People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research

University KASDI Merbeh-Ouargla
Faculty of New Information Technologies and Communication
Department of Computer Science and Information Technology

Thesis
In view of obtaining the diploma of
MASTER ACADEMIC
Topic :

Arabic Speech Emotion Recognition Building an Arabic Emotion Speech Data set.

Supervisor :
Dr.Mourad Belhadj

Before the jury :
Dr.Abdelhakim Chriet
Dr.Kadidja Amer

Presented by :
Elalia Lakhdari
Ilham Bendellali

University year : 2021/2022

Abstract

This project aims to build and improve the Arabic emotional speech data set for speech emotion recognition systems. The study went through two main phases: Collecting audio data and subjecting it to post-processing Then verify their validity and reliability based on human perception To make the database accessible to the next stage of building a complete recognition system.

The output of this work is a data set of KASDI-MERBAH University for emotional speech in Arabic, which includes 5000 audio recording files divided equally into each of the emotions of anger, happiness, sadness, fear, and neutrality, and a total of 10 sentences among the most used in communication. Five hundred representatives from the university community participated in recording the data(student, professors and staff) and 56 others evaluated the validity of identification and reliability, including specialists in clinical psychology.

Where did the evaluation give the following results: 75.08% of Accuracy of identifying emotions for the database, where the data that included neutrality took the highest evaluation rate estimated at 87.7%, followed by happiness, anger, fear, and sadness with the following evaluation results, respectively (81.2% ,79.7 %65.8 % , 61.1%) And with a moderate reliability rate of 53.4% in the evaluation results of the category and emotion's severity

.

Keywords: speech Emotion recognition , Arabic language, audio , emotional speech database.

ملخص

الهدف من هذا المشروع هو بناء وتحسين مجموعة بيانات للكلام العاطفي العربي موجهة لانظمة التعرف على المشاعر الكلامية. مرت الدراسة بمرحلتين رئيسيتين: جمع البيانات الصوتية وإخضاعها للمعالجة اللاحقة ثم التحقق من صلاحيتها وموثوقيتها بناءً على الإدراك البشري لجعل قاعدة البيانات قابلة للوصول إلى المرحلة التالية من بناء نظام تعرف كامل.

كان ناتج هذا العمل بناء مجموعة بيانات جامعة قاصدي مرباح للكلام العاطفي باللغة العربية والتي تضم 5000 ملف تسجيل صوتي مقسمة بالتساوي في كل من عاطفة الغضب السعادة الحزن الخوف والحياد وباجمالي 10 جمل من بين الأكثر إستخداما في التواصل ضمن اللغة.

شارك في تسجيل البيانات 500 جهة فاعلة من المجتمع الجامعي لقاصدي مرباح (أساتذة ,طلبة وموظفين) وقام 56 آخرون بتقييم صحة التعرف والموثوقية من بينهم متخصصين بعلم النفس العيادي.

أعطى التقييم النتائج التالية : نسبة 75.08% لصحة التعرف على عاطفة قاعدة البينات حيث أخذت البيانات المتضمنة للحياد أعلى نسبة تقييم تقدر ب 87.7% يليها على التوالي السعادة ,الغضب , الخوف ,والحزن بنتائج التقييم التالية (81.2%,79.7%,65.8%,61.1%) وبنسبة موثوقية تقدر ب 53.4% لفئة العاطفة .

كلمات مفتاحية : انظمة التعرف على المشاعر في الكلام ,اللغة العربية قاعدة بيانات الكلام العاطفي

Acknowledgments

First of all, we would like to thank Almighty ALLAH for helping us do this humble work.
We extend our sincere thanks to our supervisor, **Dr. Mourad Belhaj**, Assistant Professor A at the Kasdi Merbah University of Ouargla, who spared no effort in making this thesis possible. We are grateful to him for his guidance, encouragement, and above all, for helping us complete this letter.

We would like to warmly thank **Dr. Khadija Amer**, Head of the Department of Computer Science at the University of Kasdi Merbah, Ouargla, who encouraged us. Furthermore, we would like to express our deep gratitude to her for the interest and advice she gave us despite her multiple responsibilities.

We also extend our sincere thanks to all the jury members for the honor they gave us by accepting the judgment on our work, as well as to all the professors of the university course who contributed to our training.

We also extend our complete gratitude and thanks to **Mis. DALAL Djeridi** for all the support she provided us in all stages of the work.

Finally, it would be challenging to fail to thank all those who contributed directly or indirectly to this work and who, we hope, will find in these few lines an expression of our sincere thanks.

Dedication

“

After conciliation from God Almighty

before everyone, Thank you my master, teacher, and messenger, Muhammad (may God bless him and grant him peace),Because you taught us the meaning of diligence and the meaning of being in search of a goal.
It is your instruction before it is my graduation thesis.

A special feeling of gratitude to my father and mother ,this work and All I am today It was because of you I dedicate it to you because it is the fruit of your upbringing

For all those who made my life better without exception
Thanks for your fingerprint.

”

ilham

“

For all of my little family.

To my mom in particular because she gave me life and all the support.

Your good teaching, advice, and the blessings of your supplications never fail; May Allah rewards you well on our behalf.

For their love, understanding, and patience, I would like to dedicate this humble deed to you to all my dear sisters.

”

- Elalia

Contents

List of Figures

List of Tables

List of abbreviations and acronyms

SER	*Speech Emotion Recognition*
KEDAS	*Kasdi-merbah Emotional Database in arabic Speech*
BAVED	*Basic Arabic Vocal Emotions Dataset*
EYASE	*Egyptian Arabic Speech Emotion*
ADED	*Algerian Dialect Emotion Database*
MFCC	*Mel Frequency Cepstral Coefficient*
LLDs	*Low Level Descriptors*
SVM	*Support Vector Machine*
HMM	*Hidden Markov Model*

General introduction

Context:

Machine learning is a required field branching off from artificial intelligence, which is a technology that enables building systems based on data analysis for learning and performance improvement without being explicitly programmed. The performance of these systems includes any possible approach between human intelligence to machine intelligence, including prediction, inference, and classification. Speech emotion recognition systems have also been among the active areas of studies where human speech processing can provide information such as language and dialect, speaker gender, age, health status, and many other important characteristics about the speaker in creating appropriate interactive systems. We mention the speaker's emotional state among the essential characteristics because human reactions and communication are mainly based on knowing how the other party feels in different languages. Expression of emotion corresponds to changes in body language and the quality of ideas expressed, such as textual vocabulary that can be written down, facial features or tone, and intensity of voice using hands eye movement.

As the discursive emotion recognition systems represent all techniques and methods that process and analyze the speech signal to classify the feelings involved, which encouraged great interest in this approach and the emergence of many applications for it in the field, such as recommendation algorithms adopted and developed by Netflix, Amazon and Facebook, fraud detection applications And self-driving cars

- The idea of developing a system to identify the emotion of speech stems from building a database within the appropriate standards as the main factor that determines the quality of performance, taking into account the number, type of feelings included, the context of

the database, and the Category.

-The second step: is reprocessing the data before extracting the signal features because the speech samples contain irrelevant information such as noise and silence; Where are some of the environmental differences removed at this stage to produce pure samples of a quality that allows their approval.

Motivation:

Among the objectives of this research is to achieve a scientific addition to the Arabic speech emotion recognition systems, which, like any other recognition system, depends on the availability of the raw material of phonetic data in the language.
Looking at previous studies aimed at the same goal, we find a significant limitation and deficiency in the available databases in the Arabic emotional discourse, and this has resulted in a gap in the development of recognition systems and studies directed toward the Arabic language compared to parallel research in other global languages such as English and Chinese, which have exceeded their phonetic data 3000 hours of recording, although Arabic is ranked among the ten most spoken languages in the world.

Contribution:

This work adds two main contributions represented in a study of the main works in the databases of Arabic discourse available on the scene In addition to establishing a new database for speech emotion recognition systems to bridge the gap in the lack of databases available and not exploit the Arabic language as a resource in modern technical studies.

Description of work:

This thesis is organized as follows: We will start with a general introduction where we will present the field of study, the topic of research, and the objectives of the work The first chapter deals with the background of the information related to the speech signal,

its applications and techniques, and the concepts related to the Arabic language. At the end of this chapter, we address the current emotion recognition systems.

The second chapter also aims to explain the systems for recognizing discursive emotion and its structure, where we will discuss three main axes represented in studying the theories of emotional modeling adopted in the system, database standards, and the scope of work on them with a review of the essential studies contained within the Arabic language systems in addition to addressing the techniques for extracting features and mechanisms Category. During the third chapter, a model was prepared to design a new audio database for the Algerian Standard Arabic, where the mechanism of its implementation, evaluation, processing, and improving the quality of the data obtained is explained and clarified, in addition to discussing the final results of the submitted work. Finally, we conclude our thesis with a general conclusion, outlining the prospects for future work.

Chapter 1

Background information

1.1 Introduction

The interest in developing interactive systems as a means of facilitating human automated communication has increased recently, and since spoken speech in various languages of the world, including Arabic, is the most natural and common way of communication between humans to express ideas, Voice technology has made parallel advances in research, including the studies, on systems for speech recognition, speaker recognition, or systems that detect the emotions of spoken speech. From this point of view, this chapter will discuss the nature of the Arabic language, the nature and characteristics of the audio signal, its processing techniques, and the most important techniques and systems for recognizing emotions from speech.

1.2 Speech

It is the oral expression used to convey ideas, and it is the spoken aspect of communication language. It is represented in the information produced by the human speech system, which is transmitted by sound waves through space from the speaker to the listener.(Akçay 2020) In general, speech consists of three essential parts: sound, pronunciation, and fluency.

- **The voice:** is how the vocal folds are used and the breathing process to produce sound levels of different intensity and varying pitches between soft and sharp.

- **Pronunciation:** Producing sounds using the pronunciation device, such as the tongue, larynx, lips, etc.

- **Fluency:**It is the smoothness in the rhythm of speech, which can be determined by the lack of stuttering and unwanted pauses during speech or the repetition of sounds without necessity.(FNC 2013)

1.2.1 Speech signal properties

According to Schetelig T. and Rabenstein R. in May 1998, a speech signal is a multidimensional sound wave (shown in Figure 2.3) that provides information about the words or message being spoken, the identity of the speaker, spoken language, and physical and mental health The race, age, gender, education level, religious orientation and background of the individual (**Mourad**), here are some of the essential characteristics of an audio signal, Which is illustrated in Figure 1.1 :

- **Amplitude** represents the maximum displacement taken between the two ends of the cycle.necessity.

- **Cycle** every audio signal is trans-versed in the form of cycles. One complete upward movement and downward movement of the signal from a cycle.

- **Wavelength** The wavelength is the length of one complete cycle of the wave.

- **Speed** In the case of a sound wave, speed is the distance traveled from a point on the wave in a certain period of time.

- **-Frequency** frequency refers to how fast a signal changes over a period of time (FNC 2013).

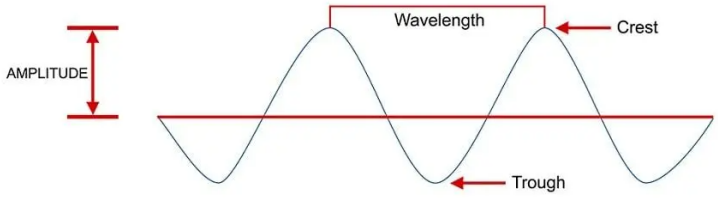

Figure 1.1: Representation of wave properties. (Maryam 2020).

1.2.2 Signal representation

The natural sound perceived by human hearing is a continuous analogue signal, which is not suitable for working on it and storing it in the digital environment. Therefore, the continuous signal must be converted into separate samples at specific time intervals in the form of a matrix of values that can be dealt with by recognition systems and learning algorithms(Smales 2019) . Figure 1.2 shows the representation of the signal from analog to digital.

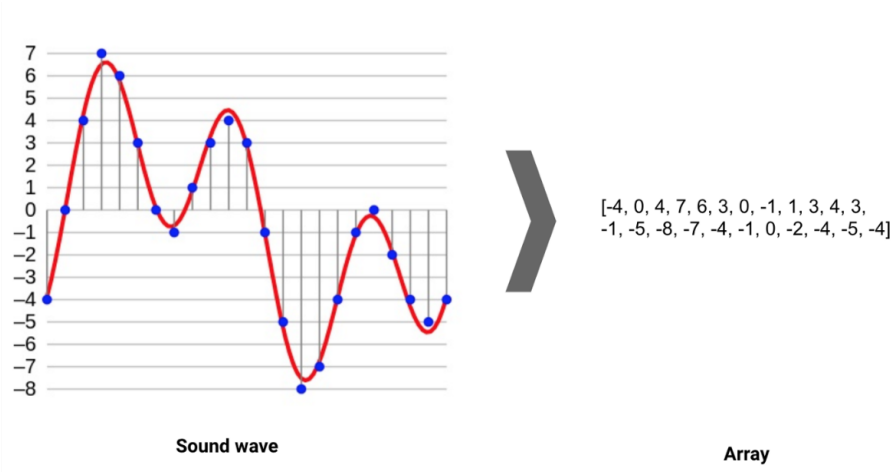

Figure 1.2: Digital representation of sound signal (Smales 2019).

1.2.3 Audio signal processing

It is one of the techniques for improving the audio signal, which makes deliberate changes to it to extract the required features and vital information without others, so it is enhanced to become suitable for adoption in recognition systems. We mention the following (Wani 2021):

Noise Reduction

The speech signal contains the speech of the target speaker, background noise, the voices of non-target speakers, and an echo of the voice of silence, which is a form of noise.

In general, noise is considered an unwanted signal.
It appears as a result of some errors in the recording process, the transmission of the signal or the environment from which it is taken, such as the pulse noise that occurs due to scratches and the great sensitivity to the captured sounds, which makes it unsuitable for use by recognition systems.
Therefore, noise removal is among the most critical processing techniques implemented on the audio signal. By detecting and identifying hidden interference between the applicable original audio samples and removing it(Wani 2021).

Speech segmentation (framing)

Speech signals continue to change and are unstable over time, but it is possible to capture a stable signal for a brief period of milliseconds (20 milliseconds, for example)
Therefore, framing, or splitting the original speech signal into subsections, will allow the extraction of quasi-static local characteristics for each frame(Wani 2021).

As it dramatically affects the performance of a classification system, a millisecond difference in frame size will give a different classification rate.(Ozseven 2018)
There are two types of automatic segmentation algorithms:
The first category considers the knowledge of the linguistic material of speech externally. The other class does not require it and depends on acoustic properties and signal analysis using MFCC or FFT coefficients, fast Fourier transform. These algorithms are often used in speech recognition systems or speech emotion recognition systems(Sakran 2017).

Signal windowing

Since the data is infinite and subject to lose and leakage due to discontinuities at the edge of the signals in the framing phase, the windowing process will reduce this resulting spectral leakage(Wani 2021).

Voice activity detection

It is difficult to accurately detect the limits of target speech in a dynamic environment in the presence of other unwanted sounds such as silence and noise.
The voice activity detection technology works on this, as it detects the presence or absence of human speech in the signal by determining the threshold or maximum limits that the speech signal reaches, which is greater compared to the background noise signal.
Thus the audio Activity can be detected at the limits of the maximum values of the signal, and it is recalculated Activity level for each window on this basis. Bearing in mind that silence also appears in the signal with a low representation, such as noise(Tanyer 2000).

Normalisation

is a necessary step used to reduce the contrast in audio features and adjust its volume to a standard level. The most common method of normalisation is z-score normalisation.

$$z = \frac{x - \mu}{\sigma} \tag{1.1}$$

x: is the value to be normalized,z: normalization value, μ: is the mean and σ: is the mean (Ayadi 2011).

1.3 Emotion recognition

With the massive trend in human-machine interaction (HMI) technology, emotion recognition systems have recently received much attention.
These systems represent automated techniques that allow identifying different emotional states and describing them by studying their expressions in physiological changes and behavioural reactions. It uses multimedia information such as audio, video, text, etc., as input, which it processes to identify and describe emotions. Among these models are :

1.3.1 Textual Emotion Detection

With the presence of the Internet, there is a massive amount of text data that We can treat to extract emotions such as emails, comments, and tweets on social networking

sites, articles and others. Recognizing emotions from text depends either on data-based methods or on machine learning classifiers.

Data-driven methods are based on discovering the most frequently occurring keywords in the textual material and then matching them with a pre-defined vocabulary set into the desired emotional categories, While machine learning methods are taking a broader form today, these systems classify text inputs into different feelings based on a pre-trained algorithm that applies different machine learning theories to determine the category of feelings (Shivhare 2012).

1.3.2 Recognize emotion from facial expressions

Gestures and facial expressions also carry much information about the person's emotional state, such as the eyes and eyebrows, and the movement of the upper section is more apparent in cases of anger and fear.In contrast, the movement of the mouth in the lower section of the face is more apparent in cases of happiness and disgust, and sudden suggestions and expressions can reach both of my sections of the face. These systems process digital images in the form of arrays with specific dimensions while training selected machine learning algorithms to classify facial expressions into a specific emotional category(Guo 2018) .

1.3.3 Recognize the emotion of multimedia

Multimodal emotion recognition systems are more complex than the ones we mentioned previously, which process one type of data and rely on additional support methods to reinforce the emotional information extracted.
It combines several linguistic, audio, and visual data features to extract and define a more accurate emotional classification(Najadat 2018) .

1.4 Language

Language is a normative system that characterizes the human race and is used to express ideas exchanged in daily communication between people.
The semantic interpretation of expressions and their meanings results from social and cultural influence, and their formation is subject to a specific construction methodology that controls the way of expression and communication(Fasold 2014).

1.4.1 Dialect

A dialect represents a sub-language sub-system derived from the original language; it is less normative, less formal and relates more complex to the narrow social environment of the dialect so that people do not understand languages that are different from their own while they can often understand other dialects derived from the same language(Fasold 2014) .

1.4.2 Arabic language

Arabic was ranked the sixth most widely spoken language globally for 2022, with more than 295 million speakers(**ethnologue**). It has several characteristics that make it more different and powerful compared to other languages, the most important of which are:

- The vocal system uses the tongue, larynx, throat and various parts of the articulation apparatus to articulate the sounds of letters.

- The Arabic linguistic lexicon has been ranked among the wealthiest dictionaries of the language, with a rate of one million words.

- Word formation is based on deriving additional vocabulary from a single source word.

- The construction of Arabic sentences is divided into two categories: nominal and phrasal verbs and each of them has special grammar rules that must be used in it, which makes the Arabic language a strong structure in terms of grammar, expressions and drafting methods. (**Mourad**)

Arabic can be divided into three main categories: Classical Arabic - Standard Modern Arabic, and Arabic dialects of the original language.

Classical Arabic

Classical Arabic represents the first language taken at the beginning of the Arab literary era, which was mentioned in the Holy Qur'an, and it was specialized in literary writings and poetic writings in an artistic and very developed way in terms of rhetoric(Ryding 2005).

Modern Standard Arabic (MSA)

Modern Standard Arabic represents the updates that the classical language underwent after the entry of Western writing methods and the spread of modern literary practices such as journalism and theatrical literature, which led to the classical language taking a more straightforward form in terms of style and vocabulary, making it innovative, lively and not highly codified like the classical language. This facilitated its integration as a language Informative and literary with a contemporary model that facilitates communication between sections of Arab speakers of all levels(Ryding 2005).

Arabic dialects

Arabic dialects, like other dialects, represent a mode of communication that is less normative than the classical language, which differs entirely from the formalities used in literature, academic works, and others. On the contrary, the Arabic dialects derive their roots from the original language, but they are affected by different culture-history and related to a specific geographical area and environment. Among the geographical divisions of the Arabic dialects, there are the distinctive Maghreb dialects of the countries of the Maghreb, with their differences, Egyptian, Levantine and Iraqi dialects, and the dialects of the Arab Gulf etc(Harrat 2016).

1.5 Conclusion

During this chapter, the background of concepts related to language and spoken language (the Arabic language in particular) and the nature of its use in the technical world in terms of representation of the speech signal and the basics of processing it automatically was presented, With Standing on the emotion recognition systems that scientific studies have touched upon to date, which relied on voice analysis in addition to analyzing textual data, facial features and image processing as parallel systems within the interactive systems approach. The second chapter will detail Arabic emotional speech recognition systems and a review of the research received about it.

Chapter 2

Speech Emotion Recognition System

2.1 Introduction

Advances in modern machine applications make it necessary to use speech emotion recognition systems in several human use areas.
SER is the study of methods that process speech and classify its signals to detect the emotions involved.
Since it is a vast field of research that keeps pace with the technical development of today's world, and as a field of study in this project, the second chapter will explain the main concepts related to the structure and development of SER systems and review previous studies that were conducted on the Arabic language in this field. with Referring to the models of emotional representation that are most relied upon in today's studies, databases, and Recognition mechanisms.

2.2 Speech emotion recognition system

Since the end of the 1950s, the study of emotions has been supported by modern technologies and has received wide interest in neuroscience, psychology, linguistics, and computer science. Proceeding from the fact that the speech signal is the most prominent and used way among humans to communicate with each other, this motivated researchers to extend this communication to humans and machines, where this machine could be (computers, telephones, robots).

On the other hand, this requires the machine to have enough intelligence to match the speaker's emotional state, which is influenced by many information such as language, gender, age, emotion, marital status, etc.

Despite the efforts of experts in this context, the field of research is still relatively young because the task of distinguishing emotional speech is very difficult for the following reasons:

first, because the same expression may appear with more than one emotion and because it is not clear which feature of speech is the strongest In the process of distinguishing between feelings.

Also, the great difference from one person to another in speaking styles, phrases used, rate of speech, etc., are among the obstacles that directly affect the majority of the extracted speech features, such as power and basic frequency.(Akçay 2020) Figure 2.1 shows the basic units of emotion recognition systems.

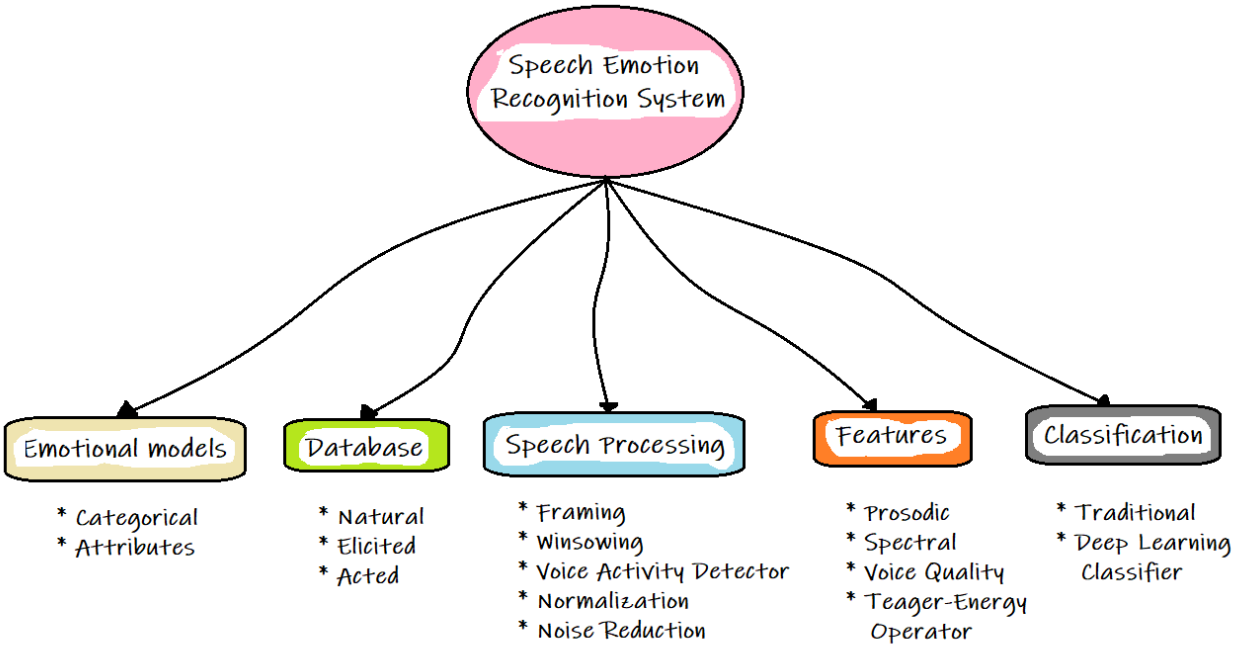

Figure 2.1: The structure of a speech emotion recognition systems.

2.3 Emotion Theories

Recognizing emotions in human speech is our primary focus in this work. For this, we must be able to control a certain number of emotions and identify the differences between them. However, this has been a door of difference and a large field of study in psychology as it is difficult to define each emotion separately from the others. Alternatively, describe it clearly and accurately, and this is because of the great diversity of human feelings and the presence of significant overlap between them. As a result, studies have not presented a single model that includes all emotions and allows them to be described and differentiated independently, as we will not be able to cover them all , but we can classify and differentiate them based on their similarities and interrelationships. The most common models are the discrete feelings model and the feelings model dimensions.

2.3.1 Discrete models of emotion

This theory defines a group of feelings that do not share characteristics. Although scientists differ in the interpretation of feelings, they tend to form a subgroup of independent feelings among themselves, called the group of pure or primary emotions, from which the rest of the other feelings are formed. The general assumption is that all different feelings arise from different groups and change the composition of basic feelings. This is the most preferred and widely used model in verbal emotion recognition systems. Among the most critical theories mentioned in it, we mention the following(Ferro 2017).

Ekman model:

Ekman created his separate theory, postulating that emotions are divided into a small number of families. Each one contains many emotions that are pretty similar to each other. Five families of his emotions are anger, sadness, joy, fear, and disgust. He can develop this anger into wildness, indignation, revenge, hatred, and other feelings.
He claims that these differences are social. He also defines basic emotions as discrete emotions that have their origins in evolutionary adaptations to the environment. His argument for families of feelings is based on his study of human facial expressions, which revealed that facial expressions of feelings within one family are similar(Ekman 1993).

Tomkins Model

Eight basic emotions are included in this discrete model. These basic feelings are felt with varying degrees of intensity. Tomkins gave each of his eight basic feelings two names, one for a weaker version of the same primary emotion and one for a more robust version of the same primary emotion. Shame / humiliation, anger / rage, distress / anguish, contempt / disgust, fear / horror, surprise / startle, enjoyment / joy, and interest / excitement.(Tomkins 1981)

Plutchik model

Fear, anger, sadness, acceptance, contempt, delight, expectation, and surprise are all included in the Plutchik model. However, every basic feeling has an opposing primary emotion: surprise contrasts with expectation, disgust with acceptance, sadness with joy, and anger contrasts with fear, see figure 2.2.

Figure 2.2: Plutchik's wheel of emotions.
(El-Naggar 2017)

Unlike Tomkins' model, its intensity was rated with three words instead of just two. Additionally by drawing each of the eight emotions on a circle and flipping the circular pattern:
this representation allows you to indicate not only that each primary emotion is in opposition to the other, but also that each primary emotion is close to the other two. Then by mixing each pair of emotions, we can categorize 32 emotions in total e.g:
optimism = expectation + joy
Aggression = anger + fear
remorse = sadness + disgust...etc (Ayadi 2011).

2.3.2 Dimensional emotion model

The dimensional model is a better alternative for identifying high-level or complex emotions rather than the basic pivotal one, where it uses a small number of dimensions to describe emotion in everyday communication, rather than identifying Define separate, unique, and special categories for each .of these dimensions(Koduru 2020):
- Equivalence: which describes an emotion as a positive or negative emotion.
- Excitement: This dimension describes the strength of the feeling involved in speech, Such as the difference in the degree of interaction and emotion between excited and indifferent or bored(Koduru 2020). and Among the emotional models of the most popular dimensions, there are :

The two-dimensional model

It is one of the most preferred models because it uses the excitation dimension versus the activation dimension, or it uses equivalence as a dimension and evaluation or excitement in the other dimension(Koduru 2020).

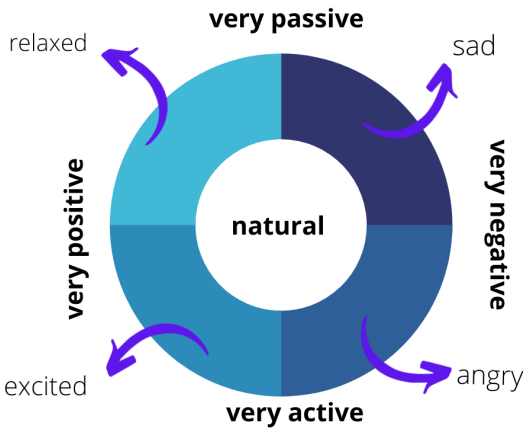

Figure 2.3: two-dimensional model.

Three-dimensional model

This model includes a third dimension representing the apparent strength of emotion, as it defines an apparent description of it between strong and weak. This dimension can distinguish the emotion of anger from fear, for example, by looking at weakness and strength in expressing emotion.

However, in general, one of the disadvantages of the emotional dimension model is that it is not sufficiently intuitive and inferential. Therefore, it requires special training for speech emotion recognition systems to extract the emotions involved accurately. Some of these emotions have both Negative and positive equivalence, such as surprise, where their equivalence varies according to the context in which they appear(Koduru 2020).

2.4 Speech Emotion Recognition Database

Emotional speech databases are the basis for any system that recognizes speech emotions, and they are the main factor in determining its efficiency. As a result, it is necessary to consider whether the database is quite sufficient to allow accurate classification conclusions to be drawn. Thus the quality of the evidence influences the success of the discrimination process. The database of emotional speech is a collection of inspirational sayings, each of which one assigned a specific emotion. They are separated into categories based on how they were created. Emotional speech databases are divided into three groups based on their nature: acted, induced, and natural. Where The acted emotional database

is one in which emotional discourse is elicited through the use of representation, the evoked emotional database is one in which the speakers' emotions are stimulated by external supervised elements such as a movie, photo, or other media. Finally, a natural emotional speech database in which the sayings record a conversation, such as a television or radio interview. Each type of database has, as expected, its own set of advantages and limitations, which we will discuss in this chapter.

2.4.1 Natural emotional databases

Natural (or spontaneous) emotional databases can be collected in various ways, including through television programs, call center conversations, and even recorded cockpit discussions from crashed planes. Aside from the obvious benefit of natural emotional databases being authentic and thus a reference, there are a few less obvious benefits. One of the implicit benefits is examining the expression of emotions in different circumstances. We will return to this feature when examining the context. On the other hand, these databases are characterized by the presence of many simultaneous emotions a mixture of primary emotions(Koolagudi 2012).

The experience of mixed emotions does not always mean that their expression is also a combination of the expressions of each of the primary emotions. If the expression of mixed feelings is a mixture of expressions of both feelings, then a single database of emotional speech will be complete. It contains only the basic feelings, making mixed feelings unnecessary. What are the disadvantages of the Natural Emotional Speech Database? The first limitation is the time and work required for a single task: first of all, it is unethical to register someone without his permission. Second, when the speakers realize they are being recorded, the context changes, and the way they convey their feelings changes also(Ferro 2017).

Third, all emotions may not be adequately expressed, and even if they are, some may not be adequately represented. Another limitation is the poor quality of speech. Microphones are not always of excellent quality, and vocalizations from multiple sources are not always identical. Furthermore, these recordings will always be buried in noise, which poses a significant challenge when dealing with classification algorithms afterward. Finally, since nested words are impractical for analytic propositions, The number of emotional speech has decreased(Koolagudi 2012).

2.4.2 Induced emotional databases

Natural data is artificially stimulated (or elicited) in an emotionally induced (or extracted) database. The goal is to have more control over the data while improving the quality of speech material regarded as credibility. The choice of the moment of registration and the location significantly contributes to improving the quality. As a result, the location may be a soundproof studio with a sophisticated microphone. Although the researcher will not have complete control over the emotions that can be accessed, he will be able to influence them. Some examples of ways are to make the speaker watch a movie or listen to a song, or involve the speaker in a conversation about his dreams, accomplishments, struggle, etc. Despite this, the use of a stimulus is not always simple: depending on the speaker, a single stimulus may lead to a variety of emotional responses. As a result, overlapping feelings will be communicated, whether they are wanted or not. This technique has a significant disadvantage that is speakers may express themselves less honestly if they are aware that they are being recorded (Koolagudi 2012).

2.4.3 Acted emotional speech databases

The advantages of represented (or simulated) emotional speech databases are controllable. The researcher may not only choose the place and time of the recording, as with the evoked emotional speech database, but he can also choose which sentences the actors have to utter. As far as this strategy is concerned with control, the feelings being expressed are unlikely to appear normal. This is because the actor is likely to exaggerate some vocal features while ignoring others. This is also because when the speaker experiences a particular feeling, such as struggling with anxiety over a personal issue, he may inadvertently express those feelings during the act.So what appears to us here is an important question which is what does the actor think and feel when he acts? Because the more he feels the desired emotions and thinks as if he is experiencing this feeling, the more accurate the database of emotional speech acted out(Ferro 2017).

As a result, the more natural the database is, the better the representation. Also, an interesting experiment that has yet to be completed is having a mixed database with both natural emotions and performed one, with clips reviewed by expert judges, who will determine if each clip was represented or natural. Such a test will reveal whether the unique vocal features of the performance are present, and it will certainly reveal that the better the actor, the fewer the specific affecting vocal features. As a result, Stanislavsky's method is the recommended and widely used method of representation. The objective is to recall a distant memory in which the desired feeling was experienced in order to relive that experience. The actor will be directly immersed in acting as he is experiencing the desired emotions. As a result, the representation will be more natural. Actors are clearly free to be more creative by subjectively inducing desired emotions, such as by listening to music this motivating method differs from the stimulating method that can be used in the evoked emotional database, where speakers are stimulated by researchers rather than self-made. It is not contextualized because it requires longer recording time, greater budget, and greater acting experience(Ferro 2017).

2.4.4 Data Quality Standards

Among the essential criteria for the accuracy of the database, we mention:

- Real Emotions: The less we act on emotions and take them from a real-world environment, the best the quality of the data becomes, like: talk on the radio.
- The quality of the speakers acted in the data recording.
- Quality of words and language materials.
- The way to simulate feelings if the data is circulating and not actual.
- Balanced distribution of sayings about emotions: the number of sayings between feelings is equal.

Last but not least, Classification performance should be focused on the style of speech and not the semantic meaning of the words.(Ayadi 2011)

2.4.5 A review of Arabic Emotional Speech databases

We have decided to create a database of emotional discourse in Standard Arabic. Thus, read Chapter Three for more information on these, such as the style of acting and the choice of actors, textual material, and other essential questions.
We will now begin to discuss The scope of the rules of evidence for emotional discourse in the Arabic language. Table 2.1 represents a comprehensive survey of existing Arabic discourse databases.

Name	Type	MSA or dialect	Speakers	Linguistic material	Emotions	acc
KSU Emotions[2014, 2018]	Acted	MSA	20 actors	16 sentences	sadness, happy, and questioning.	private
REGIM_TES [2015-17]	Acted	Tunisian	12 actors	-	Happiness, Anger, Neutral, Fear, and Sadness.	-
Egypt[2017]	Acted	MSA	7 actors	500 sentences	happiness, sadness, fear, anger, inquiry ,and neutral.	private
ANAD [2018]	Natural	Dialectal	6 actors	Eight videos of live calls	happiness, anger, and surprise.	public
BAVED [21-09-2019]	Elicited	MSA	61 actors	- words	Exhausted,Natural,happiness, and joy.	public
EYASE [04-2020]	Acted	Egyptian dialect	6 actors	Egyptian TV series	angry, happy, neutral, and sad.	-
ADED[08-02-2021]	Natural	Algerian dialect	32 actors	Tv shows	anger, fear, sadness, and neutral.	-
Saudi database[09-2021]	Natural	Saudi dialect	-	YouTube videos	anger,happiness, sadness, and neutral.	-

Table 2.1: Datasets of Arabic emotional speech

KSUEmotions

The Linguistic Data Consortium created the Saudi Representative Database in 2017. The first Arab version contains 5 hours and 10 minutes of recording, divided into 3276 audio files, with the participation of 20 representatives from three different Arab nationalities and an equal number of men and women aged between .20-37. The linguistic material was designed from 16 sentences divided into ten selected sentences, two words (yes, no), and interrogative keywords. The database includes five feelings of neutrality, happiness, sadness, surprise, and anger. It was subjected to expert evaluation in two stages, where the second stage is an evaluation extension of the results of the first stage, where neutrality gave the highest evaluation percentage, estimated at 93.2%, and sadness and surprise gave an equal percentage of 92.6% and 86.0% for anger. At the same time, the results of happiness evaluation were the lowest at an average of 77.4%(Meftah 2021).

Algerian Dialect Emotion Database

evidence base is a category of natural emotional discourse rules using six famous films in Algerian dialects that include four different emotions: fear, sadness, neutrality, and anger, with the participation of 32 actors (16 females, 16 males) from different regions of Algeria and from the age group (18 to 60 years) It consists of 200 audio files with a length of 0.2 to 0.3 seconds, 52 of which are for anger and fear, and 48 files for each of neutrality and sadness. Twelve sentences were used at a rate of three sentences for each emotion. When applying the KNN deep learning algorithm to the ADED database, an overall evaluation quality of 84.26% was obtained when working on the four previously mentioned emotion categories(Dahmani 2019).

2.4.6 Emotional context database

Context is the variable that causes the same feelings to be expressed differently by the same person. For example, a person expresses his feelings at home but does not express them at work in the same way. It is the degree to which feelings are hidden from others. Therefore, when developing an emotional database, one should always specify context. Unfortunately, insofar as it was done when creating emotional-nature databases, the context was neglected when dealing with represented emotional-speech databases. Cowie identifies four types of context: semantic context, structural context, intermodal context, and temporal context(Ferro 2017).

2.5 Features extraction

The efficiency of emotional speech classification in emotional recognition systems is related mainly to extracting appropriate signal features that allow accurate identification of emotional content without relying on the speaker's nature or the linguistic and lexical content of the signal. So that the more carefully designed the set of traits, the better the system performance in the rate of discrimination where SER systems have used different categories of features. Still, no specific category is better evaluated and more accurate than others, and all studies on this are considered experimental so far(Akçay 2020).

The issue of features extraction takes into account that the field of extraction is general and comprehensive or local to each frame that is fragmented from the total signal, Where many researchers believe that the global features are better in terms of classification accuracy and duration because their number is much less than the local features, which requires less processing time.
In addition, it is excellent and effective in classifying emotions of varying intensity, such as anger and surprise, whose intensity is very high compared to the severity of sadness, and this is one of its disadvantages at the same time, as it fails to distinguish between feelings of similar intensity accurately to distinguish between emotions of similar intensity accurately(Ayadi 2011).
The type of classification algorithms used can affect the choice of using or not using generic features. For example, they are not suitable for adoption in complex classifiers such as HMM and SVM due to the absence of the speech signal's temporal information altogether, which is not suitable for some algorithms that work on a lot of different features but are chronologically homogeneous. On the other hand, we will find that the local features are the most common in speech signal processing, as they take into account that the audio signal is not permanently stable. However, it can stabilize for a minimum period estimated at fractions of a second (20/30 milliseconds), and This is the principle of the framing process, whereby the total signal is divided into sub-signals into small units of time in order to extract the features of each frame separately(Ayadi 2011).

2.5.1 Types of speech features

There are many advantages that can be taken from the sound, among which we mention the following :

Prosodic features

The Prosodic features are the global features that can be perceived by human hearing, such as rhythm, intervals of silence, a voice in speech, and intonation, from the Typical examples of intonation are changes in the tone of voice or intonation that is added at the end of a sentence when a question or exclamation is asked. This feature type is also called pseudo-linguistic because, as we mentioned earlier, they are extracted from a large sign

unit representing integrated sentences and words. Therefore, these features are long-term and relatively related to the linguistic aspect of speech(Akçay 2020).

- **fundamental frequency (f0) :** Frequency is the expressive feature of a signal and an important characteristic. It is formed due to vibrations of the vocal cords that create a specific intonation and rhythm of speech.

- **energy :** The energy produced by speech can be defined as the size or intensity of the signal and a representation of the changes in the amplitude of the signal over time where emotions of high excitement such as anger and happiness produce increased and more energy compared to those that result from sadness, for example

- **Duration :** Duration is related to the time required to build words and pronounce vowels, the average duration of silent zones and vocal zones, and the signal(Akçay 2020).

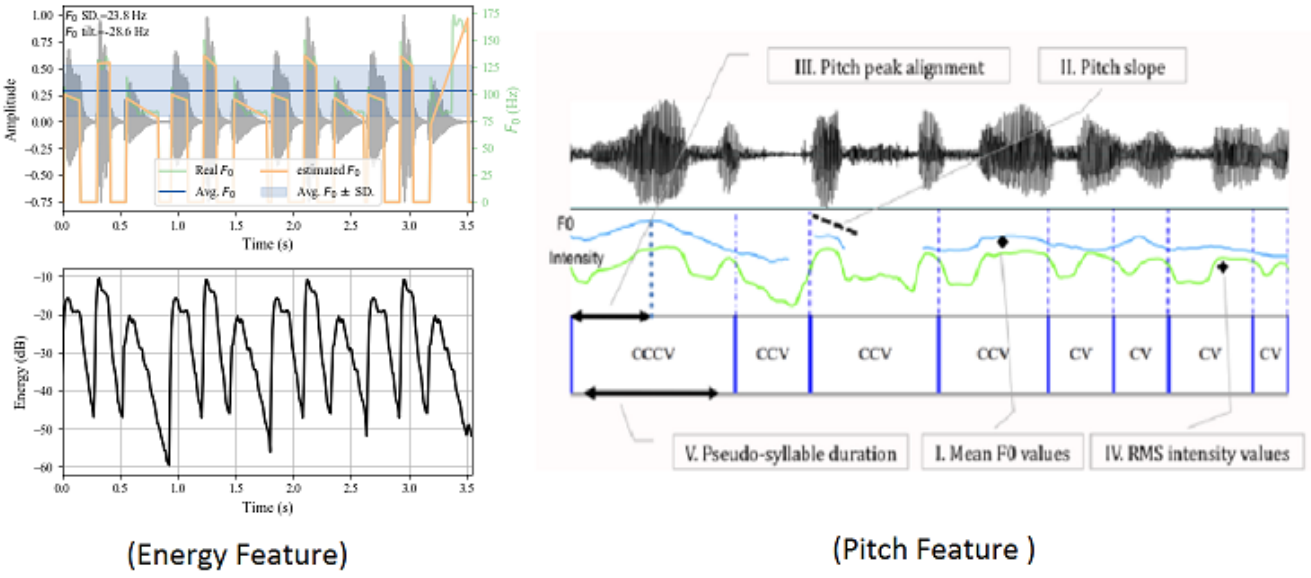

Figure 2.4: energy and pitch Prosodic features
(Biadsy 2009) (Vasquez-Correa 2018)

Spectral features

In contrast to the generic features, the Spectral features operate on small-range audio frames and convert them from the time domain to the frequency domain using the Fourier transform. For example, emotional content undoubtedly affects spectral energy distribution in the frequency domain. We note that the happiness emotion signal contains higher spectral energy than the fear, which has low energy in the same field(Ayadi 2011). In general, there are several techniques that allow extracting spectral audio features such as Mel Frequency Cepstral Coefficient (MFCC) , which is the most common extraction technique as it contains 39 different spectral features. And the Low Level Descriptors (LLDs) features, which includes 988 features, including intensity, pitch, loudness, 12 MFCC features, and others (Akçay 2020).

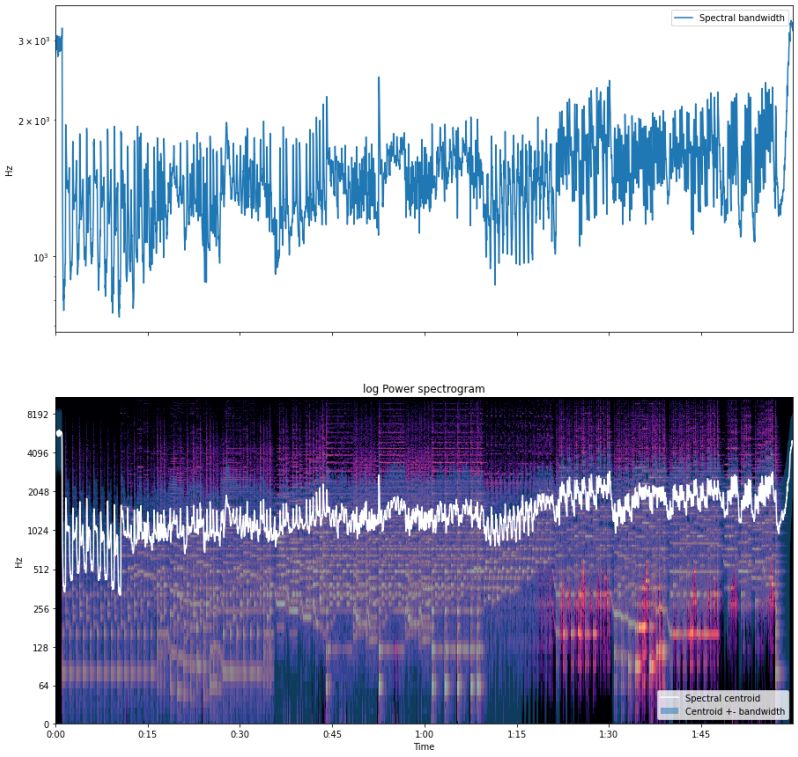

Figure 2.5: Example of spectral features

(Verma 2021)

Voice quality features

There is a close relationship between the perception of the emotional content of speech and the quality of the voice, where the quality of the voice is determined by physical characteristics such as:
(1) voice level: signal amplitude, energy, and duration be reliable measures of voice level;
(2) voice pitch;
(3) phrase, phoneme, word, and feature boundaries;
(4) temporal structures.
Quality features show the variation of audio and provide information about the linguistic composition of speech; for example, it selects the shouting based on the existence or lack of the vowel 'A' Or using the noise to determine the end of the speech content or the use of the breathing sound as a variable linked to both anger and happiness, While sadness is associated with the quality of the "ringing." . It is recognized that speech content affects spectral energy distribution via frequency. We find that talking with happiness has high energy in the high-frequency range while talking with the emotion of sadness has small energy in the same range.

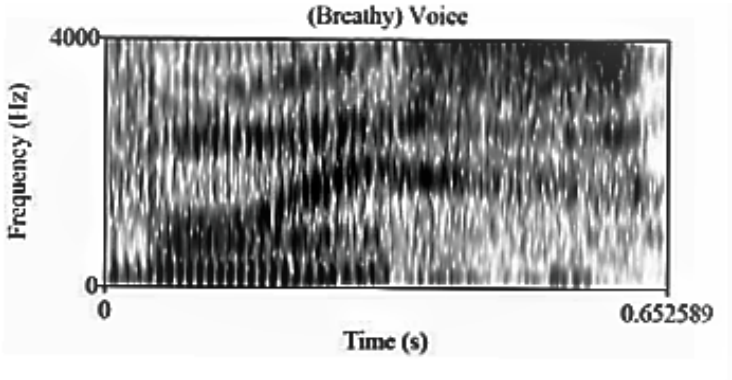

Figure 2.6: Breathy voice feature
(Yoon 2006)

2.6 Classification

After the data is collected, improved, processed, and then the appropriate features are extracted, the next step will be to select the appropriate classifiers to distinguish between sentiments. These classifiers represent Supervised Machine Learning algorithms, whether

classic traditional algorithms or deep learning algorithms such as SVM, Hmm. These classifiers take the data set as initial inputs and train to predict and determine its emotional classification. There is no prior and specific answer to the question related to choosing the best classifier because the selection criterion is linked to several variables. Among them is the nature of the classification task, the size of the data, the number of emotions, and the type of features extracted; some classifiers are more efficient with a specific type of features than others(Ayadi 2011). Among the machine learning algorithms used for classification are the following:

2.6.1 Support Vector Machine (SVM)

SVM is a supervised machine learning technique that may be used for classification or regression tasks. However, it is most commonly seen in classification issues. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is the number of features you have) with each feature's value. Then we perform classification by finding the hyperplane, which differentiates the two classes very well(Djeridi 2020).

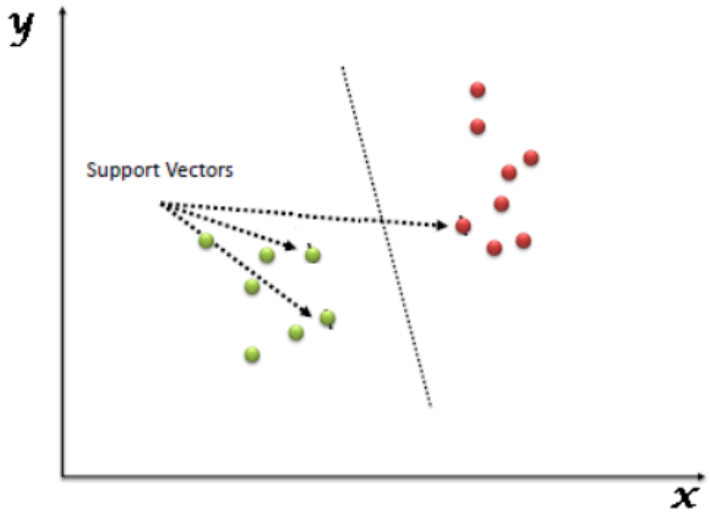

Figure 2.7: Support Vector Machine

(Djeridi 2020)

2.6.2 K-Nearest Neighbors (KNN)

KNN is one of the more basic machine learning techniques. This algorithm can be used for classification and regression problems, but it is most commonly used for classification. This algorithm is a model that classifies data points based on their similarities. It makes an "educated guess" about what an unclassified point should be classified as based on test data. KNN is a non-parametric algorithm, which means it makes no assumptions about the underlying data. It is also known as a lazy learner" algorithm because it does not immediately learn from the training set; instead, it stores the dataset and performs an action on it during classification. The following basic steps are taken in KNN (Fig. 2.6):
1-Calculate the distance.
2-Find the closest neighbors.
3-Vote for labels.
(Djeridi 2020).

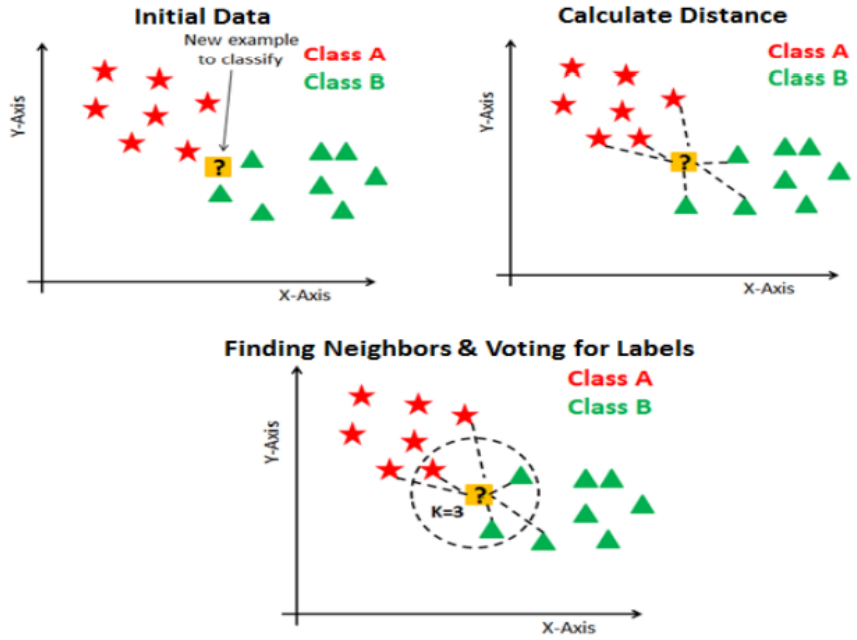

Figure 2.8: How does the KNN algorithm work
(Djeridi 2020)

2.6.3 Logistic regression (LR)

Logistic regression is one of the most widely used machine learning algorithms. When studying predictive modeling, one of the first few topics that people choose is logistic regression. It is not a regression algorithm but rather a probabilistic classification model that assigns discrete classes to observations. In contrast to linear regression, which produces continuous number values, it transforms its output using the logistic sigmoid function (Figure 2.7) to return a probability value that can be generalized to two or more discrete groups. Using distance scales such as the Euclidean distance, you can find the distance between points to the closest related point(Djeridi 2020).

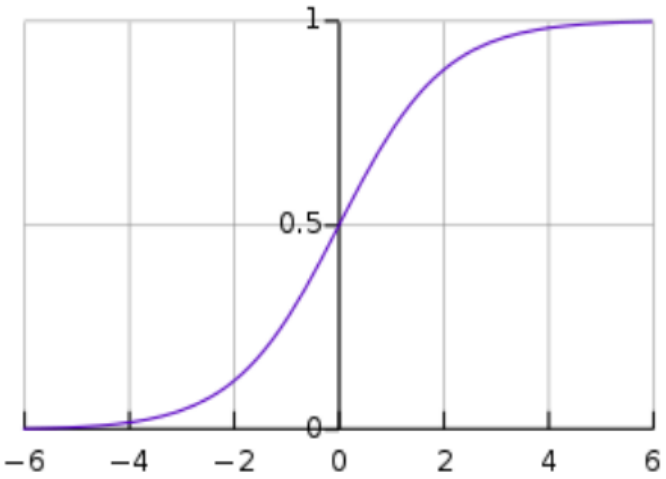

Figure 2.9: Graph for the sigmoidal function
(Djeridi 2020)

2.6.4 Speech Emotion Recognition Systems Applications

In addition to allowing more natural communication between humans and machines, it has been shown that the recognition of vocal emotions has a variety of uses:

- To keep drivers alert, use the onboard vehicle driving system..
- To improve the quality of call center services and analyze their customers' emotions during conversations.
- To create emotional interactive films, games, and electronic courses: these would be more interesting, if they were adapted to the listeners or the emotional state of the speaker.
- to index music or video by emotional content.
- To analyze recorded or recorded telephone conversations between criminals.
- To analyze emergency service calls and evaluate reliable calls. This is a useful tool for fire brigades and ambulance services.
- In the cockpit of aircraft, systems trained in stressed speech perform better than those trained in normal speech.
- to develop automatic natural speech for speech translation systems.
- to develop robots that act as companions, teachers and assistants. but not least,To create lie detectors, such as the commercialized X13- VSA PRO Voice Lie Detector 3.0.1 PRO (Ramakrishnan 2012).

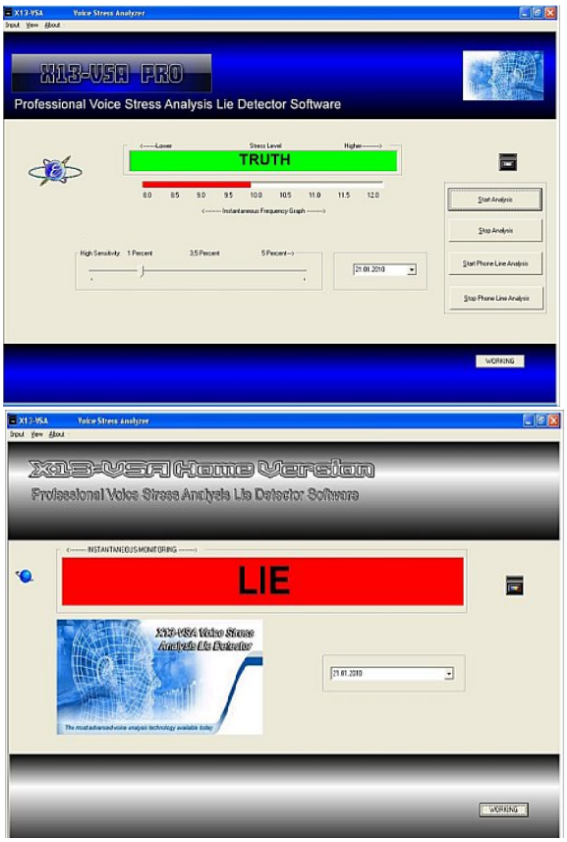

Figure 2.10: X13-VSA PRO Voice Stress Analysis Lie Detector Software (Client 2022)

2.6.5 Conclusion

Speech emotion recognition systems are among the most exciting technologies in recent times that serve the global trend in the development of human-machine voice interaction. During this chapter, we touched on the molecules related to SER systems, including the theories of emotion modeling that were reached by psychology, the studies that contributed to the work in Arabic speech in this field, Explaining the extent of the impact of databases and the nature of the speech features used on the performance of the system and the work of the workbooks.

Chapter 3

Experimental Settings

3.1 Introduction

Building and processing a valid database is the first step and essential toward building a high-performance speech emotion recognition system. The content of the database directly determines the quality of the final automatic classification. For this purpose, this section will present the methodology of building and processing the kasdi-Merbah University database for emotional speech and the nature of the tools and implementation conducted in this study.

3.2 Define speech database framework (scope)

Building the KEDAS database went through seven primary stages: preparing phrases, choosing feelings, identifying speakers, collecting audio recordings, preparing them, and pre-processing them to be evaluated finally. This section describes each step of this work.

Figure 3.1: Stages of building a Kedas database

3.2.1 Determine the emotions represented

Humans have a limited number of basic emotions and are based on the theory of discrete emotions, as mentioned in the first chapter. Therefore, we chose 4 of the basic emotions: fear, anger, sadness, and happiness.

Without addressing surprise and disgust, there is a significant overlap between them and between happiness and anger 1] and neutrality (no emotion) as a control for the KEDAS data set. Since the data set contains a balanced number of recordings in each of these emotions.

3.2.2 Linguistic material

The linguistic material used in the creation of the KEDAS database, in cooperation with the Linguistic Research Unit and Arabic Language Issues in Algeria, affiliated with the University of Kasdi Merbah, Ouargla, was determined under a study aimed at identifying the most famous phrases issued by people in different emotional situations. The study relied on selecting sentences in Modern Standard Arabic that are included Under the feelings of fear, anger, sadness, happiness, and neutrality.

To provide a documented scientific reference on which scientific studies are based,Taking into account the quality and size of the sentences and the context in which they are used to be closer to the customary use by the participants.

The proposed sentences were reviewed, discussed, and evaluated through three experimental stages to reach the required linguistic material to ensure adequate access to the emotional state of the actors. The result of the study that was approved by the research unit is as follows:

• Stop now	• توقف الآن
• Please act politely	• تصرف بأدب من فضلك
• I am so sorry	• أنا آسف حقا
• I know this is hard	• أعلم أن هذا صعب
• Be careful it falls down	• إحذر سيسقط منك
• Is there another solution?	• آلا يوجد حل آخر
• This is really great	• هذا رائع حقا
• Thank you very much	• شكرا جزيلا لك
• The second office is on the right	• المكتب الثاني على اليمين
• What can i do for you?	• تفضل

Note: The word "تفضل " does not have an exact meaning in the English language, so the sentence that is used has been placed in the closest context to it.

3.2.3 Actors selection

Taking into account the difference in the quality, type of voice, the style of speech, and the way of expressing emotion is distinct and specific to each person compared to others, so we will need a good number of participants and an appropriate diversity in the data to create an independent and effective recognition system Figure3.6.

The study community and the complete sample of participants were identified as part of the crew of the University of Kasdi Merbah - Ouargla - Algeria, including "students - workers - administrators - professors" so that the total number of participants within this work is five hundred actors of both sexes (male and female) and from different age groups ranging from [20-70] years old.

All representatives were native speakers of the Arabic language and from different geographic areas within the borders of Algeria, and this is one of the motives that make the university community an ideal study sample to provide the advantage of diversity and attract individuals from different Algerian provinces, and thus the diversity of speaking dialects.

Where the participants were reached through media promotion through social networking sites, in addition to presenting the project through the activities of the conference on Digitization and Digital Transformation, which was launched at the University of Ouargla on 05/10/2022.

The actors included a content maker on social media and some amateur theater workers. In addition to active public speakers in the field of singing and a group of memorizers of the Noble Qur'an, Because they are the people of the language, undoubtedly, this is due to the comprehensive sample of various scientific, social, and professional orientations that did not belong to Previous professional acting experience to preserve the natural ratio of performance and avoid exaggeration that You will appear at work if professional representatives are adopted Or semi-professionals.

All contributions to this study's database registration were made without charge, with the owners' approval, and within the bounds of scientific research ethics and applicable legislation.

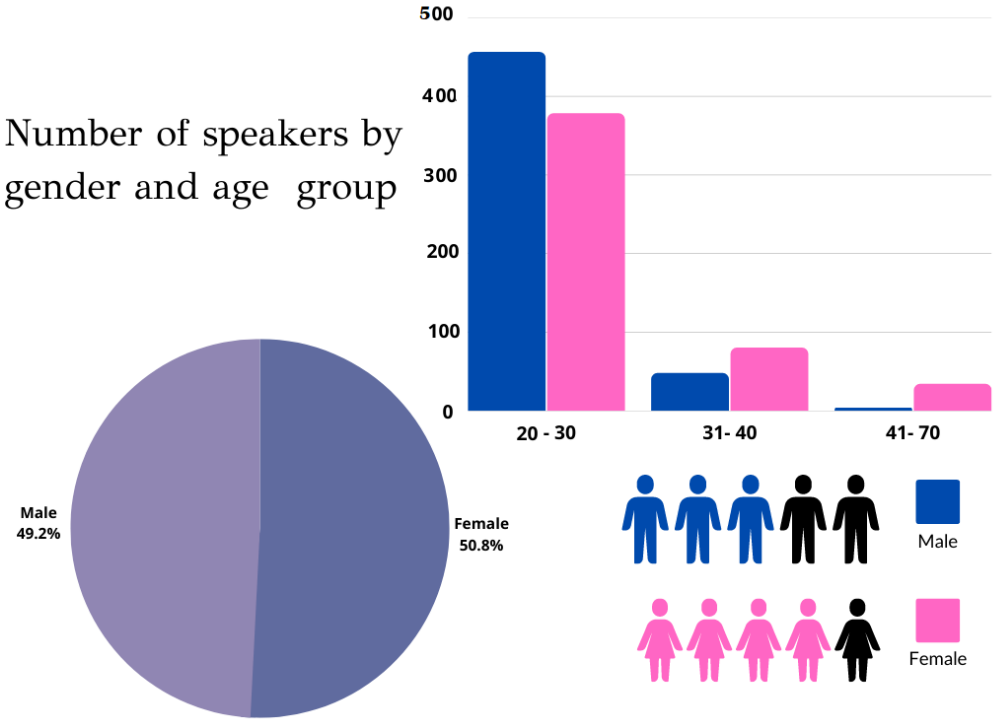

Figure 3.2: Statistics of the age and gender of the actors.

3.2.4 Ethics declaration :

Participants in the KEDAS creation experiment were treated by the Declaration of Helsinki, where written consent was obtained before registration from all participants. These individuals gave written informed consent for the participants' data (name, surname, age group, etc.) Moreover, how to use the recorded audio files, as described in the PLOS model.

3.2.5 Download and Accessibility

The primary goal of KEDAS is to contribute to enriching the body of evidence for the emotional discourse of the Arabic language.To this, the KEDAS database will be free and available to researchers and interested parties.

The KEDAS database can be viewed on the Linguistic Data Consortium site, and it is an open consortium of universities, companies, and government research laboratories. It

creates, collects, and distributes speech and text databases, lexicons, and other resources for linguistics research and development purposes.

3.2.6 Acting and recording method

The actors first receive a full explanation about the sentences that they will record and the nature of the performance required of them, with an explanation of the part related to the fact that the performance should be closer to their natural psychological This makes the repetition of reading the sentences and seeing them more than once contribute to the approximation of its use and the emotional stimulation of the participants. The recording was done individually to ensure the comfort of the experience for each participant unless the actors wanted to share the experience with their friends or companions; they were also free to adopt any method they saw that might help them enter the required emotional state, They had an absolute choice in determining the appropriate time for their readiness to register and enter the required psychological state so that the maximum time for the participant to take is about 15 minutes, during which the participants varied into four categories:

First category

A good part of the participants was asked to listen to an experimental sample to be able to evoke the natural elaboration of the sentences while clarifying that the sample presented represents itself only and alerting the actors about their need to separate themselves and their performance from the performance of the experimental sample he listened to, as the experimental recording was prepared by a team The research after observing that it helps a specific type of participant who sees that they need to listen to some form of emotional performance in order to be able to determine their performance.

Second category

The second type of participant was more motivated by having an open conversation with the research team, inquiring more deeply about the nature of using sentences, attempting to project it onto his own life, and identifying similar situations in which he could add

similar linguistic material as a way of approaching and simulating more spontaneous performance.

Third category

This category is characterized by greater privacy than others, as they prefer to remain entirely isolated to record their performance. Therefore, the audio recording tool was equipped with the applicable mechanism. As a result, the participant was left alone during the entire recording process until he gave a signal to the research team.

Fourth category

The fourth section of the participants did not find it difficult to enter the appropriate psychological state to start recording and saw that the sentences were closer to nature; this made them emotionally self-motivated.
The repetition of performing sentences was a necessity that all participants underwent before registration; they were also asked to represent each sentence twice during the main recording to ensure the smoothness of using the sentences and more excellent quality of performance, in addition to the periodic evaluation of each recording by the research team where the initial quality is determined To record and return it in case of damage due to noise in the background of work or other. There was also a sample of participants who preferred to listen to their recordings after completion. To assess the correctness of their performance and its repetition, they found that they could provide a realistic performance that satisfies them and represents them more. Finally, the collection of audio recordings was through two stages, In the first stage, 300 participants were reached, including people who listened to an audio sample as an example before recording their acting performance. During the second phase, stimulating subscribers was dispensed entirely with using the pre-acoustic stimulus to ensure sincerity and not negatively or positively affect performance. Therefore, registration was carried out during this stage with an additional 200 subscribers of both sexes to be the total number of participants in the registration KEDAS database 500 subscribers collected in over 52 separate recording sessions.

3.2.7 Recording conditions

The audio recording was in ordinary non-soundproof halls where professors' halls, university auditoriums, computerized media laboratories, and university residence rooms were sites for recording participants. The microphone was placed approximately 25 cm in front of the actor at a 90-degree angle of capture. As for recording, the Zoom H8 device was used. It contains all the essential features and characteristics of a professional audio device. It has been specifically designed for field recordings. In addition, it allows the possibility of verifying the recording process (recording, reading, storing, naming, tracking the interruption or discontinuation of recordings, and determining the file format) thanks to the easy-to-use graphic interface. Where the microphone was directed at the angle of capture 90 degrees, and the degree of sound capture was set between 3-6 degrees. Recorded files were taken at a frequency of Wav 44.1 kHz/24bit.

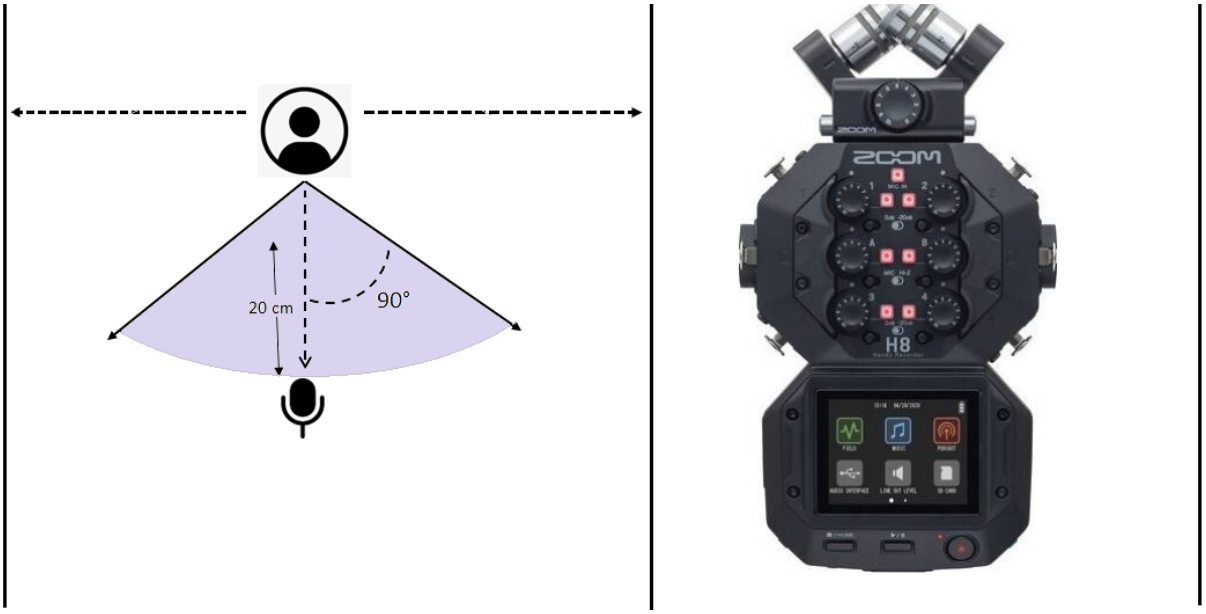

Figure 3.3: Recording environment.

3.2.8 KEDAS filenames

Each audio file in KEDAS has a unique name consisting of five identifiers consisting of numbers and letters separated by hyphens, for example (134-m-02-E1-01.wav), indicating the actor's number - the actor's gender - the age group - the emotion represented - Registration number.

Identifier	Coding description of factor levels
Actor	001 = First actor, . . ., 500= Five hundred actor.
Gender	m =male ,f=female.
Age group	01=20-30 years,02=30-40 years,03=40-50 years.
Emotion	E1=Angry,E2= Sadnes,E3= Fear,E4=happiness,E5=neutrality.
Statement	01=first statment ,02=second statement.

Table 3.2: Description of factor-level coding of KEDAS filenames.

3.2.9 Post-processing of data

The recordings were processed using the Audacity program Where the audio clips were divided, and the best recording was chosen in terms of sound quality from among two recordings for each representative sentence, as defects were discovered in the recording of 157 audio files, which were divided as follows: 52 files for the emotion of anger, 35 files for the emotion of fear, 32 files An emotion for happiness, 28 files for an emotion of neutrality, ten files for emotion for sadness.

3.2.10 Description of database

Kasdi Merbah University database is an audio data-set of emotional speech in the Arabic language specially developed for training on speech recognition systems. KEDA contains 5,000 audio files distributed parallel to the five basic emotions of anger, sadness, fear, happiness, and neutrality. These were expressed by 500 non-professional speakers, including 254 female and 246 male .

The speakers' ages range from [18-70 years], including 417 people from the youth category

and 64 people from the middle and old category.

Each actor utters ten different sentences so that each emotion contains two sentences. The phrases are lexically identical to Standard Arabic and were selected based on a study conducted by the Linguistic Research Unit and Arabic Language Issues of the Kasdi Merbah University of Ouargla .

This results in 5,000 audio files with 1,000 recordings per emotion and ,The following table provides comprehensive statistical information about KEDAS.

Data-set name	kasdi-merba emotional database in Arabic speech.
Year of creation	June 2022
Data-set type	Acted
File Type	Audio only
File format	.wav
Sampling rate	44.1 KHz
Number of speakers	500
Number of female speakers	254
Number of male speakers	246
Age of speakers	[20-70 year]
Number of emotional states	05 emotions
Number of words	26 word
Number of audio clips	5000
Size of the data-set	1.47 Go
Average duration of each clip	[0.5 s to 2.5s]
Utilized Software	Audacity
Utilized hardware	Zoom h8
Number of statements	10= [9 sentences + 1 word]
Duration	01 :20
Accessibility	public

Table 3.4: Data-set description.

3.3 Spectral analysis of data samples

To represent the sound wave, we can use the spectrogram; it is a very suitable visual technique that allows showing the strength of the sound and the change of energy levels over time. It is represented by both the time dimension and the frequency dimension, taking the amplitude as a third dimension that is color-coded Each audio spectrum has its color system, but the color scale is generally red, green, and blue. Blue corresponds to low amplitudes, red corresponds to high amplitudes, green and yellow represents the gradation between the two values.

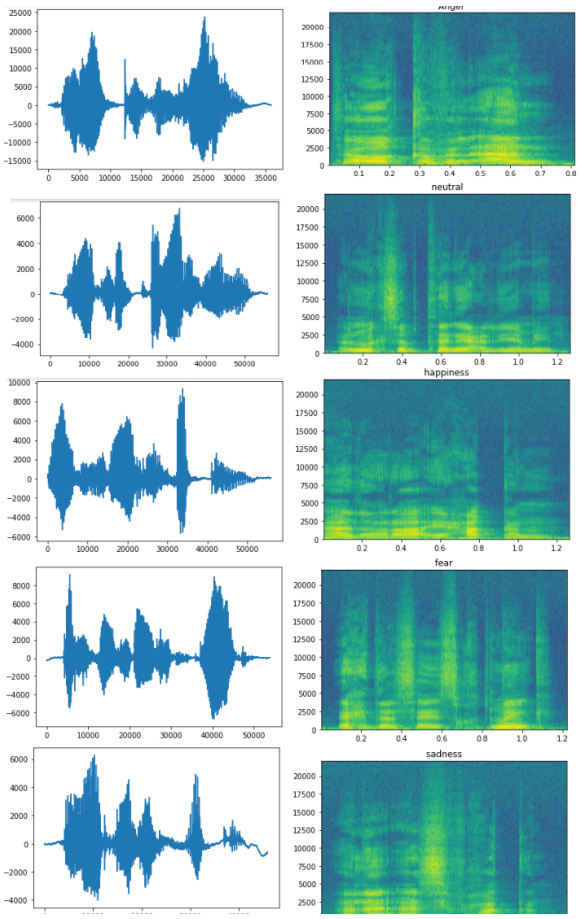

Figure 3.4: A waveform plot of a randomly chosen sample of each emotional state

3.4 Result and evaluation

3.4.1 Statistical evaluation

Evaluating and determining the validity of the database aims to clarify the extent to which natural persons can identify and categorize the feelings contained in the audio files as a first stage that leads us to anticipate the possible results in automatic recognition. Therefore, 56 native Arabic speakers participated in the assessment of the KEDAS database, including 11 assessors specialized in clinical psychology, 45 non-professional doers, including participants in the audio recordings, and random groups of Kasdi Merbah University students. all residents in good health (physical and mental fitness). They do not have any hearing impairment or diseases that prevent focus or assimilation of good information, and the fact that they are all over the age of 20 years. The evaluation was conducted through 5 separate sessions, with an average of 10 to 11 assessors in each session. The duration of each session ranged from one hour to two and a half hours, during which a total of 500 audio files were evaluated. The residents at each session are in a closed room and far apart while controlling the limits of using phones and external factors, including controlling communication between residents to ensure that there is no mutual influence between them With an explanation that the evaluation principle should be based on the actor's vocal performance away from the linguistic context, Each listener fills out a written form divided by the number of actors and the number of audio files for each emotion and actor. He asked the assessor the following:

- **Determining the category of emotions :** This cognitive test measures the recognition rate and human classification of the five emotions included in the database [happy, afraid, neutral...].

- **Determining the clarity's intensity of emotion:** point Likert scale was adopted to assess the intensity of emotion into five levels [feeble (1) - weak (2) - medium (3) - strong (4) - very strong (5)] where the intensity of high emotion corresponds to a higher expression quality on the part of the actor, and thus its clarity with the listeners.

Validity task

For the recognition validity test, We extracted the confusion matrix to quantify the emotion identification test results as a tool oriented to machine learning systems that allow measuring the accuracy of classification predictions compared to the correct classifications of the data. The results are encoded with a value of 1 if the evaluators prediction result matches the original rating of the audio file. In contrast, we encode the value to 0 if the original rating conflict with the evaluators choice. Therefore, this matrix will allow comparing the results issued by 56 judgments with the actual rating for 500 audio files, as we mention that ten evaluators test each audio file. The confusion matrix is expressed as the five predicted emotions versus the original categories; This allows the accuracy of the KEDAS classification to be measured based on the average overall test results. The diameter of the matrix represents the correct classification locations where the predicted category matches the correct category. Based on the chromatic map system, the darker the color, the more it corresponds to a high predictive value, and All values located outside the diameter are false predictions made on the data, such as evaluating one of the audio recordings as being in the sadness category when it is in the anger category, for example. We have calculated the following transactions:

TP :true positive

TN: true Negative

FN: False Negative

FP: False positive

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

$$Recoll = \frac{TP}{TP + FN} \tag{3.3}$$

$$F1\_Score = 2 * \frac{Recoll * Precision}{Recoll + Precision} = \frac{2 * TP}{2 * TP + FP + FN} \tag{3.4}$$

The results obtained are shown in the table3.6 ,and The obtained confusion matrix is shown in figure[3.5] which gave an accuracy rate of 75.08%. By extracting the results of the precision of recognition for each emotion, neutrality gave the highest accuracy rate of 87.7%, followed by happiness, anger, fear, and sadness, respectively.

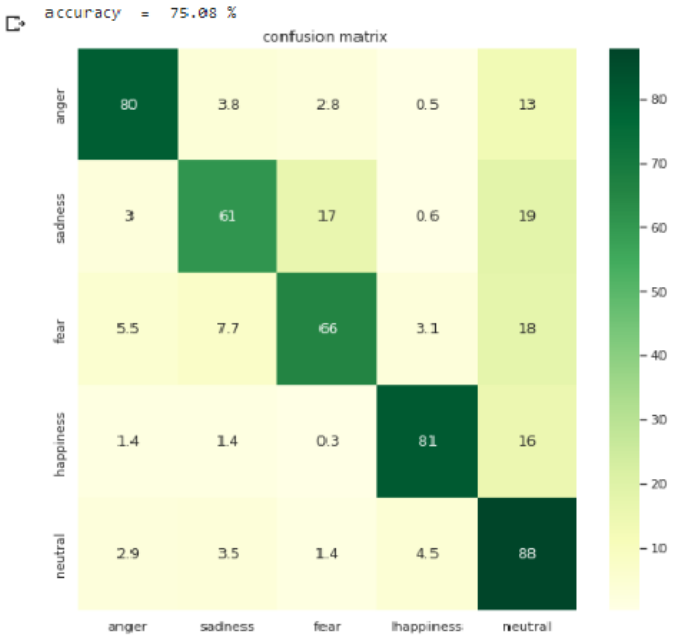

Figure 3.5: KEDAS confusion matrix.

3.4.2 Reliability rating

Fleiss' kappa (Fleiss 1971) was used to assess the reliability of the database test, as this scale represents the percentage of the assessors' agreement on classifying the emotions represented and determining their intensity, A matrix containing 500 randomly selected audio recordings was used against five emotions, where each audio clip was evaluated by ten assessors, which gave a result of $\kappa = 0.53$ as a total result for the reliability of the KEDAS database arbitration; in addition, kappa scores were calculated for each emotional category.

The results are shown in Table 1. Kappa values are interpreted according to the guidelines set by Landis and Koch (Landis 1977), where values <0 indicate poor agreement, 0.01–0.20 slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement, and 0.81 1 indicates almost complete agreement.

Emotions	precision	recall	f1-scor	κ
Sadness	61.1 %	86.2 %	0.82	0.022
Fear	65.8 %	78.8 %	0.68	0.093
Anger	79.7 %	91.4 %	0.76	0.022
Happiness	81.2 %	89.18 %	0.85	0.046
Neutral	87.7 %	51.9 %	0.65	0.016
Total	accuracy :75.08%	-	-	kappa: 0.53

Table 3.6: Evaluation result.

3.4.3 Discussing the evaluation results

KEDAS includes 5000 audio files evenly distributed into four basic emotions in addition to the neutral emotion. Five hundred randomly selected files were evaluated without any bias to be tested by 56 assessors, including clinical psychologists, some of the representatives participating in the KEDAS recordings, and a random sample of Kasdi Merbah University students. The evaluation included a test of validity and reliability about the correctness of identifying and categorizing audio files correctly in human perception and evaluating the intensity and clarity of each audio file that was displayed. The outcome of this evaluation study will be the following: The corresponding figure represents the percentage of the precision of recognition of KEDAS audio data for each emotion. Evaluating the validity of the database indicates a good percentage estimated at 75.08% of the cases that the evaluators were able to perceive correctly, that is approximately 379 audio files out of 500 files. While calculating the accuracy rate for each emotion, neutrality took the highest judging in terms of health. An average of 88 audio files were correctly rated out of 100, Followed by happiness, anger, and fear, respectively. While sadness gave the lowest precision rate of61%, where there was a significant overlap between the emotion of sadness and fear Alternatively, sadness and neutrality, and through the recording sessions, we noticed that one of the phrases related to the emotion of fear might enter into more than one context of the speech. However, most of the speakers used it with sadness and neutrality. In addition, the representation of sadness and fear required more training time and a more stable psychological state than the participants had. Where were they located

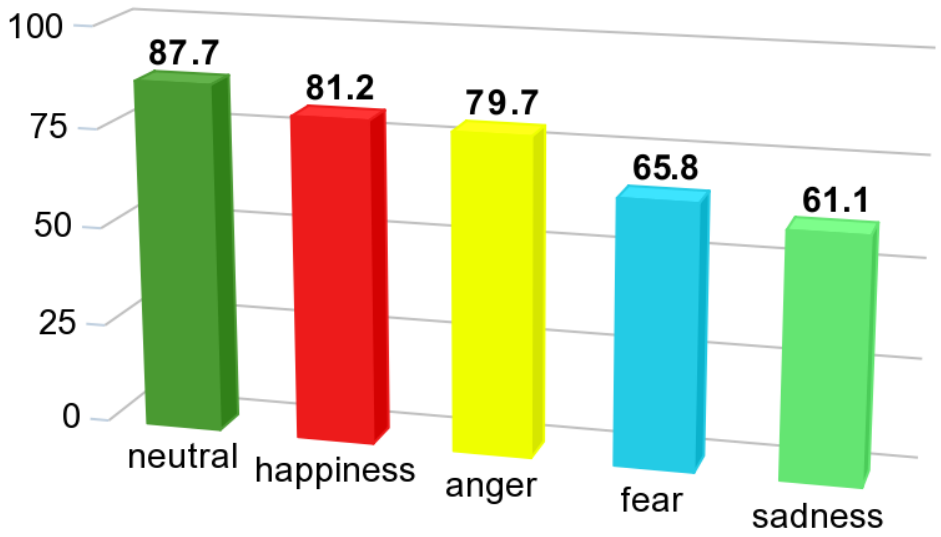

Figure 3.6: Precision recognition of each emotion.

in a very dynamic and lively environment, such as the university and work offices, and this represents a kind of restriction explicitly in presenting the situation required by the actors. On the contrary, we noticed that the actors' emotions of happiness and anger had a faster response in terms of expression, which is what the evaluation results proved.

3.5 Speech Emotion Classification Method

In this section, we extracted the MFCC features, and the SVM classifier was applied to 500 audio clips from KEDAS, which was validated as an additional procedure through which we wanted to estimate the validity of the data in speech emotion recognition systems. Specifically with the SVM algorithm as one of the most widely used and efficient algorithms in the field.

3.5.1 Feature extraction :

Mel-Frequency Cepstral Coefficients (MFCC) : is the most used feature in recognition systems that deal with the speech signal because it takes into account the parallel characteristics of the human hearing system, in addition to the good results that he achieved

compared to the rest of the spectral features

To obtain MFCC, the following must be done:

(1)- framing and windowing The spoken speech signal into small segments is estimated at fractions of a second, and this is because the signal continues to change over time. Still, it can be semi-stable for a short period.

(2)- Each segment is transformed into the frequency domain using a short discrete Fourier transform(DFT),It is given by the following equation :

$$X(t) = \frac{A_0}{2} + \sum_{n=1}^{N}(A_n \sin(\frac{2\pi nt}{P} + \phi_n)) = \sum_{n=-N}^{N}(C_n e^{\frac{2\pi jnt}{P}}) \tag{3.5}$$

(3)- the sub-band energies are calculated using a Mel filter bank; This allows the signal to be separated and analyzed into multiple components.

(4)- The logarithm of these sub-domains is calculated.

(5)- The inverse Fourier transform is applied to get MFCC

During this work, we have selected 32 mfcc features to work on with classification algorithms.

3.5.2 Data preparation

Before implementing the machine learning algorithm on the extracted features, we did the following:

- Download the data in CSV format

- Then, we create the input and output data where x represents all the input columns, and y takes the corresponding labels.

-The next step is data segmentation, where we want a symmetrical distribution among all the applied algorithms. The data that was processed and extracted reached 564 audio

files. Therefore, a division of 80 percent of them was adopted as the training set, and the remaining 20 percent was adopted for testing the model.

3.5.3 Support Vector Machine (SVM)

The classification results given by the SVM algorithm were as follows:
Figure 3.7 shows confusion matrices for (MFCC features), so we can see that the highest known value was given at 91% The emotion of neutrality is matched by the emotion of sadness, with the lowest health rate equal to 40% all results of precision and return and a value of f1-score appear in the table 3.8

Figure 3.7: Confusion matrix of SVM result

Emotions	precision	recall	f1-scor
Fear	0.73	0.81	0.77
Anger	0.65	0.83	0.73
Happiness	0.46	0.50	0.48
Neutral	0.91	0.71	0.80
Sadness	0.40	0.25	0.31
TOTAL	acurcy : 0.62	-	-

Table 3.8: Results of SVM.

Subsection K-Nearest Neighbors (KNN) Set the number of neighbors in the algorithm to 4, The classification results given by the KNN algorithm were as follows: Figure reference Figure: knn shows confusion matrices for (MFCC features), so we can see that the highest known value is given at 92 %. With an emotion of sadness, with the lowest healthy percentage equal to 40 %, all accuracy and return results and the value of the score f1 appear in the table 3.10.

Figure 3.8: Confusion matrix of Knn result

Emotions	precision	recall	f1-scor
Fear	0.71	0.89	0.79
Anger	0.76	0.79	0.78
Happiness	0.41	0.83	0.39
Neutral	0.92	0.79	0.85
Sadness	0.40	0.33	0.36
TOTAL	accuracy: 0.63	-	-

Table 3.10: Results of knn .

3.5.4 Logistic regression (LR)

The classification results given by the RL algorithm were as follows: Figure 3.9 shows confusion matrices for (MFCC features), So we can see that the highest precision value was for neutrality like the rest of the algorithms, but it was fully recognized (100 %) The emotion of neutrality is matched by the emotion of sadness, with the lowest health rate equal to 38 % all results of precision and return and a value of f1-score appear in the table 3.12

Figure 3.9: Confusion matrix of RL result

Emotions	precision	recall	f1-scor
Fear	0.59	0.85	0.70
Anger	0.66	0.88	0.75
Happiness	0.40	0.33	0.36
Neutral	1.0	0.64	0.78
Sadness	0.38	0.21	0.27
TOTAL	accuracy : 0.58	-	-

Table 3.12: Results of RL

Note

All the code used in extracting and selecting features and training the classifiers was made by Librosa and Scikit-learn Learn Python libraries, which are libraries dedicated to audio data processing and Dealing with machine learning algorithms.

3.6 Challenges and recommendations

The following are several challenges and recommendations that the team faced during the development of the KEDAS database, which aims to transfer simple knowledge to each researcher who will work on parallel work in the future.

- Weak culture of contribution to scientific experiments Lack of adequate recording equipment, high-quality loudspeakers, and soundproof rooms
- The lack of dealing with a Standard Arabic in the target community
- The difficulty of providing a place with the minor noise or a sound echo in a vital environment such as the university
- The great reservation in contributing audio recordings from individuals
- Implementation of the work was within the period that is witnessing tremendous pressure for the university staff
- The difficulty of reaching a stable psychological state for an individual in a very dynamic and volatile environment such as work or study could have been reduced by giving more time to the participants

The problem of overlapping emotions for many participants The following recommendations can help researchers working in a work environment with similar conditions to shorten some problems:

- The participants' appreciation of the project idea plays a significant role in the extent of their interaction with the instructions and how to register.
- It is better to adopt a formal environment such as a university if the study is about a standard language such as Standard Arabic to find better ability and quality in terms of fluency and sound pronunciation compared to the performance of people from another environment
- Choose appropriate recording times away from peak times to ensure a less noisy background recording

- Recording by phone or non-professional equipment is not valid for this type of study; where will bad results be issued after the data pre-processing process
- The Periodic evaluation of each recording saves a lot of time Within this work, both the scientific research team and the participant were evaluating the performance and recording in each session, and this is one of the highly recommended points

General conclusion

This study focuses on building an audio database oriented toward a system to recognize human emotion for speech in the Arab world to strengthen limited databases in the field. The goals we set for conducting this study are first to build the Kasdi-Merbah emotional Database in Arabic Speech (KEDAS) by relying on a proposed methodology; It includes choosing the adopted emotional model, which includes anger, sadness, happiness, fear, and neutrality, then specifying the linguistic material with a total of ten sentences approved by the Linguistic Research and Arabic Language Issues Unit in Algeria - University of Ouargla. The database contains 5000 audio files recorded with the participation of 254 females and 246 males from the university community and varying age groups ranging from [20-70] years. Secondly, the validity of KEDAS data is verified by fifty-six (56) judges, including specialists in clinical psychology and a group of untrained individuals.

The test revealed reasonable emotional validity and reliability rates, as KEDAS data gave an accuracy rate of 75.08%, Where was neutrality the highest emotion in terms of recognition with a rate of87.7% It was followed by happiness, anger, and fear, respectively $(81.02\%, 79.7\%, 65.8\%)$, and sadness with the lowest percentage, 61.1%. The reliability of the Cohen's kappa test for each of the recognition and emotion intensity tests obtained a moderate agreement score estimated at (0.53 and 0.026), respectively.

The validated data set is small and contains only 500 audio recording files for the sum of the five emotional states. So, For further verification, a complete evaluation of the database is recommended. as well, We hope that good training results for learning algorithms that can be applied to the KEDAS database will be reached during future studies.

Bibliographie

Akçay (2020). “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers”. In: *Speech Communication* 116, pp. 56–76.

Ayadi, El (2011). “Survey on speech emotion recognition: Features, classification schemes, and databases”. In: *Pattern recognition* 44.3, pp. 572–587.

Biadsy (Jan. 2009). “Using prosody and phonotactics in Arabic dialect identification”. In: pp. 208–211.

Dahmani (2019). “Natural Arabic language resources for emotion recognition in Algerian dialect”. In: *International Conference on Arabic Language Processing*. Springer, pp. 18–33.

Ekman (1993). “Facial expression and emotion.” In: *American psychologist* 48.4, p. 384.

Fasold (2014). *An introduction to language and linguistics*. Cambridge university press.

Ferro (2017). “Speech emotion recognition through statistical classification”. PhD thesis.

Fleiss (1971). “Measuring nominal scale agreement among many raters.” In: *Psychological bulletin* 76.5, p. 378.

FNC (2013). “Vibrations Waves and Sounds”. In:

Guo (2018). “Dominant and complementary emotion recognition from still images of faces”. In: 6, pp. 26391–26403.

Harrat (2016). “An algerian dialect: Study and resources”. In: *International journal of advanced computer science and applications (IJACSA)* 7.3, pp. 384–396.

Koduru (2020). “Feature extraction algorithms to improve the speech emotion recognition rate”. In: *International Journal of Speech Technology* 23.1, pp. 45–55.

Koolagudi (2012). “Emotion recognition from speech: a review”. In: *International journal of speech technology* 15.2, pp. 99–117.

Landis (1977). “The measurement of observer agreement for categorical data”. In: *biometrics*, pp. 159–174.

Meftah (2021). “King Saud University Emotions Corpus: Construction, Analysis, Evaluation, and Comparison”. In: *IEEE Access* 9, pp. 54201–54219.

El-Naggar (2017). “Sentiment analysis of modern standard Arabic and Egyptian dialectal Arabic tweets”. In: *2017 Computing Conference*. IEEE, pp. 880–887.

Najadat (2018). “Multimodal sentiment analysis of Arabic videos”. In: *Journal of Image and Graphics* 6.1, pp. 39–43.

Ozseven (2018). “Evaluation of the effect of frame size on speech emotion recognition”. In: *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. Ieee, pp. 1–4.

Ramakrishnan (2012). “Recognition of emotion from speech: A review”. In: *Speech Enhancement, Modeling and recognition–algorithms and Applications* 7, pp. 121–137.

Ryding (2005). *A reference grammar of modern standard Arabic*. Cambridge university press.

Sakran (2017). “A review: Automatic speech segmentation”. In: *International Journal of Computer Science and Mobile Computing* 6.4, pp. 308–315.

Shivhare (2012). “Emotion detection from text”. In: *arXiv preprint arXiv:1205.4944*.

Tanyer (2000). “Voice activity detection in nonstationary noise”. In: *IEEE Transactions on speech and audio processing* 8.4, pp. 478–482.

Tomkins (1981). “The quest for primary motives: biography and autobiography of an idea.” In: *Journal of personality and social psychology* 41.2, p. 306.

Wani (2021). “A comprehensive review of speech emotion recognition systems”. In: *IEEE Access* 9, pp. 47795–47814.

Yoon (2006). “Voice quality dependent speech recognition”. In: *International Symposium on Linguistic Patterns in Spontaneous Speech*. Citeseer.

Webographie

Client (2022). *X13-VSA PRO Voice Stress Analysis Lie Detector Software*. URL: https://www.lie-detection.com/products.html (visited on 06/11/2022).

Djeridi (2020). *Emotion recognition in Arabic speech signal*. http://dspace.univ-ouargla.dz/jspui/handle/123456789/28896. [Online; accessed 23-04-2022].

Maryam (2020). *Properties of Sound Waves: Speed, Reflection and Echo*. URL: https://www.informationpalace.com/properties-of-sound-waves/ (visited on 04/24/2022).

Smales, Mike (2019). *Sound Classification using Deep Learning*. URL: https://mikesmales.medium.com/sound-classification-using-deep-learning-8bc2aa1990b7 (visited on 04/28/2022).

Vasquez-Correa, J. C. (2018). *Prosody Features*. URL: https://disvoice.readthedocs.io/en/latest/Prosody.html.

Verma, Yugesh (2021). *A Tutorial on Spectral Feature Extraction for Audio Analytics*. URL: https://analyticsindiamag.com/a-tutorial-on-spectral-feature-extraction-for-audio-analytics/ (visited on 04/24/2022).