide the number of neighbors in KNN?	29
2.8	What is Logistic Regression(LR)?	30
2.8.1	Types of Logistic Regression	30
2.9	Conclusion	31
3	Implementation and Results	32
3.1	Introduction	33
3.2	Tools	33
3.2.1	Operating System	33
3.2.2	Programming language	33
3.2.3	Scikit-learn Python library	34
3.2.4	PyTorch library	34
3.3	Implementation and Results	35
3.3.1	Preparing the Data	36
3.3.2	Results of Learner Based Feature Selection	37

3.3.3 Results of Rough set theory	41
3.4 Evaluation	45
3.4.1 Results analysis	46
3.5 Conclusion	47
General Conclusion and Future Work	48
Bibliography	50

List of tables

1.1	The fundamental emotions by Robert Plutchik [22]	10
2.1	Datasets of Arabic emotional speech	17
2.2	Videos used to build the corpus [32]	18
3.1	Classification Report of MLP	37
3.2	Classification Report of SVM	38
3.3	Classification Report of KNN	39
3.4	Classification Report of LR	40
3.5	Test Accuracy by Model with WEKA features selection	41
3.6	Classification Report of MLP	42
3.7	Classification Report of SVM	42
3.8	Classification Report of KNN	42
3.9	Classification Report of LR	43
3.10	Test Accuracy by Model with Rough set features selection	43

List of figures

1.1	Representation of properties of sound wave[51]	9
1.2	Plutchik’s wheel of emotions[22]	11
2.1	Screenshot of the OpenSMILE command console during feature extraction	21
2.2	Screenshot of the WEKA during feature selection	22
2.3	Explain rough set theory[38]	24
2.4	Multilayer Perceptron (MLP)[43]	27
2.5	Support Vector Machine[57]	28
2.6	How does the KNN algorithm work?[9]	29
2.7	Graph for the sigmoidal function[2]	30
3.1	Results of MLP	37
3.2	Results of SVM	38
3.3	Results of KNN	39
3.4	Results of LR	40
3.5	Results	44
3.6	Train-Val Accuracy/Epoch plot	45
3.7	Normalized confusion matrix of (LLDs_MLP)	45
3.8	Normalized confusion matrix of (MFCC_MLP)	46

General Introduction

General Introduction

Imagine you are in a noisy, public place and someone wants to communicate with you, and because of the noise you can't hear him well, but you can deduce his emotional state through his facial expression. Now imagine another situation and this person you can't see him but you can hear his voice by phone for example, and therefore you can know how he feels just by talking.[3]

In recent years, many audio treatments have been introduced to identify different things such as age, gender, translation, and many human characteristics, and among the most important of these characteristics is emotion, because knowing how a person feels solves many problems, and emotion is one of the main factors in communication between humans although different languages. In our lives, we need to express our emotions through many actions such as our faces, our hands, and our voices, The latter has a huge impact on our dealings, take for example a baby can distinguish the emotion of his parents from the tone of their voice, he feels panic if the sound is frightening, and he feels happy if the voice is compassionate or funny despite his lack of understanding of the true meaning of the spoken words. Therefore the sound is a good standard for knowing human emotion.

Motivation

With the rapid development of technology, recognition of human emotion through the device has become a reality. We expect the machine in the future has become superior to humans in this field, which will lead to a large degree of natural communication between humans and the machine. So automatic emotion recognition systems can help people in many research areas, such as diagnosing the psychological state, smart games, detection about lying, smart call center, and educational programs.

Aims and Objectives

We aim through this study to extract emotions from the Arabic language. Comparing the Arabic language with its English or Chinese counterparts, we find that it is almost constrained. Knowing the emotions generated by the speech through the machine makes the interaction smoother and helps solve many problems in the Arab world. In general, in this work we have depended on three major steps in detecting emotion, the first is

selecting of emotional states which means what are the principal emotions we need to know, or we need to detect, this includes anger, happiness and surprise. The second step is the extraction of features from the speech : LLDs, MFCC, then select the useful features which eliminate unwanted features and organize the dataset. The last step is selecting classifiers to recognize emotions, which means the application of machine learning models on this dataset and give the results, here we worked with MLP, SVM, KNN and LR.

Objectives :

- To suggest the development of Android applications which can be used to detect peoples' emotions for better health.
- To offer better services and also better interaction between humans and machines.

Description of the work

The rest of the thesis is organized as follows: in the last part of the general introduction, we explain the related works with the study, including Arabic and external studies. Chapter I describe the necessary information related to the Arabic speech signal. Chapter II presents the Materials and methods used for the dataset and in writing code. The Implementation of the code ,results analysis and evaluation showed in Chapter III. We concluded the thesis with the general conclusion that explains the essential steps that we discussed in this work and what can be applied in the future.

Related Work

Arabic studies

Recognizing a person's emotion through speech Also known as Speech Emotion Recognition (SER) is the ability to distinguish feelings from one another through the tone and pitch of this voice. This phenomenon is used when young children and animals to know human emotions.[16]

Recognition emotion from Arabic Speech is new field research. The main reason for this is the absence of a comprehensive Arabic database; there are few types of research in the Arabic language, and this is a group of recent Arab studies:

Egyptian work: They created a semi-natural Egyptian Arabic speech emotion (EYASE)

database from the award-winning Egyptian TV series that includes pronouncements from 3 male and 3 female professional actors considering four emotions: angry, happy, neutral and sad. Prosodic, spectral, and wavelet functions are measured for emotion recognition from the EYASE database. In addition to the commonly applied classical pitch, intensity, formants, and Mel-frequency cepstral coefficients (MFCC) for SER, this work also considers the long-term average spectrum (LTAS) wavelet parameters.[1]

Emirati work: They use a two-step system for describe the unknown emotion. The first stage focuses on identifying the speaker who pronounced the unknown emotion. In contrast, the second stage focuses on identifying the unknown emotion conveyed in the identified speaker's prior stage. This proposed framework was evaluated on a speech database based on Arabic Emirates spoken by fifteen speakers per gender. Mel-Frequency Cepstral Coefficients (MFCCs) were used as the extracted features and Hidden Markov Model (HMM) was used as the classifier in this study. [53]

KSU work: This research aims to explore the potential use of speech rhythm metrics as a new function for recognition of speech emotion, gender identification and identification of regional accents. Also, it seeks to determine a new corpus of Arabic speech emotion. The speech corpus of the King Saud University Emotions (KSUEmotions) contains five emotions: neutral, sad, happy, surprising and anger. In this research, acoustic features of speech are extracted and used to identify the emotions of the speakers. All classification results were obtained using neural networks and support vector machine (SVM) classifiers and the multilayer perceptron (MLP).[40]

Algerian work: They construct an Algerian dialect (ADED) emotional database. This database has four emotions: anger, fear, sadness and neutral. Pitch, intensity, length, unvoiced photos, jitter, shimmer, HNR , MFCCs are the characteristics extracted in this study. The method used for classification is the KNN (K-Nearest Neighbor).[24]

External studies

In general, there are many types of research in the world for emotion detection by speech using different programs and methods, we are going to show some new researches in this field.

In [45] a review from **Indians** about Techniques for Speech Emotion Recognition uses Convolutional Neural Network and Long Short-Term Memory(LSTM) for testing emotion

capturing from various standard speech representations, such as Mel Spectrogram, magnitude Spectrogram and Mel-Frequency Cepstral Coefficients on two popular datasets EMO-DB and IEMOCAP. The good result obtained is 82.35%, achieved for CNN + BLSTM architecture with MFCC as input.

In [36] presents a comparative study of speech emotion recognition (SER) systems. Mel-frequency cepstrum coefficients (MFCC) and modulation spectral (MS) features are extracted from the speech signals and used to train different classifiers. **Berlin** and **Spanish** databases are used as the experimental data set. This study shows that all classifiers for the Berlin database achieve an accuracy of 83% when a speaker normalization (SN) and a feature selection (FS) are applied to the features. For Spanish database, the best accuracy (94 %) is achieved by recurrent neural network (RNN) classifier without SN and with FS.

In [45] an experimental study on the detection of emotion from speech. This study utilizes a corpus containing continuous emotional speech comprised of 9 different emotions: Activation, Valence, Power, Intensity, Amusement, Happiness, Sadness, and Anger. They created an RNN based model and a CNN-based model. The algorithm makes use of MFCC features, pitch, magnitudes, and spectral roll-off. These features are fed into a basic neural net and an LSTM.

In [22] presents the implementation of accurate emotion recognition with the deep learning model of Convolutional Neural Networks (CNN). The architecture is an adaptation of an image processing CNN, programmed in Python using Keras model-level library and Tensor-Flow back end. The theoretical background that lays the foundation of the classification of emotions based on voice parameters is briefly presented. According to the obtained results, the model achieves the mean accuracy of 71.33% for six emotions (happiness, fear, sadness, disgust, anger, surprise).

In [54] the efficacy of convolutional neural networks in recognition of speech emotions has been investigated. Wide-band spectrograms of the speech signals were used as the input features of the networks. The networks were trained on speech signals that were generated by the actors while acting a specific emotion. The speech databases with different languages were used to train and evaluate their models. The training data on each database were augmented with two levels of augmentations. The dropout technique was implemented to regularize the networks. The results showed that the gender-independent,

language-independent CNN models achieved the state-of-the-art accuracy, outperformed previously reported results in the literature, and emulated or even surpassed human performance over the benchmark databases.

Chapter 1

Background Information

1.1 Introduction

Speech is an effective way to express human needs. Through the tone of a person's words, we deduce what he wants and precisely what he feels. The human voice is characterized by different characteristics depending on his feelings or needs. With the Speech processing, we can collect different features that let us differentiate between the voices.

This chapter explains the various terms and techniques used in the field of sound processing and extraction of its features, and the field of emotions with the allocation of the Arabic language as a focus of study.

1.2 Speech

Speech is the ability to move thoughts, ideas, or other information through articulating sound into meaningful words.[34] Speech includes:

Articulation: How we make speech sounds using the mouth, lips, and tongue. For example, we need to be able to say the "r" sound to say "rabbit" instead of "wabbit."

Voice: How we use our vocal folds and breath to make sounds. Our voice can be loud or soft or high- or low-pitched. We can hurt our voice by talking too much, yelling, or coughing a lot.

Fluency: This is the rhythm of our speech. We sometimes repeat sounds or pause while talking. People who do this a lot of may stutter.[8]

1.2.1 Speech Signal

According to Schetelig T. and Rabenstein R. in May 1998, the speech signal is a multidimensional acoustic wave (as shown in Figure 1.1) which provides information about the words or message being spoken, speaker identity, the language spoke, physical and mental health, race, age, sex, education level, religious orientation and background of an individual.[51]

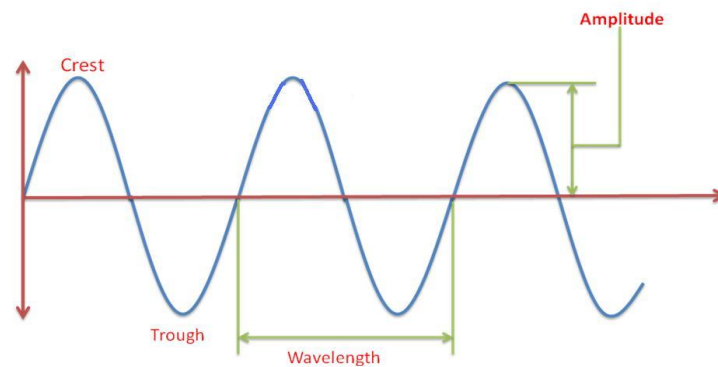


Figure 1.1: Representation of properties of sound wave[51]

1.2.2 Speech Signal processing

Speech processing is the study of speech signals and signal processing methods. The signals are typically interpreted in a digital form, and it is possible to consider speech processing as a particular case of digital signal processing, applied to speech signals. Aspects of speech processing include receiving, controlling, storing, transmitting, and providing voice signals. The input is called speech recognition, and the output is called speech synthesis.

Applications

- Interactive Voice Systems
- Virtual Assistants
- Voice Identification
- Emotion Recognition
- Call Center Automation
- Robotics[63]

1.3 Emotions

There are many difficulties in defining the notion of emotion. According to Don Hockenbury and Sandra E. Hockenbury in the book "Discovering Psychology", an emotion is a complex psychological state that involves three distinct components: a subjective experience, a physiological response, and a behavioral or expressive response.[30]

The aroused state accompanies the characteristics of emotions in the organism, usually accompanied by physiological changes and much energy is released in every emotion, ex-

cept grief. Emotions are divided into two types: positive emotion like happiness, trust; negative emotion like aggression, fear, sadness...etc. Emotion generates changes in the body an external and an internal. External changes are facial expressions, vocal expression, and posture expression (Body Language); internal changes are the heartbeat, blood pressure, blood chemistry, galvanic skin response, metabolic changes, and brain waves.[29]

1.3.1 Taxonomy of Emotions

In psychology, there is no consensus classification of emotions, and this is one of the barriers encountered in research on discovering emotions. Psychology distinguishes feelings, emotions, and influences (depending on severity, duration, and persistence).

Among the most famous classifications in the world is the classification of the American psychologist Robert Plutchik, who has classified emotions into eight primary categories (Table 1.1) grouped into four pairs of opposites, resulting in their combination of secondary emotions. Figure 1.2 illustrates this representation of emotions of varying values. So that strong emotions are in the center.[22]

joy	sadness
trust	disgust
fear	anger
anticipation	surprize

Table 1.1: The fundamental emotions by Robert Plutchik [22]

1.3.2 Emotion Recognition

Audiovisual emotion recognition is not a new problem; it is extensive research. There has been much work in visual pattern recognition for facial emotional expression recognition and signal processing for audio-based detection of emotions and many multimodal approaches combining these cues or writing texts. However, improvements in hardware, availability of datasets and wide-scale annotation infrastructure made it possible to create real effective systems a reality. We now see applications across many domains, including robotics, HCI, healthcare, and multimedia.[7]



Figure 1.2: Plutchik's wheel of emotions[22]

1.4 Speech Emotion Recognition System

Speech Emotion Recognition (SER) is the task of recognizing Speech's emotional aspects irrespective of the semantic contents[35]. There are five main modules in the speech emotion recognition system consist of an input is the emotional speech, there is feature extraction, feature selection, the classifier of the emotions, and recognized emotional output. The structure of the Speech emotion recognition system illustrated in Figure 1.3.[61]



Figure 1.3: Structure of the Speech Emotion Recognition System

1.5 Arabic Language

The Arabic language is distinguished from all other international languages by a set of characteristics:

Sounds: one of the main characteristics of the Arabic language is that the pronunciation system is considered one of the essential systems of linguistic Speech, so the tongue,

throat, and throat are used to pronounce letters and words based on their sounds, and the sounds in the Arabic language are divided into a group of sections, such as occlusal voices and throat sounds, and others.

Vocabulary: these are the words that make up the Arabic language, and the private lexicon is classified as one of the most vocabulary rich in vocabulary and structures; it contains more than a million words. The original vocabulary in the Arabic language is considered to be the triple root of the other words, and one linguistic root produces many words and vocabulary.

Pronunciation: it is how the words of the Arabic language are pronounced, and the words are pronounced using language movements, and it is called the formation. The specific word in each word changes based on the nature of its formation; that is, the movements written on its letters, and the term includes special spelling in the letters, which every person who wants to learn Arabic learns it; so that it is easier for him to understand it, and to deal with its words and sentences correctly.

Exchange: it is the method associated with vocabulary, as it depends on the system of the roots of words that are often triple, and may become quadrilateral at times, as the Arabic language is distinguished from many other languages by the presence of formulas for its own words, it is possible to convert the single word into a mean, or Plural, and other methods that the Arabic language uses to classify words.

Syntax: is the basis of the sentence in the Arabic language, and the Arabic sentences are divided into two types, namely: the nominal sentence and the actual sentence, and each type of these sentences has grammatical foundations and rules that must be used in writing and formulating them to contribute to the transmission of their ideas, and also the grammar is adopted in the Arabic language on the use of a set of tools that link the sentences, and many other means that maintain the integrity of its structure, therefore it classifies the Arabic language as one of the languages that maintains its grammatical system, and helps in expressing its sentences and showing their writing methods.[15]

So far, the Arabic language and its dialects are a relatively missing resource when compared to other languages such as English, and Arabic has three forms:

- **Classical Arabic or literary Arabic:** Literary Arabic, usually called Classical Arabic, is essentially the form of the language found in the Qurʾān, with some

modifications necessary for its use in modern times; it is uniform throughout the Arab world.[44]

- **Modern Standard Arabic(MSA):** is the direct result of modern Arabic and a standardized version of the language. It is the type of Arabic used in universities, Arabic language schools, audiovisual and written media, and other formal contexts.
- **Colloquial Arabic(dialect):** In contrast to modern standard Arabic or MSA, we find another type of Arabic far from the formalities of academic settings, diplomatic relations, and the media. Colloquial Arabic is the Arabic dialect specific to each area. Although most of its vocabulary and grammatical roots come from the MSA, it also incorporates its lexicon as a result of its historical past. Therefore, colloquial Arabic gives rise to many variants in the same country and even within the same region.[4]

1.5.1 Language and express emotion

Another source of emotions outside the true sense is the sound of a voice, that is, the spoken language's phonetic and acoustic properties. Thus, a cracking voice, for example, is evidence of high excitement. Recognition of acoustic emotions is the principal focus of this thesis. The vocal parameters that were best investigated in psychological studies and the most intuitively important ones are prosody (pitch, intensity, speaking rate) and voice quality.[18]

1.6 Conclusion

The conversion of speech into a signal is considered the first step in identifying emotions through speech, This step follows signal processing and eliminating noise. The difference in expressing emotions in the language results in a difference in the resulting signal. This is why the Arabic language was chosen as the focus of the study to distinguish it from other languages.

Chapter 2

Materials and Methods

2.1 Introduction

Many studies have focused on developing human-machine interaction by making them more flexible and effective. Among these studies, the machine can distinguish emotions through your voice. These studies need a dataset of the voices of people whose characteristics vary according to the needs and objectives of the study. Then the sound features are extracted and finally the identification of emotion. There are many available databases in the world related to this area, including dataset containing a combination of languages. Some of them focus on one language.

2.2 Emotional Speech Dataset

Getting good emotional performance depends on choosing an appropriate database. To prepare a satisfactory database must be respected four criteria: scope, physical presence, contents and actual chosen language. These criteria include the number of speakers, the type of speaker, the type of emotions, the number of dialects, and the type of language and age. Speech emotion databases are widely available in languages like German, English, Japanese, Danish, Spanish, Hebrew, Swedish, Russian, Chinese, and Greek.[58]

2.2.1 Types of Emotional Speech Dataset

Most verbal emotional databases are not enough to simulate emotions clearly and naturally, and the quality of data affects the success of the recognition process. In-complete, low-quality, or insufficient data may lead to incorrect expectations; Therefore, the data should be designed and collected carefully. Databases for recognizing emotions can be divided into three types:

- **Acted (Simulated) speech emotion databases :** It is relatively easier to create such a database than other methods. The words that are represented are merely identical to the type of emotions. Some speech sets represented consist of recordings with professional actors and in some other recordings with mature artists in soundproof studios. However, researchers have stated that acted speech cannot adequately convey real life emotions, and may be overrated. This reduces recognition rates for real emotions.

- **Elicited (Induced) speech emotion databases :** In elicited speech, some emotions are stimulated by using certain specific actions that must provoke certain emotions. It is also possible to put the topic in a situation intended to provoke a particular emotion and then record the speaker’s speech. Although emotions are not fully clear, they are close to real emotions.
- **Natural speech emotion databases:** Natural speech databases are mostly obtained from talk shows, call center records, radio chats and similar sources. The precise emotional circumstances of the speakers are recorded by hidden mechanisms without the speakers knowing, so that they behave spontaneously. But collecting data for natural emotions is not so easy. [5, 21]

2.2.2 Arabic Emotional Speech Datasets

Arabic Datasets for speech recognition are very few compared to their counterparts from other languages, Table (Table 2.1) illustrates the scarcity of Arabic language resources. Also, it confirms that little work has been devoted to analyzing emotional Speech in the Arabic language.

Name	Type	MSA or dialect	Speakers	Linguistic material	Emotions
KSU Emotions (2014, 2018)	acted	MSA	20 actors	16 sentences	neutral, sadness, happy, surprised, and questioning
Egypt(2017)	acted	MSA	7	500 sentences	happiness, sadness, fear, anger, inquiry, neutral
REGIM_TES (2015-17)	acted	Tunisian	12 actors		Happiness, Anger, Neutral, Fear, Sadness

Name	Type	MSA or dialect	Speakers	Linguistic material	Emotions
Egyptian Arabic speech emotion (EYASE)	acted	dialect	6 actors	Egyptian TV series	angry, happy, neutral and sad
Emotional database of the Algerian dialect (ADED)	acted	dialect	32 actors	6 movies in Algerian dialect	anger, fear, sadness and neutral
Arabic-Natural Audio-Dataset (2018)	Natural	Dialectal: Egyptian, Jordan, Gulf, Lebanese	6 actors	Eight videos of live calls	happiness, anger, and surprise

Table 2.1: Datasets of Arabic emotional speech

In this study, the Lebanese database (Arabic-Natural Audio-Dataset (2018)) was chosen to be the focus of the study because it is the only natural database in order to obtain results comparable to reality.

Arabic-Natural Audio-Dataset (ANAD): Among the Arabic talk shows on the Internet, eight video clips were collected between the anchor and a person outside the studio. Then the video clips were named by human designers after listening to them either happy, angry, or surprised, and the average result is taken into consideration. Silence, laughter, and loud pieces have been removed. Each piece was divided into two-second speech units to form the final set of 1384 records with 505 happy, 137 surprises and 741 angry units. Table (Table 2.2) summarizes the characteristics of the videos used to build the body.[32]

ID	Dialect	Gender	Length (s)	#of chunks	Emotion perceived
1	Egyptian	Male	114	9	Happy
2	Egyptian	Male	78	6	Surprised
3	Gulf	Female	73.8	6	Happy
4	Jordan	Male	210	17	Angry
5	Gulf	Male	198	34	Angry
6	Egyptian	Female	23.4	2	Surprised
7	Lebanese	Female	504	24	Angry
8	Egyptian	Female	430.8	87	Happy

Table 2.2: Videos used to build the corpus [32]

2.3 Feature Extraction

Feature extraction consists of extracting key features from speech samples and creating large vectors, theoretically, it should be possible to recognize speech directly from the digital wave form. However, due to the large variation in the speech signal, it is better to make some extraction features that will reduce this contrast. In particular, eliminating different sources of information, such as whether the sound is expressed or not, and if expressed, it eliminates the effect of frequency or volume, the expansion of the excitement signal, the basic frequency etc., to extract features from speech samples, OpenSMILE data mining tools have been used.[56]

2.3.1 OpenSMILE toolkit

It is a modular and flexible feature extractor for signal processing and machine learning applications. The primary focus is clearly put on audio-signal features. However, due to their high degree of abstraction, OpenSMILE components can also be used to analyse signals from other modalities, such as physiological signals, visual signals, and other physical sensors, given suitable input components. It is written purely in C++, has a fast, efficient, and flexible architecture, and runs on various main-stream platforms such as Linux, Windows, and MacOS. OpenSMILE is designed for real-time online processing and can also be used offline in batch mode for processing large datasets. OpenSMILE can extract features incrementally as new data arrives. Using the PortAudio library, OpenSMILE features platform independent live audio input and live audio playback, which enabled the extraction of audio features in real-time.[20]

In this study we compared two types of features, we use OpenSMILE toolkit to extract these features from the speech signal. The first type of audio features, extract through the configuration file "emobase.conf" that enables us to extract 988 features which are the following low-level descriptors(LLDs), the second type is Mel Frequency Cepstral Coefficient (MFCC) through the "MFCC12_0_D_A.conf" configuration file which computes 13 MFCC , 13 delta and 13 acceleration coefficients are appended to the MFCC, In total we get 39 features.

2.3.2 Low Level Descriptors (LLDs)

The feature set specified by emobase.conf contains the following low-level descriptors (LLDs):Intensity, Loudness, 12 MFCC, Pitch (F0), Probability of voicing, F0 envelope, 8 LSF (Line Spectral Frequencies), Zero-Crossing Rate. Delta regression coefficients are computed from these LLDs, and the following functionals are applied to the LLDs and the delta coefficients: Max./Min. value and respective relative position within input, range, arithmetic mean, 2 linear regression coefficients and linear and quadratic error, standard deviation, skewness, kurtosis, quartile 1–3, and 3 inter-quartile ranges.[20] These features are obtained by performing the following instruction:

```
SMILExtract -C config/emobase.conf -I inputN.wav -O output.csv inputN -classes {Angry, Happy, Surprised} -classlabel Angry
```

2.3.3 Mel Frequency Cepstral Coefficient (MFCC)

MFCC are the most widely-used acoustic feature for speech recognition, speaker recognition, and audio classification; it takes into account specific properties of the human auditory system.[46]

The following steps explain how to calculate MFCC :

- 1- Frame the signal into short frames.
- 2- For each frame calculate the periodogram estimate of the power spectrum.
- 3- Apply the Mel filter-bank to the power spectra, sum the energy in each filter.
- 3- Take the logarithm of all filter bank energies.
- 4-Take the DCT (Discrete Cosine Transform)of the log filter-bank energies.
- 5- Keep DCT coefficients 2-13, discard the rest.

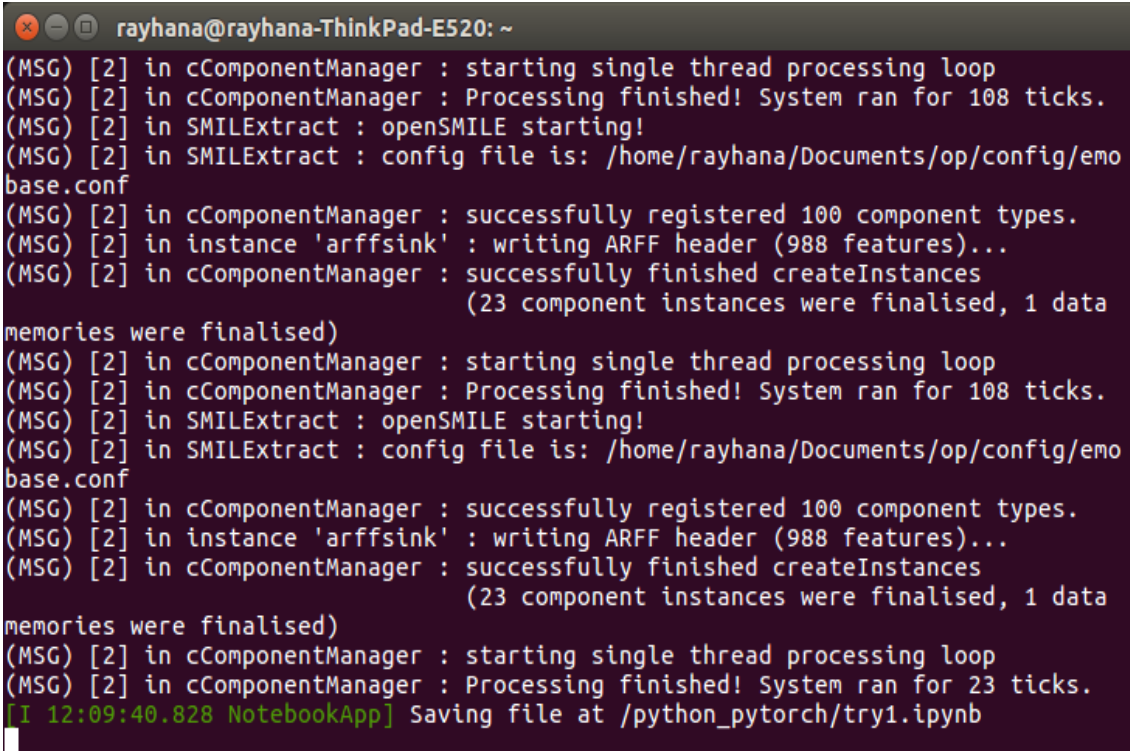
There are a few things that are commonly done, we applied it by attached the features of Delta and Delta-Delta, Also known as differential and acceleration coefficients. The MFCC feature vector describes only the power spectral envelope of a single frame, but it seems like speech would also have information in the dynamics i.e. what are the trajectories of the MFCC coefficients over time. It turns out that calculating the MFCC trajectories and appending them to the original feature vector increases ASR (Automatic Speech Recognition) performance by quite a bit (if we have 12 MFCC coefficients, we would also get 12 delta coefficients, which would combine to give a feature vector of length 24).[48]

All these features are obtained by performing the following instruction:

```
SMILExtract -C config/MFCC12_0_D_A.conf -I input.wav -O output.mfcc.csv
```

Features (LLDs, MFCC) are extracted via OpenSMILE through the console commands (Figure 2.1), and this complicates the process of extracting the features and takes a lot of time, especially since the number of voices is not a little. Therefore we used a Python code that allows us to process a set of sounds in an iterative way and in record time :

```
[ ]: import os
for i in range (1,138):
    os.system("SMILExtract -C /home/user/Documents/op/config/MFCC12_0_D_A.
↵conf -I /home/user/Documents/arabic-natural-audio-dataset/Surprised/
↵"+str(i)+".wav -O /home/user/Documents/arabic-natural-audio-dataset/
↵Surprised1/surprised"+str(i)+".csv")
```

```
rayhana@rayhana-ThinkPad-E520: ~
(MSG) [2] in cComponentManager : starting single thread processing loop
(MSG) [2] in cComponentManager : Processing finished! System ran for 108 ticks.
(MSG) [2] in SMILEExtract : openSMILE starting!
(MSG) [2] in SMILEExtract : config file is: /home/rayhana/Documents/op/config/emo
base.conf
(MSG) [2] in cComponentManager : successfully registered 100 component types.
(MSG) [2] in instance 'arffsink' : writing ARFF header (988 features)...
(MSG) [2] in cComponentManager : successfully finished createInstances
(23 component instances were finalised, 1 data
memories were finalised)
(MSG) [2] in cComponentManager : starting single thread processing loop
(MSG) [2] in cComponentManager : Processing finished! System ran for 108 ticks.
(MSG) [2] in SMILEExtract : openSMILE starting!
(MSG) [2] in SMILEExtract : config file is: /home/rayhana/Documents/op/config/emo
base.conf
(MSG) [2] in cComponentManager : successfully registered 100 component types.
(MSG) [2] in instance 'arffsink' : writing ARFF header (988 features)...
(MSG) [2] in cComponentManager : successfully finished createInstances
(23 component instances were finalised, 1 data
memories were finalised)
(MSG) [2] in cComponentManager : starting single thread processing loop
(MSG) [2] in cComponentManager : Processing finished! System ran for 23 ticks.
[I 12:09:40.828 NotebookApp] Saving file at /python_pytorch/try1.ipynb
```

Figure 2.1: Screenshot of the OpenSMILE command console during feature extraction

Executing this code will produce a set of CSV files with a header containing all the feature names and one instance, which contains a feature vector for the specified input file. The CSV file will have a fake class named Emotion, containing one category angry, happy or surprised. The `-classlabel` parameter specifies the category / value of the calculated instance of the currently given entry (-I).

2.4 Feature Selection

After extracting 988 features(LLDs) from the Speech Signal, we applied an optimization process to these features due to the presence of many unrelated features and containing information that might hinder recognition rates, which allows us to obtain better performance of the algorithm, for that we chose to compare two methods for features selection:

2.4.1 Learner Based Feature Selection

We use WEKA environment, which is an open source software, It is a collection of machine learning algorithms for data mining tasks. It is undoubtedly a powerful tool that we will use for the processing stage of our features selection. It is widely used for teaching, research, and industrial applications, contains a plethora of built-in tools for standard machine learning tasks, and additionally gives transparent access to well-known toolboxes such as scikit-learn, R, and Deep learning.[62]

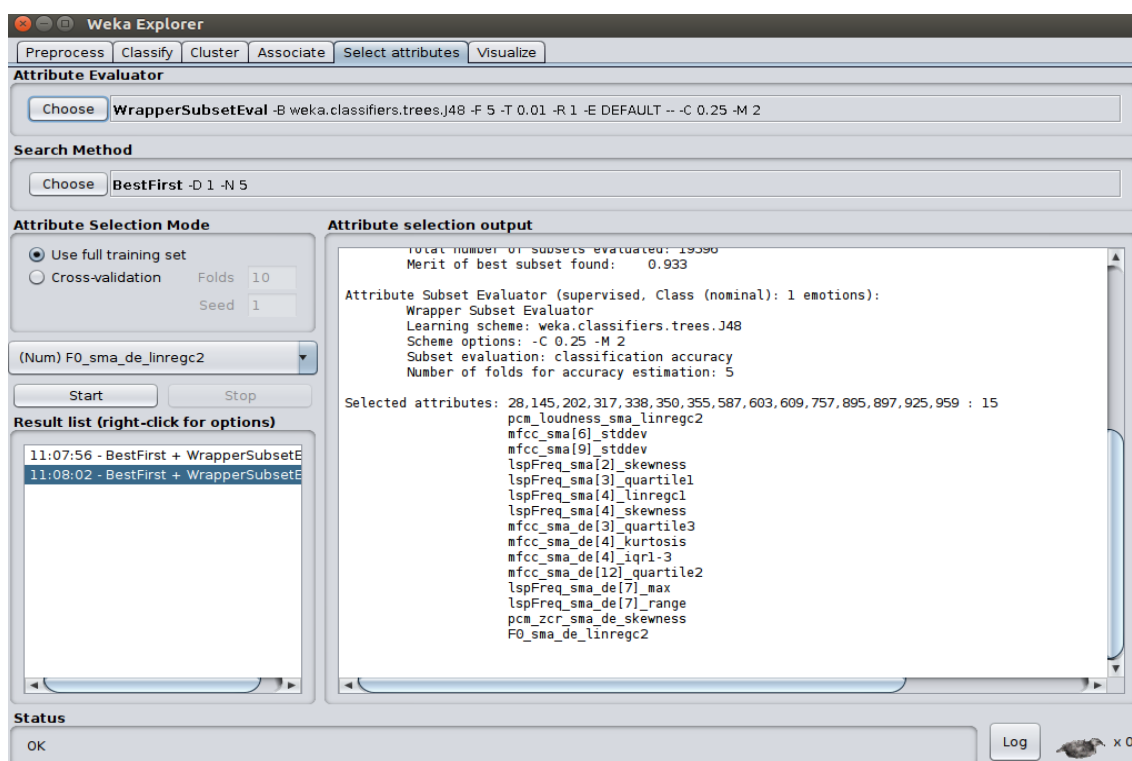


Figure 2.2: Screenshot of the WEKA during feature selection

Feature selection is divided into two sections: **Attribute Evaluator**, **Search Method**. Each section has multiple techniques from which to choose.

The attribute evaluator is the technique by which each attribute in your dataset (also called a column or feature) is evaluated in the context of the output variable (e.g. the class). The search method is the technique by which to try or navigate different combinations of attributes in the dataset in order to arrive at short list of chosen features. Some Attribute Evaluator techniques require the use of specific Search Methods.

As shown in the Figure (Figure 2.2), we apply the **Learner Based Feature Selection** which is a popular feature selection technique is to use a generic but powerful learning

algorithm and evaluate the performance of the algorithm on the dataset with different subsets of attributes selected. The subset that results in the best performance is taken as the selected subset. The algorithm used to evaluate the subsets does not have to be the algorithm that you intend to use to model your problem, but it should be generally quick to train and powerful, like a decision tree method.

We followed the steps below to select features:

- 1-Open the Weka GUI Chooser.
- 2-Click the “Explorer” button to launch the Explorer.
- 3-Open the dataset.
- 4-Click the “Select attributes” tab to access the feature selection method then click “WrapperSubsetEval”.
- 5-Click on the name “WrapperSubsetEval” to open the configuration for the method.
- 6-Click the “Choose” button for the “classifier” and change it to J48 under “trees”.
- 7- Click “OK” to accept the configuration.
- 8-Change the “Search Method” to “BestFirst”.
- 9-Click the “Start” button to evaluate the features.[25]

After applying the previous steps, we got 15,7 features for LLDs, MFCC respectively.

2.4.2 Rough set for Features Selection

Rough set theory is still another approach to vagueness. Rough set theory can be viewed as a specific implementation of Frege’s idea of vagueness. The rough set concept can be defined quite generally using topological operations, interior and closure, called approximations.

2.4.2.1 The basic concepts of rough set theory

U and A , two finite, non-empty set.

U is the universe.

A is a set of attributes.

$a \in A$ we associate a set V_a , of its values, called the domain of a .

B is a subset of A .

- *Indiscernibility relation* is a binary relation $I(B)$ on U , and is defined as follows: $x I(B) y$ if and only if $a(x) = a(y)$ for every $a \in A$, where $a(x)$ is the value

of attribute a for element x .

- The B -lower approximation of X is subset X of the universe U is denoted by $B_*(X)$. Can be defined as follows,

$$B_*(X) = \{x \in U : B(x) \subseteq X\}$$

- The B -upper approximation of X is subset X of the universe U is denoted by $B^*(X)$. Can be defined as follows,

$$B^*(X) = \{x \in U : B(x) \cap X \neq \emptyset\}$$

- The B -boundary region of a set X is:

$$BN_B(X) = B^*(X) - B_*(X)$$

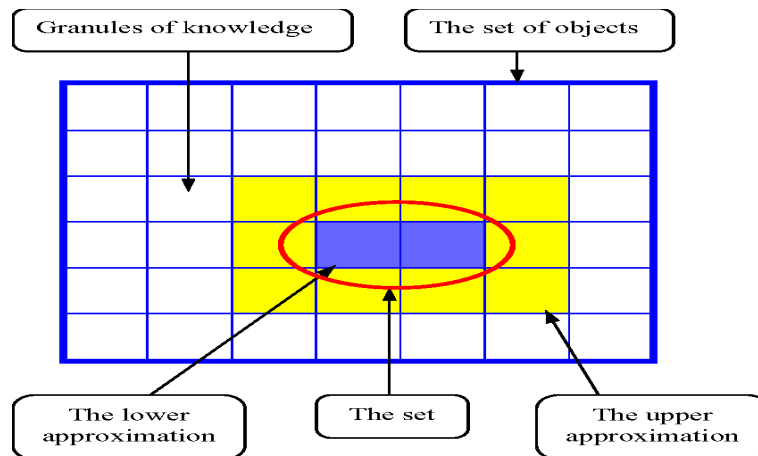


Figure 2.3: Explain rough set theory[38]

2.4.2.2 Decision Tables and Decision Algorithms

In an information table, two classes of attributes called condition and decision (action) attributes.

the row of a table of decisions defines a decision rule that describes decisions (actions) that should be made when conditions defined by condition attributes are met. Objects in a decision table are used as labels of decision rules.

The rules have the same conditions, but different decisions are *inconsistent*; otherwise,

the rules are *consistent*.

In a decision table, The number of consistent rules to all rules can be used as a consistency factor denoted by $\gamma(C, D)$, where C and D are condition and decision attributes respectively. Thus if $\gamma(C, D) = 1$ the decision table is consistent and if $\gamma(C, D) \neq 1$ the decision table is inconsistent.

Decision rules are often presented as implications called “if...then...” rules. A set of decision rules is called a *decision algorithm*. Decision table is a collection of data; decision algorithm is a collection of implication.

2.4.2.3 Dependency of attributes

Set of attributes D *depends totally* on a set of attributes C if there is a functional dependency between the values of D and C denoted by $C \Rightarrow D$.

The *partial dependency* means that values of C determine only some values of D.

Dependency can be defined in the following way:

D and C be subsets of A.

We say that D depends on C in a degree k ($0 \leq k \leq 1$), denoted $C \Rightarrow_k D$, if $k = \gamma(C, D)$.

- If $k = 1$ we say that D depends totally on C
- If $k < 1$ we say that D depends partially (in a degree k) on C.

The coefficient k expresses the ratio of all elements of the universe, which can be properly classified to blocks of the partition U/D , employing attributes C.

If D depends totally on C then $I(c) \subseteq I(D)$

If D depends in degree k, ($0 \leq k \leq 1$), on C, then

$$\gamma(C, D) = \frac{|POS_c(D)|}{|U|}$$

where

$$POS_c(D) = \cup_{X \in U/I(D)} C_*(X)$$

$POS_c(D)$ called a positive region of the partition U/D with respect to C.

2.4.2.4 Reduction of Attributes

the table contains some superfluous data for this we need to remove this data from a data table preserving its basic properties.

We need some auxiliary notions to express the idea more precisely. Let B be a subset of A and let a belong to B .

- We say that a is *dispensable* in B if $I(B) = I(B - \{a\})$; otherwise a is *indispensable* in B .
- Set B is *independent* if all its attributes are indispensable.
- Sub set B' of B is a *reduct* of B if B' is independent and $I(B') = I(B)$.

There are some essential properties of the reducts. We'll discuss two of them in what follows.

Core of attributes: Let B be a subset of A . The core of B is the set off all B 's indispensable attributes.

$$Core(B) = \cap Red(B),$$

where $Red(B)$ is the set off all reducts of B .

relative reduct: Let $C, D \subseteq A$. Obviously if $C' \subseteq C$ is a D -reduct of C , then C' is a minimal subset of C such that

$$\gamma(C, D) = \gamma(C', D)$$

- We will say that attribute $a \in C$ is D -dispensable in C , if $POS_C(D) = POS_{C-a}(D)$; otherwise the attribute a is D -indispensable in C .
- If all attributes $a \in C$ are C -indispensable in C , then C will be called D -independent.
- Subset $C' \in C$ is a D -reduct of C , iff C' is D -independent and $POS_C(D) = POS_{C'}(D)$.

The set of all D -indispensable attributes in C will be called D -core of C , and denoted by $CORE_D(C)$. In this case we have also the property

$$CORE_D(C) = \cap Red_D(C),$$

where $Red_D(C)$ is the family of all D -reducts of C .

If $D = C$ we will get the previous definitions [38].

We applied the notions of this theory, we got 11 and 13 features of LLDs and MFCC respectively.

2.5 What is MultiLayer Perceptron (MLP)?

A multilayer perceptron (MLP) is a deep, artificial neural network. It consists of more than one perceptron. They consist of an input layer to receive the signal, an output layer that makes a decision or prediction about the input, and in between input and output layer exist an arbitrary number of hidden layers that are the true computational engine of the MLP (Figure 2.4). In the forward pass, the signal flow moves from the input layer to the output layer through the hidden layers, and the decision of the output layer is measured against the ground truth labels.[14] The idea of the backpropagation algorithm is, based on error (or loss) calculation, to recalculate the weights array W in the last neuron layer towards the previous layers, that is, to update all the weights W in each layer, from the last one until arriving the input layer of the network, for this doing the backpropagation of the error obtained by the network. In other words, we calculate the error between what the network predicted to be and what it was indeed. Always intending to decrease the neural network error.[59]

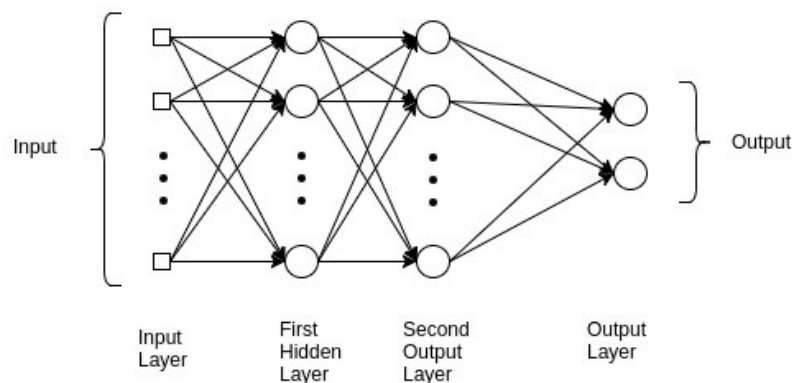


Figure 2.4: Multilayer Perceptron (MLP)[43]

2.6 What is Support Vector Machine (SVM)?

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n -dimensional space (where n is number of features you have) with each feature's value being the value of a particular coordinate. Then we perform classification by finding

the hyperplane which very well differentiates the two classes.[57] The Figure 2.5 below explains this.

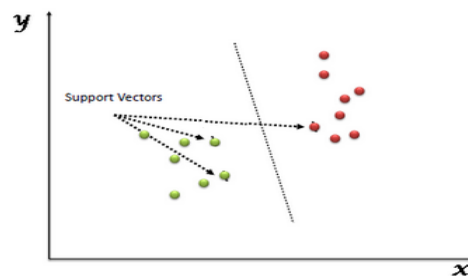


Figure 2.5: Support Vector Machine[57]

2.6.1 SVM Kernels

The SVM algorithm is implemented in practice using a kernel. A kernel transforms the given dataset input data into the required form. SVM uses a technique called the kernel trick. It converts nonseparable problems to separable problems by adding more dimension to it. It is most useful in a non-linear separation problem. Kernel trick helps you to build a more accurate classifier.

Linear Kernel A linear kernel is the dot product between the input(x) and each support vector (x_i) calculated as follows:

$$K(x, x_i) = \text{sum}(x * x_i)$$

Polynomial Kernel A polynomial kernel is a more generalized form of the linear kernel. The polynomial kernel can distinguish curved or nonlinear input space.calculated as follows:

$$K(x, x_i) = 1 + \text{sum}(x * x_i)^d$$

Radial Basis Function Kernel RBF RBF can map an input space in infinite-dimensional space.[10]calculated as follows:

$$K(x, x_i) = \exp(-\text{gamma} * \text{sum}((x-x_i)^2))$$

2.7 What is K-Nearest Neighbors(KNN)?

KNN is one of the more simple techniques used in machine learning. This algorithm can be used for Classification as well as for Regression but mostly it is used for the Classification problems. This algorithm is a model that classifies data points based on the points that are most similar to it. It uses test data to make an “educated guess” on what an unclassified point should be classified as.

KNN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.[39, 26]

Using distance measures such as Euclidean distance, Hamming distance, Manhattan distance and Minkowski distance, you can find the distance between points for closest related points. The following basic steps are taken at KNN (Figure 2.6):

- 1- Calculate distance
- 2- Find closest neighbors
- 3- Vote for labels

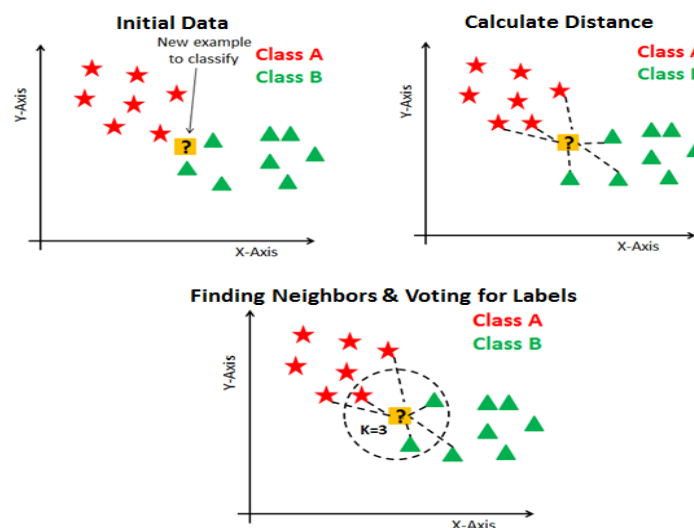


Figure 2.6: How does the KNN algorithm work?[9]

2.7.1 How do you decide the number of neighbors in KNN?

Neighbors(K) number in KNN is a hyperparameter you need to use when constructing a model. For the prediction model you can think of K as a controlling variable.

Research has shown that no optimal number of neighbours is suitable for all forms of data sets. Every dataset has needs of its own. For general, if the number of groups is even, data scientists select as an odd number. You can also inspect and test their output by generating the model on different values of k.[9]

2.8 What is Logistic Regression(LR)?

One of the most basic and commonly used Machine Learning Algorithms is logistic regression. Logistic regression is typically one of the first few topics that people select when studying predictive modelling. It is not an algorithm of regression but a probabilistic classification model used to assign discrete classes to observe. It transforms its output using the logistic sigmoid function (Figure 2.7) to return a probability value that can then be generalized to two or more discrete groups as opposed to linear regression that outputs continuous number values.[13, 2]

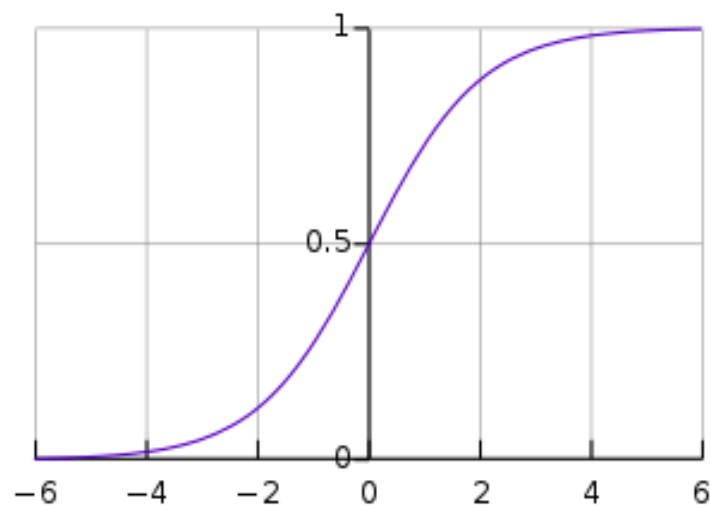


Figure 2.7: Graph for the sigmoidal function[2]

2.8.1 Types of Logistic Regression

Binary Logistic Regression: The target variable has only two possible outcomes such as Spam or Not Spam, Cancer, or No Cancer.

Multinomial Logistic Regression: The target variable has three or more nominal categories, such as predicting the fruit type.

Ordinal Logistic Regression: the target variable has three or more ordinal categories

such as restaurant or product rating from 1 to 5.[11]

2.9 Conclusion

In this chapter, all the resources and methods that we used during our study were explained, starting with building two different databases in the features extracted (LLDs, MFCC) from the Lebanese audios using the OpenSMILE program, then we selected the critical features using two methods : Learner Based Feature Selection and Rough set theory. Finally, we moved on to explain the machine learning's models which we will use them for predict our emotion recognition system: MLP, SVM, KNN and LR.

Chapter 3

Implementation and Results

3.1 Introduction

Studies of understanding emotions through Speech know a great difference in terms of the approaches followed to obtain stronger and more similar outcomes to reality. Take the difference in the programming language, for example, so it is important to choose the programming language and the functions and libraries before writing the code because of its great importance in strengthening the study. The Python language was chosen in this study, as it is currently one of the strongest languages, especially in the field of machine learning. Most machine learning models have produced amazing results in many fields, so we focus on applying them to identify and compare emotions.

3.2 Tools

3.2.1 Operating System

For Operating System (OS) we chose Linux / Ubuntu 16.04 LTS with processor Intel® Core™ i3-3110M CPU @ 2.40GHz × 4 , mainly due to its flexibility when running applications from the console command line. Some primarily required software applications were developed to run under Linux, and even offer fewer conflicts if installed via console command entries rather than Windows platforms. This is the case of OpenSMILE. However, other used software applications, such as Rough set, also have strong and reliable behavior under Win OS.

3.2.2 Programming language

The programming language chosen in this study is the Python language in edition Anaconda 3.6. We wrote the code in Jupyter Notebook: version 6.0.1.

Python language is a widely-used general-purpose, high-level programming language. It was created by Guido van Rossum in 1991 and further developed by the Python Software Foundation. It was designed with an emphasis on code readability, and its syntax allows programmers to express their concepts in fewer lines of code. Python is a programming language that lets you work quickly and integrate systems more efficiently. There are two major Python versions: Python 2 and Python 3. Both are quite different.[23]

Anaconda is a data science and machine learning platform for the Python and R programming languages. It is designed to make the process of creating and distributing projects simple, stable and reproducible across systems and is available on Linux, Windows, and OSX. Anaconda is a Python based platform that curates major data science packages including pandas, scikit-learn, SciPy, NumPy and Google's machine learning platform, TensorFlow. It comes packaged with conda (a pip like install tool), Anaconda navigator for a Graphical User Interface (GUI) experience, and spyder for an Integrated Development Environment (IDE).[12]

Jupyter Notebook is an open-source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebooks are a spin-off project from the IPython project, which used to have an IPython Notebook project itself. The name, Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R. [41]

3.2.3 Scikit-learn Python library

The Scikit-learn Python library, first published in 2007, is widely used to solve machine learning and data science issues — from start to finish. The flexible library provides an uncluttered, reliable, and secure API as well as comprehensive documentation online.

Scikit-learn is an open-source Python library, with robust data analysis and data mining tools. It is built on the machine learning libraries that follow:

- **NumPy** is a library for manipulating multi-dimensional arrays and matrices. It also has an extensive compilation of mathematical functions for performing various calculations.
- **SciPy** is an ecosystem consisting of various libraries for completing technical computing tasks.
- **Matplotlib** is a library for plotting various charts and graphs.

We can use it in main ways: Classification, Regression, Clustering, Dimensionality reduction, Model selection and Preprocessing. [19]

3.2.4 PyTorch library

PyTorch, Created by Facebook in October 2016, it is an open-source machine learning library for Python, based on Torch.[47, 37] We can install this library by instruction:

```
conda install pytorch torchvision cpuonly -c pytorch[50]
```

In [37], they compared the parameters that distinguish PyTorch, Keras, and TensorFlow like the level of Application Programming Interface (API), speed, architecture, debugging, dataset, and popularity. Which they found Keras is most suitable for rapid prototyping, small dataset, and multiple back-end support; counter to, TensorFlow is most suitable for a large dataset, high performance, functionality, and object detection; PyTorch is most suitable for flexibility, short training duration and debugging capabilities. [37]

3.3 Implementation and Results

In this part, we will show the characteristics of each model that has been applied. As for the results, they will be presented with two properties:

1- Confusion Matrix : The confusion matrix has three axes:

- 1- Prediction label (class).
- 2- True label.
- 3- Heat map value (color).

The prediction label and true labels show us which prediction class we are dealing with. The matrix diagonal represents locations in the matrix where the prediction and the truth are the same, so this is where we want the heat map to be darker.

Any values that are not on the diagonal are incorrect predictions because the prediction and the true label don't match. To read the plot, we can use these steps:

- Choose a prediction label on the horizontal axis.
- Check the diagonal location for this label to see the total number correct.
- Check the other non-diagonal locations to see where the network is confused.[17]

We create a dataframe from the confusion matrix and plot it as a heatmap using the seaborn library.

2- Classification Report: Is used to measure the quality of predictions from a classification algorithm. How many predictions are True and how many are False. More specifically, True Positives, False Positives, True negatives and False Negatives are used to predict the metrics of a classification report[42].

In classification report we have :

Precision : True Positive / Total Positive

Recall : True Positive / (True positive + False Negative)

F1- score : Is the harmonic mean of precision and recall

Support : The support is the number of samples of the true response that lie in that class.

In other words precision tells you how good is your model at rejecting the wrong labels and recall tells you how good is your model at assigning correct labels to the true classes.[49]

Accuracy : The sum of true positives and true negatives divided by the total number of samples. This is only accurate if the model is balanced. It will give inaccurate results if there is a class imbalance.[31]

Macro-average: A macro-average will compute the metric independently for each class and then take the average (hence treating all classes equally).[33]

weighted average : averaging the support-weighted mean per label(The weighted metrics consider the number of instances of each class).[52]

3.3.1 Preparing the Data

We load the csv file as a data frame then we create input and output data in which the input X is all columns except the emotion's column which is y. The CSV file does not contain any information about the data type to be inferred which may lead to errors. To specify data types for different columns, we use the astype parameter.[55]

We split the data into train and test using `train_test_split()` from Sklearn. Because there's a class imbalance, we want to have equal distribution of all output classes in our train and test sets. To do that, we used the stratify option in function `train_test_split()`. For MultiLayer Perceptron We divided the database in another way to train, validation and test using `train_test_split()`. So first we will split the data two times, the first to train+validation and test sets , second to train and validation.

Many machine learning algorithms work better when features are on a relatively similar scale and close to normally distributed. `MinMaxScaler` is a scikit-learn method to preprocess data, On the other hand Neural networks need data that lies between the range of (0,1) which allowing the network to learn the optimal parameters for each input node more quickly.[27, 6, 28]

3.3.2 Results of Learner Based Feature Selection

3.3.2.1 MultiLayer Perceptron MLP

In our work has used five layers:

- **In the input layer:** 15, 7 neurons which are the number of features of LLD, MFCC respectively.
- **In the first hidden layer:** 250 neurons.
- **In the second hidden layer:** 150 neurons.
- **In the third hidden layer:** 35 neurons.
- **In the output layer:** 3 neurons are the number of classes.

Train model:

model.train() tells PyTorch that you're in training mode and model.eval() in testing mode, We apply Validation set in training part.

Model Parameters: NUM_EPOCHS = 700, LEARNING_RATE = 0.0007

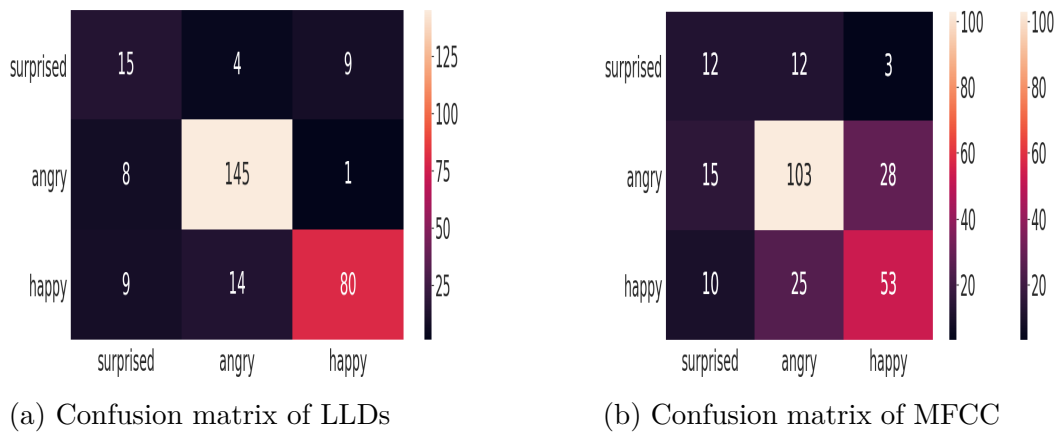


Figure 3.1: Results of MLP

		precision	recall	f1-score	support
Surprised	LLDs	0.47	0.54	0.50	28
	MFCC	0.32	0.44	0.038	27
angry	LLDs	0.89	0.94	0.91	154
	MFCC	0.74	0.71	0.72	146
happy	LLDs	0.89	0.78	0.83	103
	MFCC	0.63	0.60	0.62	88
macro avg	LLDs	0.75	0.75	0.75	285
	MFCC	0.56	0.58	0.57	261
weighted avg	LLDs	0.85	0.84	0.84	285
	MFCC	0.66	0.64	0.65	261
accuracy	LLDs	84%			
	MFCC	64%			

Table 3.1: Classification Report of MLP

3.3.2.2 Support Vector Machine (SVM)

After dividing the data into sets for training and testing. We train the training data on our SVM. Scikit-Learn contains the svm library which includes integrated classes for various SVM algorithms. Since we will perform a classification task, we use the class of support vector classifier, which is written as SVC in the svm library of the Scikit-Learn. This class will take one parameter, which is the kernel type. That is extremely important. We simply set this parameter as "linear" for example in the case of a simple SVM. The SVC class's fit method is called to train the algorithm on the training data, which is passed as a parameter to the fit method.[60] We chose the linear kernel with LLDs features, and sigmoid kernel with MFCC features because they gave the best results.

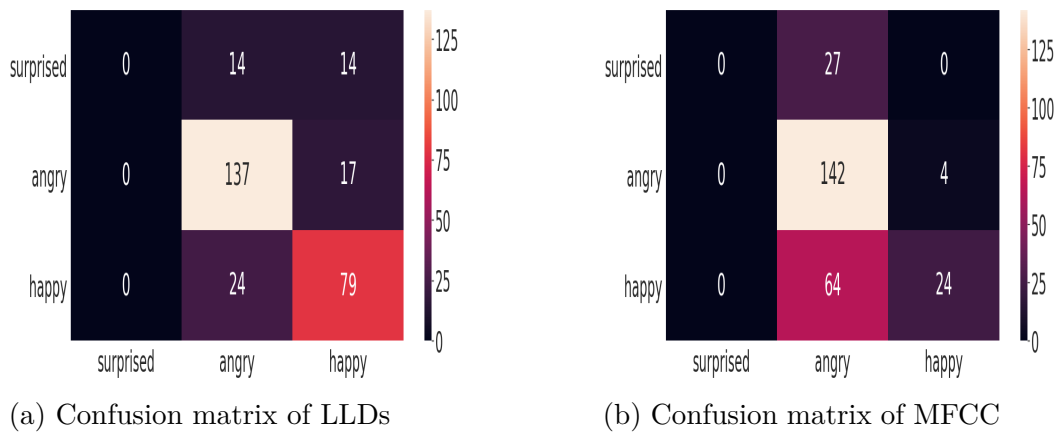


Figure 3.2: Results of SVM

		precision	recall	f1-score	support
Surprised	LLDs	0.00	0.00	0.00	28
	MFCC	0.00	0.00	0.00	27
angry	LLDs	0.78	0.89	0.83	154
	MFCC	0.61	0.97	0.75	146
happy	LLDs	0.72	0.77	0.74	103
	MFCC	0.86	0.27	0.41	88
macro avg	LLDs	0.50	0.55	0.52	285
	MFCC	0.49	0.42	0.39	261
weighted avg	LLDs	0.68	0.76	0.72	285
	MFCC	0.63	0.64	0.56	261
accuracy	LLDs	76%			
	MFCC	64%			

Table 3.2: Classification Report of SVM

3.3.2.3 K-Nearest Neighbor(KNN)

First, import the KNeighborsClassifier module and construct KNN classifier object by passing neighboring argument number in the function KNeighborsClassifier(). Then fit the model with fit() onto the train package.

We use KNN with multiple classes, in our case 3 emotions. We chose k=7 because it provided a better result to us.

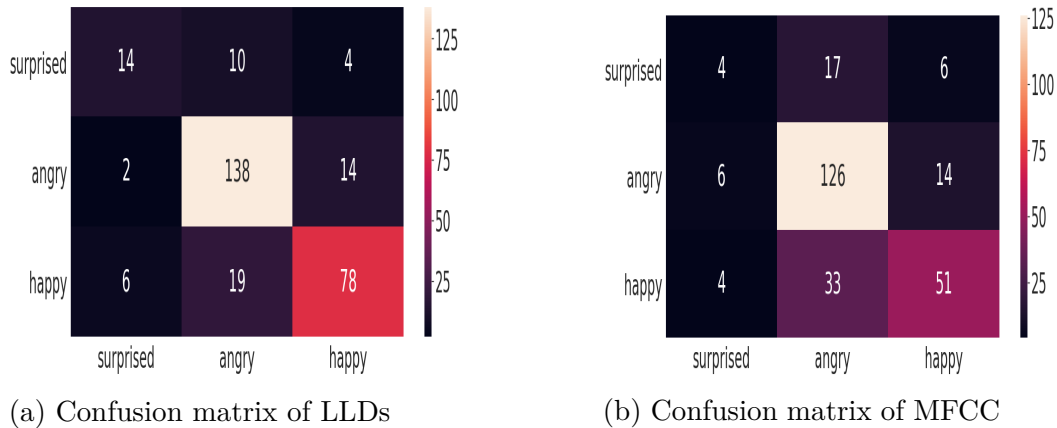


Figure 3.3: Results of KNN

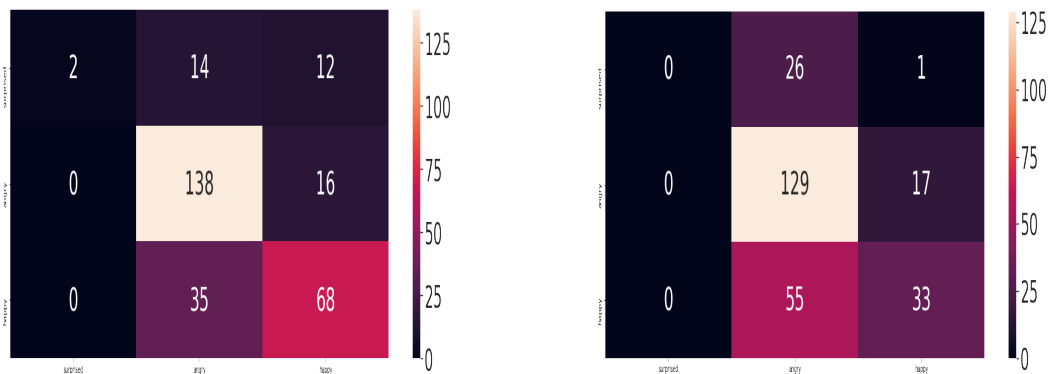
		precision	recall	f1-score	support
Surprised	LLDs	0.64	0.50	0.56	28
	MFCC	0.29	0.15	0.20	27
angry	LLDs	0.83	0.90	0.86	154
	MFCC	0.72	0.86	0.78	146
happy	LLDs	0.81	0.76	0.78	103
	MFCC	0.72	0.58	0.64	88
macro avg	LLDs	0.76	0.72	0.73	285
	MFCC	0.57	0.53	0.54	261
weighted avg	LLDs	0.80	0.81	0.80	285
	MFCC	0.67	0.69	0.67	261
accuracy	LLDs	81%			
	MFCC	69%			

Table 3.3: Classification Report of KNN

3.3.2.4 Logistic Regression (LR)

In the multi class logistic regression python Logistic Regression class, multi-class classification can be enabled/disabled by passing values in the algorithm constructor to the argument called "multi class". In the multiclass case, if the 'multi class' option is set to 'ovr' the training algorithm uses the one-vs-rest (OvR) scheme.[2]

Scikit-learn ships with five different solvers. Each solver tries to find the parameter weights that minimize a cost function. We evaluated the logistic regression solvers in a multi-class classification problem with our data and chose 'lbfgs' solver, it gave good results.



(a) Confusion matrix of LLDs

(b) Confusion matrix of MFCC

Figure 3.4: Results of LR

		precision	recall	f1-score	support
Surprised	LLDs	1.00	0.07	0.13	28
	MFCC	0.00	0.00	0.00	27
angry	LLDs	0.74	0.90	0.81	154
	MFCC	0.61	0.88	0.72	146
happy	LLDs	0.71	0.66	0.68	103
	MFCC	0.65	0.38	0.47	88
macro avg	LLDs	0.82	0.54	0.54	285
	MFCC	0.42	0.42	0.40	261
weighted avg	LLDs	0.75	0.73	0.70	285
	MFCC	0.56	0.62	0.57	261
accuracy	LLDs	73%			
	MFCC	62%			

Table 3.4: Classification Report of LR

Table 3.5 summarizes the accuracies obtained when applying the earner Based Feature Selection and as shown the best accuracy got with MLP in LLDs features(84%), and the high accuracy for MFCC features got with KNN (69%):

Features / Model	MLP	SVM	KNN	LR
LLDs	84%	76%	81%	73%
MFCC	64%	64 %	69%	62%

Table 3.5: Test Accuracy by Model with WEKA features selection

3.3.3 Results of Rough set theory

The same models applied on earner Based Feature Selection are applied with some changes in parameters Table 3.6, Table 3.7, Table 3.8 and Table 3.9 illustrate each model's classification report (MLP, SVM, KNN and LR).

Table 3.10 summarizes the accuracies obtained when applying the models with Rough set features selection respectively.

Figure 3.5 show the results of confusion matrices of the models when applying Rouf ser theory.

		precision	recall	f1-score	support
Surprised	LLDs	0.54	0.54	0.54	28
	MFCC	0.62	0.48	0.54	27
angry	LLDs	0.97	0.89	0.93	154
	MFCC	0.87	0.89	0.87	146
happy	LLDs	0.84	0.94	0.89	103
	MFCC	0.81	0.83	0.82	88
macro avg	LLDs	0.78	0.79	0.78	285
	MFCC	0.77	0.73	0.75	261
weighted avg	LLDs	0.88	0.87	0.87	285
	MFCC	0.82	0.83	0.82	261
accuracy	LLDs	87%			
	MFCC	83%			

Table 3.6: Classification Report of MLP

		precision	recall	f1-score	support
Surprised	LLDs	0.00	0.00	0.00	28
	MFCC	0.00	0.00	0.00	27
angry	LLDs	0.87	0.94	0.90	154
	MFCC	0.61	0.94	0.74	146
happy	LLDs	0.83	0.92	0.87	103
	MFCC	0.75	0.31	0.44	88
macro avg	LLDs	0.56	0.62	0.59	285
	MFCC	0.45	0.42	0.39	261
weighted avg	LLDs	0.77	0.84	0.80	285
	MFCC	0.59	0.63	0.56	261
accuracy	LLDs	84%			
	MFCC	63%			

Table 3.7: Classification Report of SVM

		precision	recall	f1-score	support
Surprised	LLDs	0.61	0.39	0.48	28
	MFCC	0.38	0.30	0.33	27
angry	LLDs	0.87	0.94	0.90	154
	MFCC	0.76	0.87	0.81	146
happy	LLDs	0.83	0.82	0.82	103
	MFCC	0.77	0.64	0.70	88
macro avg	LLDs	0.77	0.71	0.73	285
	MFCC	0.64	0.60	0.61	261
weighted avg	LLDs	0.83	0.84	0.83	285
	MFCC	0.72	0.73	0.72	261
accuracy	LLDs	84%			
	MFCC	73%			

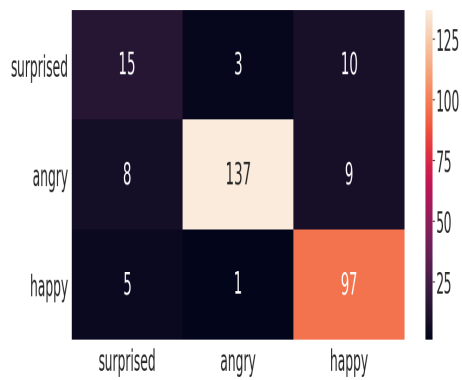
Table 3.8: Classification Report of KNN

		precision	recall	f1-score	support
Surprised	LLDs	0.00	0.00	0.00	28
	MFCC	0.00	0.00	0.00	27
angry	LLDs	0.84	0.94	0.88	154
	MFCC	0.62	0.91	0.74	146
happy	LLDs	0.85	0.91	0.88	103
	MFCC	0.71	0.39	0.50	88
macro avg	LLDs	0.56	0.62	0.59	285
	MFCC	0.44	0.43	0.41	261
weighted avg	LLDs	0.76	0.84	0.79	285
	MFCC	0.59	0.64	0.58	261
accuracy	LLDs	84%			
	MFCC	64%			

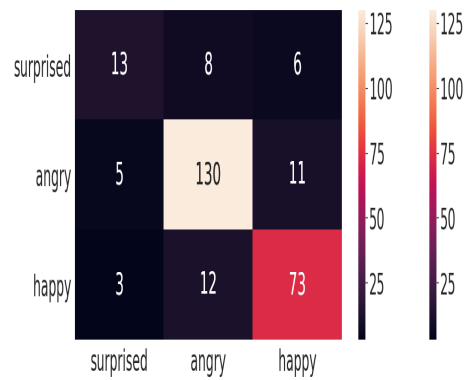
Table 3.9: Classification Report of LR

Features / Model	MLP	SVM	KNN	LR
LLDs	87%	84%	84%	84%
MFCC	83%	63%	73%	64%

Table 3.10: Test Accuracy by Model with Rough set features selection



(a) Confusion matrix of LLDs_MLP



(b) Confusion matrix of MFCC_MLP



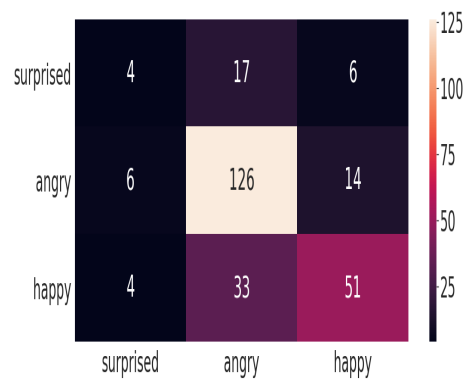
(c) Confusion matrix of LLDs_SVM



(d) Confusion matrix of MFCC_SVM



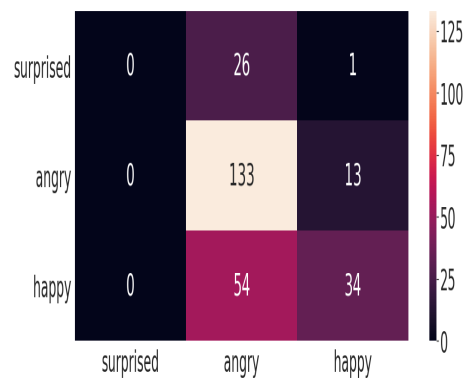
(e) Confusion matrix of LLDs_KNN



(f) Confusion matrix of MFCC_KNN



(g) Confusion matrix of LLDs_LR



(h) Confusion matrix of MFCC_LR

Figure 3.5: Results

3.4 Evaluation

We find a remarkable improvement in Rough set's results by comparing the results obtained with earner Based Feature Selection. Thus, the feature's selection of Rough set theory provided a better performance From its counterpart earner Based Feature Selection. We note the superiority Multilayer perceptron (MLP) over the rest models on Rough set features selection for the applied models. So that the results were as follows: 87% with LLDs features and 83% with MFCC features.

Figure 3.6 illustrates The average performance accuracy of the MLP model trained and validation set on the all-inclusive, with LLDs, MFCC features on 700 training epochs.

Figure 3.7 shows the Normalized confusion matrix of (LLDs_MLP), so we can notice

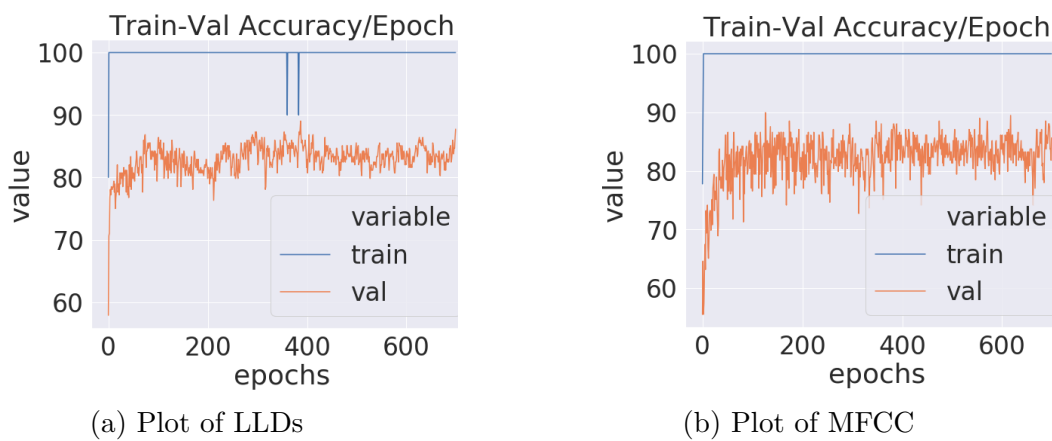


Figure 3.6: Train-Val Accuracy/Epoch plot

that the highest value is angry-angry with 96.42%, followed by happy-happy with 84.07% and the latest one is surprised-surprised with 53.33%.

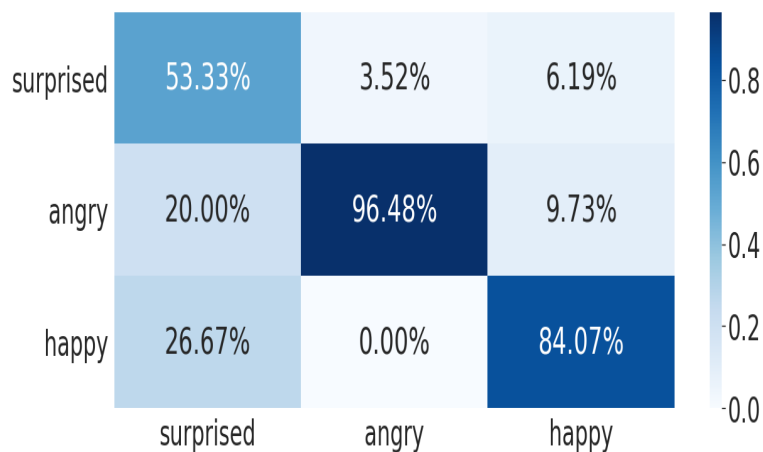


Figure 3.7: Normalized confusion matrix of (LLDs_MLP)

Figure 3.8 shows the Normalized confusion matrices of (MFCC_MLP), so we can notice that the highest value is angry-angry with 86.67%, followed by happy-happy with 81.11% and the latest one surprised-surprised is better than surprised-surprised of LLDs_MLP with 61.90%.

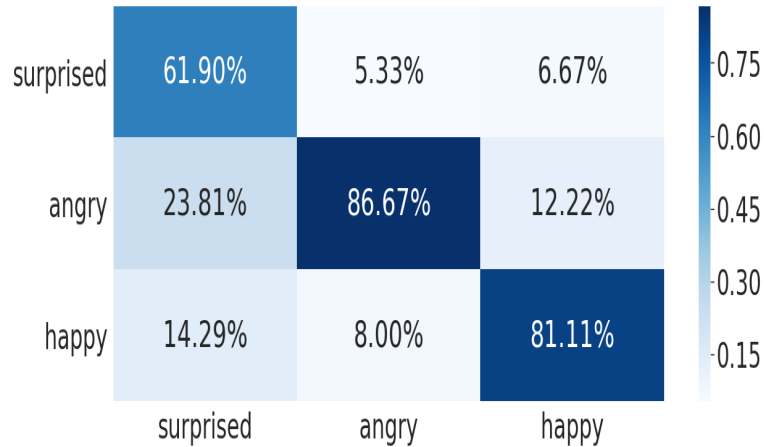


Figure 3.8: Normalized confusion matrix of (MFCC_MLP)

3.4.1 Results analysis

Experimentation results are discussing in this section when the experimental evaluation is performed on the LLDs and MFCC features. Because of the specialization of LLDs in emotions, the perpetual dominance of LLDs features over MFCC features was observed in outcomes. It contains different standard feature sets for emotion recognition, where emotions were extracted using the 'emobase' package, which is the most modern and huge.

For classification techniques, there is no conclusive response to the preference of one over the other for classification techniques; every model has its advantages and limitations. Here, we decided to compare the output of each one for this purpose. We were applying cross-validation for performance analysis when the model was trained. When tested the models, we got good precision when implemented MLP on LLDs and MFCC .

Given the results of the Dataset's confusion matrices, we find a varying difference in accuracy between emotions, with some being easier to recognize than others, where the emotion of surprise was the Less precision. Which can be due to the following reason: relative to the other two groups (angry, happy), the number of utterances for surprised speech is smaller.

3.5 Conclusion

In this part which bearing of Learner Based Feature Selection and Rough set theory, we relied on presenting the tools used in coding in terms of environment, libraries, preparing the dataset, and presenting the results of each model through the Confusion Matrix and Classification Report. Indeed, After comparing the scores, we got a noticeable improvement in the results in Rough set theory, topped by the Multilayer perceptron in both features: 87% for LLDs and 83% for MFCC. Worth noting is the features of LLDs was better than MFCC features. However, this does not mean MLP always best because none of the models is better than the other. One's superior performance is often credited to the nature of the data being worked on.

General Conclusion and Future Work

The field of speech processing is still late compared to image processing, but this does not mean superiority in the results, because there are speech studies that have good results outperformed image studies. In this study, we focused on recognizing the human emotion through Speech in the area of the Arab world, intending to make the interaction between a person and a machine more suitable than before and improving many functions. The more knowledge of a machine is about human characteristics, the easier it is for the machine to respond to human requirements. This of course includes clear and unclear specifications.

Our study of the Arabic language was devoted to the scarcity we witness in this field, as well as the availability of a few databases whether it is simulated, elicited, or natural. We chose the Lebanese database considering that the nature database is closer to reality, then we moved on to extract two types of features from the speech signal using the OpenSMILE program, the first type is the LLDs features where we obtained 988 features; the second type is MFCC features which are Frequently popular in the field of sound processing, we got 39 features. We applied an optimization process to features to improve results due to the presence of many unrelated features and containing information that might hinder recognition rates by Learner Based Feature Selection. We obtained 15, 7 features for LLDs and MFCC respectively. The result of the classification shows that the Multilayer perceptron performed better with LLDs features (84%), While the k-Nearest Neighbor(KNN) performed better with MFCC features (69%).

In order to improve our results, we proposed to apply Rough set theory for features selection and compare it with the results obtained with Learner Based Feature Selection, we got 11, 13 features of LLDs, MFCC features respectively.

Indeed, we got a noticeable improvement in the results, topped by the Multilayer perceptron in both features: 87% for LLDs and 83% for MFCC, Worth noting is the features of LLDs was the better than MFCC features. The results obtained are satisfactory, in view of the field in which we are operating, and given our first use of the OpenSMILE system, the Rough set theory and the PyTorch library, as well as the existence of a few studies relevant to our subject, in particular Arabic. The research is not fully thorough, because we focused on three forms of emotions and at the same time it is one of the works related to the Arabic language and the speech signal in particular. Can be regarded as the beginning of a deeper analysis of these studies in our Arab world.

In future studies, we aim to build an Arab database that includes numerous emotion; work to improve audio features and focus on sound processing and use new technologies. For classification, we aim to use different and more robust algorithms in the field of deep learning.

Bibliography

- [1] ABDEL-HAMID, L. Egyptian arabic speech emotion recognition using prosodic, spectral, and wavelet features. *Speech Communication* (04 2020).
- [2] ABHAY KUMAR . Multivariate Multilabel Classification with Logistic Regression. <https://acadgild.com/blog/logistic-regression-multiclass-classification>. [Online; accessed 18-August-2020].
- [3] AFFECTIVA. Introducing Affectiva’s Emotion Recognition through Speech. <https://blog.affectiva.com/introducing-affectivas-emotion-recognition-through-speech/>. [Online; accessed 17-March-2020].
- [4] AHLAN WORLD. Modern standard Arabic and colloquial Arabic, which to choose? <https://www.ahlan-world.org/modern-standard-arabic-colloquial-arabic/>. [Online; accessed 28-March-2020].
- [5] AKÇAY, M. B., AND OĞUZ, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.* 116, October 2019 (2020), 56–76.
- [6] AKSHAJ VERMA . PyTorch [Tabular] Multi-class Classification. <https://towardsdatascience.com/pytorch-tabular-multiclass-classification-9f8211a123ab>. [Online; accessed 10-June-2020].
- [7] ALBERT ALISALAH,†HEYSEM KAYA,FURKAN GÜRPINAR. Chapter 17 - Video-based emotion recognition in the wild. <https://www.sciencedirect.com/science/article/pii/B9780128146019000316>. [Online; accessed 8-July-2020].
- [8] ASHA:AMERICAN SPEECH-LANGUAGE-HEARING ASSOCIATION. What Is Speech?

- What Is Language? https://www.asha.org/public/speech/development/language_speech. [Online; accessed 28-March-2020].
- [9] AVINASH NAVLANI. KNN Classification using Scikit-learn. <https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn>. [Online; accessed 18-August-2020].
- [10] AVINASH NAVLANI. Support Vector Machines with Scikit-learn. <https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python>. [Online; accessed 12-August-2020].
- [11] AVINASH NAVLANI. Understanding Logistic Regression in Python. <https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python>. [Online; accessed 13-August-2020].
- [12] BRAD PATTON. Anaconda Python Tutorial. <https://linuxhint.com/anaconda-python-tutorial/>. [Online; accessed 9-June-2020].
- [13] CHEATSHEET. Logistic Regression. https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html#multiclass-logistic-regression. [Online; accessed 13-August-2020].
- [14] CHRIS NICHOLSON. A Beginner's Guide to Multilayer Perceptrons (MLP). <https://pathmind.com/wiki/multilayer-perceptron>. [Online; accessed 2-June-2020].
- [15] DAADACADEMY. خصائص اللّغة العربيّة. <http://daadacademy.org/arabic-properties//>. [Online; accessed 22-March-2020].
- [16] DATAFLAIR TEAM. Python Mini Project - Speech Emotion Recognition with librosa - DataFlair. <https://data-flair.training/blogs/python-mini-project-speech-emotion-recognition/>. [Online; accessed 16-February-2020].
- [17] DEEPLIZARD. CNN Confusion Matrix with PyTorch - Neural Network Programming . <https://deeplizard.com/learn/video/0LhiS6yu2qQ>. [Online; accessed 16-June-2020].
- [18] DIPL, F., AND VOGT, T. Real-time automatic emotion recognition from speech.
- [19] DR. MICHAEL J. GARBADE FEED. How to use the Scikit-learn Python library for data science projects. <https://opensource.com/article/18/9/>

- [how-use-scikit-learn-data-science-projects](#). [Online; accessed 11-August-2020].
- [20] EYBEN, F., WÖLLMER, M., SCHULLER, B. B., WENINGER, F., WOLLMER, M., AND SCHULLER, B. B. OPENSIMILE: open-Source Media Interpretation by Large feature-space Extraction. *MM'10 - Proc. ACM Multimed. 2010 Int. Conf.*, December (2015), 1–65.
- [21] FIROZ SHAH, A. Study and analysis of speech emotion recognition. 13–23.
- [22] FRANTI, E., ISPAS, I., DRAGOMIR, V., DASC, M., ALU, ZOLTAN, E., AND STOICA, I. C. Voice Based Emotion Recognition with Convolutional Neural Networks for Companion Robots. *Rom. J. Inf. Sci. Technol.* 20, 3 (2017), 222–240.
- [23] GEEKSFORGEES . Python Language Introduction. <https://www.geeksforgeeks.org/python-language-introduction/>. [Online; accessed 9-June-2020].
- [24] HORKOUS, H., AND GUERTI, M. Recognition of emotions in the Algerian Dialect Speech. 1–10.
- [25] JASON BROWNLEE. How to Perform Feature Selection With Machine Learning Data in Weka . <https://machinelearningmastery.com/perform-feature-selection-machine-learning-data-weka/>. [Online; accessed 12-July-2020].
- [26] JAVATPOINT. K-Nearest Neighbor(KNN) Algorithm for Machine Learning. <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning>. [Online; accessed 12-August-2020].
- [27] JEFF HALE. Scale, Standardize, or Normalize with Scikit-Learn. <https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02>. [Online; accessed 10-June-2020].
- [28] JEREMY JORDAN. Normalizing your data (specifically, input and batch normalization). <https://www.jeremyjordan.me/batch-normalization/>. [Online; accessed 10-June-2020].
- [29] JHALA, A. Emotions. https://www.researchgate.net/publication/320621079_{_}Emotions_{_}Psychology. [Online; accessed 25-March-2020].
- [30] KENDRA CHERRY. Emotions and Types of Emotional Responses . <https://www.verywellmind.com/what-are-emotions-2795178>. [Online; accessed 24-

- March-2020].
- [31] KENNY MIYASATO. Classification Report: Precision, Recall, F1-Score, Accuracy. <https://medium.com/@kennymiyasato/classification-report-precision-recall-f1-score-accuracy-16a245a437a5>. [Online; accessed 18-June-2020].
- [32] KLAYLAT, S., OSMAN, Z., HAMANDI, L., AND ZANTOUT, R. Enhancement of an Arabic Speech Emotion Recognition System. *International Journal of Applied Engineering Research* 13, 5 (2018), 2380–2389.
- [33] KSTACKEXCHANGE. Micro Average vs Macro average Performance in a Multiclass classification setting. <https://datascience.stackexchange.com/questions/15989/micro-average-vs-macro-average-performance-in-a-multiclass-classification-setting>. [Online; accessed 18-June-2020].
- [34] LARRY BLASER. Speech. <https://www.encyclopedia.com/literature-and-arts/language-linguistics-and-literary-terms/language-and-linguistics/speech>. [Online; accessed 28-March-2020].
- [35] LECH, M., STOLAR, M., BEST, C., AND BOLIA, R. Real-Time Speech Emotion Recognition Using a Pre-trained Image Classification Network: Effects of Bandwidth Reduction and Companding. *Front. Comput. Sci.* 2 (2020), 14.
- [36] LEILA KERKENI, YOUSSEF SERRESTOU, MOHAMED MBARKI, KOSAI RAOOF, MOHAMED ALI MAHJOUR AND CATHERINE CLEDER. Automatic Speech Emotion Recognition Using Machine Learning. <https://www.intechopen.com/books/social-media-and-machine-learning/automatic-speech-emotion-recognition-using-machine-learning/>. [Online; accessed 17-March-2020].
- [37] LIAM HURST. Keras vs TensorFlow vs PyTorch: What are the differences? <https://morioh.com/p/a78f0b9f3e60>. [Online; accessed 8-June-2020].
- [38] LIN, T. Y., CERCONE, N., AND PAWLAK, Z. Rough Sets. *Rough Sets and Data Mining* (1997), 3–7.
- [39] MADISON SCHOTT. K-Nearest Neighbors (KNN) Algorithm for Machine Learning. <https://medium.com/capital-one-tech/k-nearest-neighbors-knn-algorithm-for-machine-learning-e883219c8f26>.

- [Online; accessed 12-August-2020].
- [40] MEFTAH, A. H., QAMHAN, M., ALOTAIBI, Y., AND SELOUANI, S. Emotional speech recognition using rhythm metrics and a new arabic corpus. In *2020 16th IEEE International Colloquium on Signal Processing Its Applications (CSPA)* (2020), pp. 57–62.
- [41] MIKE DRISCOLL . Jupyter Notebook: An Introduction. <https://realpython.com/jupyter-notebook-introduction/>. [Online; accessed 9-June-2020].
- [42] MUTHU KRISHNAN. Understanding the Classification report through sklearn. <https://muthu.co/understanding-the-classification-report-in-sklearn/>. [Online; accessed 18-June-2020].
- [43] NITIN KUMAR KAIN. Understanding of Multilayer perceptron (MLP). https://medium.com/@AI_with_Kain/understanding-of-multilayer-perceptron-mlp-8f179c4a135f. [Online; accessed 19-August-2020].
- [44] OF ENCYCLOPAEDIA BRITANNICA, T. E. Arabic language. <https://www.britannica.com/topic/Arabic-language>. [Online; accessed 28-March-2020].
- [45] PANDEY, S. K., SHEKHAWAT, H. S., AND PRASANNA, S. R. Deep learning techniques for speech emotion recognition: A review. *2019 29th Int. Conf. Radioelektronika, RADIOELEKTRONIKA 2019 - Microw. Radio Electron. Week, MAREW 2019*, July (2019), 1–6.
- [46] PERTILA, P. Gammatone filter banks & Mel-frequency cepstral coefficients Introduction. *Tut* (2015).
- [47] PIOTR MIGDAL AND RAFAL JAKUBANIS . Keras or PyTorch as your first deep learning framework. <https://deepsense.ai/keras-or-pytorch/>. [Online; accessed 8-June-2020].
- [48] PRACTICALCRYPTOGRAPHY. Mel Frequency Cepstral Coefficient (MFCC) tutorial. <http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/#deltas-and-delta-deltas>. [Online; accessed 3-June-2020].
- [49] PYTORCH. Classification report output. <https://discuss.pytorch.org/t/classification-report-output/31222>. [Online; accessed 18-June-2020].
- [50] PYTORCH. Quick Start Locally. <https://pytorch.org/>. [Online; accessed 9-June-

- 2020].
- [51] ROSE, P., AND RABENSTEIN, R. CHAPTER-2: MECHANISM OF SPEECH PRODUCTION AND LITERATURE REVIEW 2.1. Anatomy of speech production. 5–17.
- [52] SCIKIT-LEARN.ORG. sklearn.metrics.classification-report. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html. [Online; accessed 18-June-2020].
- [53] SHAHIN, I. Emotion recognition using speaker cues, 2020.
- [54] SHAHSAVARANI, S. Speech Emotion Recognition using Convolutional Neural Networks (Master thesis, The University of Nebraska-Lincoln). *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 10633 LNAI* (2018), 87.
- [55] SHANELYNN. Python Pandas read-csv Load Data from CSV Files. https://www.shanelynn.ie/python-pandas-read_csv-load-data-from-csv-files/. [Online; accessed 10-June-2020].
- [56] SHRAWANKAR, U., AND THAKARE, V. M. Techniques for feature extraction in speech recognition system : A comparative study.
- [57] SUNIL RAY. Understanding Support Vector Machine(SVM) algorithm from examples (along with code) . <https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/>. [Online; accessed 12-August-2020].
- [58] SWAIN, M., ROUTRAY, A., AND KABISATPATHY, P. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology* 21 (01 2018).
- [59] TIAGO M. LEITE. Neural Networks, Multilayer Perceptron and the Backpropagation Algorithm. <https://medium.com/@tiago.tmleite/neural-networks-multilayer-perceptron-and-the-backpropagation-algorithm>. [Online; accessed 3-June-2020].
- [60] USMAN MALIK . Implementing SVM and Kernel SVM with Python's Scikit-Learn . <https://stackabuse.com/implementing-svm-and-kernel-svm-with-pythons-scikit-learn/>. [Online; accessed 11-August-2020].

-
- [61] UTANE, A., AND NALBALWAR, S. Emotion recognition through Speech. *International Journal of Applied Information Systems*, Ncipet (2013), 8.
- [62] WEKA. WEKA The workbench for machine learnin. <https://www.cs.waikato.ac.nz/ml/weka/>. [Online; accessed 12-July-2020].
- [63] WIKIPEDIA. Speech processing. https://en.wikipedia.org/wiki/Speech_processing. [Online; accessed 20-August-2020].