CATÓLICA
ESCOLA DAS ARTES
PORTO

# SPEECH EMOTION RECOGNITION THROUGH STATISTICAL CLASSIFICATION

Dissertação apresentada à Universidade Católica Portuguesa
para obtenção do grau de Mestre em Som e Imagem

*Adelino Rafael Mendes Ferro*

Porto, Julho de 2017

CATÓLICA
ESCOLA DAS ARTES

PORTO

# SPEECH EMOTION RECOGNITION THROUGH STATISTICAL CLASSIFICATION

Dissertação apresentada à Universidade Católica Portuguesa
para obtenção do grau de Mestre em Som e Imagem

- Especialização em –
Design de Som

*Adelino Rafael Mendes Ferro*

Trabalho efetuado sob a orientação de

Pedro Pestana

Porto, Julho de 2017

# Dedicatória

À minha família: pais, irmãos, tios e avó.

# Agradecimentos

Para além do meu orientador, queria agradecer a todas as pessoas da UCP que diretamente se envolveram neste trabalho ou que contribuíram para o seu sucesso.

Queria também agradecer a todos os que comentaram construtivamente o trabalho, que motivaram ou que contribuíram para a sua realização.

Os meus agradecimentos calorosos a um grupo que foi particularmente indispensável: aos atores e às atrizes que se disponibilizaram para ter a sua voz gravada, para que fosse possível a construção de uma base de dados simulada de discurso emocional Português e aos que colaboraram no preenchimento anónimo de inquéritos, que visaram avaliar e validar a respetiva base de dados.

Por último, um agradecimento especial aos meus pais, aos meus tios, aos meus irmãos e irmãs, à minha avó e aos meus amigos mais próximos, por me terem apoiado durante este trabalho e durante a minha vida.

# Resumo

O propósito desta dissertação é a discussão do reconhecimento de emoção na voz. Para este fim, criou-se uma base de dados validada de discurso emocional simulado Português, intitulada European Portuguese Emotional Discourse Database (EPEDD) e foram operados algoritmos de classificação estatística nessa base de dados.

EPEDD é uma base de dados simulada, caracterizada por pequenos discursos (5 frases longas, 5 frases curtas e duas palavras), todos eles pronunciados por 8 atores—ambos os sexos igualmente representados—em 9 diferentes emoções (raiva, alegria, nojo, excitação, apatia, medo, surpresa, tristeza e neutro), baseadas no modelo de emoções de Lövheim.

Concretizou-se uma avaliação de 40% da base de dados por avaliadores inexperientes, filtrando 60% dos pequenos discursos, com o intuito de criar uma base de dados validada. A base de dados completa contem 718 instâncias, enquanto que a base de dados validada contém 116 instâncias. A qualidade média de representação teatral, numa escala de a 5 foi avaliada como 2,3. A base de dados validada é composta por discurso emocional cujas emoções são reconhecidas com uma taxa média de 69,6%, por avaliadores inexperientes. A raiva tem a taxa de reconhecimento mais elevada com 79,7%, enquanto que o nojo, a emoção cuja taxa de reconhecimento é a mais baixa, consta com 40,5%.

A extração de características e a classificação estatística foi realizada respetivamente através dos softwares Opensmile e Weka. Os algoritmos foram operados na base dados original e na base de dados avaliada, tendo sido obtidos os melhores resultados através de SVMs, respetivamente com 48,7% e 44,0%. A apatia obteve a taxa de reconhecimento mais elevada com 79,0%, enquanto que a excitação obteve a taxa de reconhecimento mais baixa com 32,9%.

# Abstract

The purpose of this dissertation is to discuss speech emotion recognition. It was created a validated acted Portuguese emotional speech database, named European Portuguese Emotional Discourse Database (EPEDD), and statistical classification algorithms have been applied on it.

EPEDD is an acted database, featuring 12 utterances (2 single-words, 5 short sentences and 5 long sentences) per actor and per emotion, 8 actors, both genders equally represented, and 9 emotions (anger, joy, disgust, excitement, fear, apathy, surprise, sadness and neutral), based on Lövheim's emotion model. We had 40% of the database evaluated by unexperienced evaluators, enabling us to produce a validated one, filtering 60% of the evaluated utterances. The full database contains 718 instances, while the validated one contains 116 instances. The average acting quality of the original database was evaluated, in a scale from 1 to 5, as 2,3. The validated database is composed by emotional utterances that have their emotions recognized on average at a 69,6% rate, by unexperienced judges. Anger had the highest recognition rate at 79,7%, while disgust had the lowest recognition rate at 40,5%.

Feature extraction and statistical classification algorithms were performed respectively applying Opensmile and Weka software. Statistical classification algorithms operated in the full database and in the validated one, best results being obtained by SVMs, respectively the emotion recognition rates being 48,7% and 44,0%. Apathy had the highest recognition rate: 79.0%, while excitement had the lowest emotion recognition rate: 32.9%.

**Key Words:** theories of emotion, speech emotion recognition, emotional speech, speech database, statistical classification, SVM, Random Forests, ANN

# TABLE OF CONTENTS

Chapter Page

# LIST OF TABLES

Table Page

# CHAPTER 1: Introduction and Historic Context

## 1.1. The context of speech emotion recognition

The development of machine-learning is opening a window of new opportunities, the most popular ones being self-driving cars, fraud detection and recommendation algorithms, such as those developed by Spotify, Netflix, Amazon, Facebook. This dissertation will cover the field of speech emotion recognition through machine-learning, connecting the field of machine-learning, emotion expression theory, linguistics and low-level audio descriptors.

Machine-learning is a technique to build models, through database pattern recognition. The idea is to have an algorithm that will analyze data in order to identify a pattern or to create a model. Machine-learning algorithms enables tasks such as recommendation, prediction or classification. For instance, our objective (speech emotion recognition) is classification and we must climb the following four steps:

- The first step is to create a database containing several recorded audio speech files, each labeled with an emotion.
- The second step is to create a database containing several instances, each being a vector, its entries being feature values extracted from each audio speech file from the previous database.
- The third step is to have machine-learning algorithms analyze the previous database in order to discover a pattern connecting labeled emotions to recorded audio speech features.
- The fourth step is to have that algorithm classify the emotion of an audio recorded speech.

More exactly, our objectives are the creation of a Portuguese acted emotional speech database based on Lövheim's emotion model, and the statistical classification of emotional speech based on it. We will thus also discuss in depth database development methodology and, in particular, acted emotional speech database one.

## 1.2. The Structure of this dissertation

This dissertation contains five chapters: introduction and historic context, background and literature review, methodology, results discussion, and conclusion. In the next sections of this introductory chapter, we will discuss the pragmatics of speech emotion recognition, we will introduce the idea of an emotion and the different types of emotional speech databases. The following chapter will then address the state-of-the-art in emotion expression theory, speech emotion database development, audio feature extraction and statistical classification algorithms. The third chapter will address the methodology applied to build an acted Portuguese emotional speech database. In the fourth chapter, we will present and discuss the results of the quality of this database and the performance of several statistical classification algorithms based on it. Finally, the last paragraph is the conclusion of this dissertation: it is a reflection on our approach and results on emotional speech database creation and recognition, while suggesting future research on this field.

## 1.3. The pragmatics of speech emotion recognition systems

It is acknowledgeable that speech is the fastest and simpler human communication method; and, in addition to the message, it contains para-linguistic information, such as speaker, language, emotion, gender, age, etc. It is, therefore, highly desirable that human-computer interaction is performed through speech. The machine can be a laptop, a humanoid robot, a robotic pet, a phone, a game console, a car, etc. Effectively, since the late 50s there has been tremendous research on speech recognition. (El Ayadi, Kamel, & Karray, 2011) Speech recognition systems performance is however limited by its incapacity to work under real-life environments: noisy rooms, emotional speech, etc. (Koolagudi & Rao, 2012)

Furthermore, the semantics of any message is drastically altered by its para-linguistic information (emotions, speaker, age, etc.) making para-linguistic speech recognition desirable. For instance, the word “Okay” in English is used to express consent, disbelief, boredom or assertion. Therefore, speech recognition systems, not only have to be nose proof, but must be able to recognize para-linguistic information, specially emotions. Ideally, these systems ought to be language, dialect, social class, culture, gender, age and speaker independent.

It turns out that speech emotion recognition, aside from being useful to a more natural human-machine communication, has several other applications; Koolagudi et al. reported an extensive list of speech emotion recognition pragmatics, including the following ones: (Koolagudi & Rao, 2012)

(1) To keep drivers alert, using an onboard car driving system.

(2) To improve the quality of call-center services, analyzing their clients’ emotions during conversations.

(3) To create emotion-interactive movies, games and E-tutoring: these would be more interesting, should they adapt themselves to the listener’s or student’s emotional state.

(4) To index music or video by emotional content.

(5) To analyze recorded or telephone conversations between criminals.

(6) To analyze emergency services calls, evaluating the genuineness of requests; this is a helpful tool for fire brigades and ambulance services.

(7) In aircraft cockpits, systems trained to stressed speech achieve better performance than those trained by normal speech.

(8) To develop automatic natural speech to speech translation systems.

Speech emotion recognition pragmatics include also the following ones:

(9) To develop lie detectors; one example being the commercialized X13-VSA PRO Voice Lie Detector 3.0.1 PRO. (Ramakrishnan, 2012)

(10) To link languages by emotional expression similarity. (K. Scherer, 2000)

(11) To enable robot-robot communication through expression, enabling thus humans to understand what robots are communicating. (Crumpton & Bethel, 2016)

(12) To diagnose psychological disorders, such as depression. (France et al., 2000)

(13) To sort voice mail by emotional content. (Ramakrishnan, 2012)

(14) Finally, to develop robots that perform as companions, tutors and caregivers. (Crumpton & Bethel, 2016)

## 1.4. What is an Emotion?

What is an emotion? One needs a written definition for every used word, one might think. However, pragmatically, much as a word has not a written definition, if it is understood by everybody, there is no need to define it. Thus, for example, even though a scent would not be defined as what it is perceivable through the nose, for one day, mankind will probably be able to experience scents without one, that is, through some neural stimulation, we all know what a scent is: and in the same fashion, we intuitively know what an emotion is. An emotion is a phenomenon, such as a scent or a sound is.

Scents, sounds and emotions are phenomena; however, emotions distinguish themselves from the other phenomena, for it is not unusual that two persons approximately at the same moment and location will experience very similar scents, sounds or visions, but drastically different emotions. This is because this phenomenon reflects an interior state in contrast with the other previously stated phenomena that reflect exterior properties. And, in contrast to phenomena like hunger that do also reflect an interior state, emotions are related to highly potential actions. Therefore, the study of emotion expression will reveal potential actions or action patterns, also known as personality.

## 1.5. An introduction to emotional speech databases types

An emotional speech database is a collection of emotional utterances, each labelled with an emotion. And these are classified according to the method used to create them. In this section, the classification of emotional databases is discussed.

Emotional speech databases can be classified in many ways. For instance, Cowie considers five different issues: naturalness, scope, descriptors context and accessibility. (Douglas-Cowie, Campbell, Cowie, & Roach, 2003) We will introduce in this chapter the naturalness and the emotion descriptors issue. Then in the next chapter, we will discuss context. In the third chapter, while addressing the method used to create the EPEDD, we will address scope and availability.

### 1.5.1. Naturalness

In terms of naturalness, emotional speech databases are classified into three major categories: acted, induced and natural. An acted emotional database is defined as one in which the emotional speech was achieved through acting. An induced emotional database is defined as one in which the speakers have their emotions stimulated by controlled external factors, such as, for example, a movie, a picture, etc. Finally, a natural emotional speech database is defined as one in which the utterances are naturally occurred conversation recordings, such as some TV or radio interview, or a cockpit conversation of a crashed airplane, etc. (Koolagudi & Rao, 2012; Ramakrishnan, 2012; K. R. Scherer, 2003; Ververidis & Kotropoulos, 2006) As it is expected, each database type has its own advantages and disadvantages; these will be studied in the following sections.

#### *1.5.1.1. Natural emotional databases*

Natural (or spontaneous) emotional databases can be obtained in a myriad of ways, for example, using TV shows, call center conversations, or even recorded cockpit conversations of crashed airplanes. As an example, the Belfast database used TV shows. (Douglas-cowie, Cowie, & Schröder, 2000)

Apart from the obvious advantage of natural emotional databases being genuine, and thus, our reference, there are some other advantages, however more implicit.

One of the implicit advantages is the possibility of studying emotion expression on different contexts. Section 2.3.2. being exclusively dedicated to the study of the context, we will return to the discussion of this advantage then.

These databases are, on the one hand, reported to be characterized by featuring multiple concurrent emotions—a mixture of basic emotions. (Cowie & Cornelius, 2003; Koolagudi & Rao, 2012) The experience of a mixed emotion does not necessarily imply that its expression is also a mix of the expressions of both of the basic emotions. Should the expression of mixed emotions be a mix of the expression of each of the emotions, then one emotional speech database is completed if it features only the basic emotions, mixed emotions being redundant. In fact, another advantage of natural speech emotion databases is that these can feature many different emotions. In contrast, in acted speech emotion databases, the more emotions an actor is requested to enact, the better he must enact each one of them, so as to create a clear distinction between them. On the other hand, these databases are reported to predominantly feature low intensity emotions (Cowie & Cornelius, 2003): this is because speakers not always express their emotions conspicuously. Applying the same argument as above, a speech emotion database covering many degrees of should ideally be a natural one.

What are the disadvantages of a natural emotional speech database? A first disadvantage is the effort to create one (Douglas-Cowie et al., 2003): firstly, for ethical reasons, one ought not to record someone without their consent. (Koolagudi & Rao, 2012) Secondly, when a speaker knows that they are being recorded, the context being different, the way a speaker expresses their emotions might differ too. In fact, the creators of the Belfast project scanned through a myriad of TV interview shows, and found only a few in which the speakers had not been affected by the fact that they had been recorded. (Douglas-cowie et al., 2000) Thirdly, all emotions might not be covered; and, even if all would be covered, some could still not be sufficiently represented. (Koolagudi & Rao, 2012)

Another disadvantage is related to the quality of the utterances. The microphones are not necessarily high quality, and if using utterances from different sources, they probably will not be the same. Furthermore, these recordings will usually be to a certain extent drowned in noise: a possible major problem when later dealing them with classification algorithms. Finally, the overlapping of utterances being impractical for analytical proposes, the number of emotional utterances is reduced. (Koolagudi & Rao, 2012)

Last but not least, in this paragraph, we briefly review some natural speech emotional databases. Cowie et al. created the Belfast database, which is a natural and induced one using 201 clips from TV interview shows. (Douglas-cowie et al., 2000) France et al. worked with multiple natural databases featuring normal subjects, major depressed patients and high-risk suicidal patients. (France et al., 2000) Lee and Narayanan created one using call center conversations. The TUM AVIC database was created featuring five non-linguistic vocalizations. The VAM database was created featuring 47 talk show guests and 947 utterances (approximately 12 hours), applying a three-dimensional emotion theory. The RECOLA database was created featuring five social behaviors (agreement, dominance, engagement, performance, rapport) and using a two-dimensional emotion theory (approximately 7 hours). Finally, the AViD-Corpus was created so as to study minimal depression, mild depression, moderate depression and severe depression. (Valstar et al., 2013)

#### 1.5.1.2. *Induced emotional databases*

An induced (or elicited) emotional database consists in natural data created artificially. The idea is to gain more control over the data, at a cost in genuineness, while improving the quality of the speech material.

The increase in quality is largely due to the fact that the moment and the location of the recording can be chosen. So, the location can be a sound proof studio, equipped with a high-quality microphone. The microphone can also be set up at the same amount of gain and located at the same distance from the speaker, during the entire experiment.

When it comes to the available emotions, though the researcher will still not have a complete control over them, they will influence them. Some methods are, for example, making a speaker watch a movie or listen to a song, or engaging the speaker in a conversation about his dreams, his achievements, his losses and sorrows, his loves, etc. However, the use of a stimulus might not be straightforward: one stimulus may lead to different emotional reactions, depending on the speaker. Also, therefore, mixed emotions, though they may or may not be wanted, they will still be expressed.

A big drawback on this approach is the hypothesis that speakers, since they know that they are being recorded, might express themselves less genuinely. And the utterances might even contain acoustic features specific to the acknowledgment of being recorded. (Koolagudi & Rao, 2012)

Last but not least, in this paragraph, we briefly review some induced speech emotional databases. The FAU AIBO database was created featuring 51 school children interacting with Sony AIBO pet robot (approximately 9 hours) (Batliner, 2004). The eNTERFACE database was created featuring 42 subjects, in a total of 1166 video sequences, exploring six discrete emotions (anger, disgust, sadness, happiness, surprise).

#### 1.5.1.3. Acted emotional speech databases

Acted (or simulated, or posed) emotional speech databases advantages are a product of the control the research has over them. Not only is the researcher able to choose the space and the time of the recordings, as it was the case for the induced emotional speech database, but the researcher can also decide which sentences the actors must utter too. Thus, the researcher has control over the phonetic characteristics of the sentences and can have a dozen of actors uttering the same set of sentences. This is a major advantage, for it enables an accurate comparison between different emotional expressions. Moreover, the researcher is also able to decide which emotions are to be recorded; thus, the researcher can have each sentence, preferably emotionally neutral, repeated over each emotion, so as to achieve great comparability.

In contrast, much as this approach is about control, acted emotions probably might never sound as natural emotions do. That is because the actor might probably overact certain acoustic features, while ignoring more subtle ones. And that is also because when a speaker is feeling a certain emotion, for example, being in love with someone, or anxiously struggling over a personal issue, that speaker, when acting, might slightly unconsciously express his true emotion. An interesting question that arises here is what does the actor think and feels when he acts. For, the closer he feels the desired emotion and the closer he thinks like he would think, experiencing that same emotion, the more accurate might the acted emotional speech database be. Therefore, the better the acting, the more natural the database is. And, an interesting experience that is yet to be done is to have a mixed database, featuring both natural and acted emotions, and having them being judged by experienced judges; the judges would judge whether each clip is acted or natural. Such test will reveal whether there are specific acting acoustical features and probably it will reveal that the better an actor is, the fewer will be the specific acting acoustical features.

Therefore, the Stanislavski method is the recommended acting method and in fact a commonly used one. (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005; K. Scherer, 2000; Staroniewicz & Majewski, 2009) The idea is to remember a distant memory in which the desired emotion had been felt, in an attempt to feel again that emotion. The actor, feeling the requested emotion, will be more personally involved and thus the acting will be more natural. (Sawoski, 2006) And, evidently, actors are free to be more creative by self-inducing themselves the requested emotions, for example, by listening to music—this inducing method differs from the inducing method that might be applied on an induced emotional database, because on these, speakers, instead of being self-induced, they are induced by the researchers.

Typically, in acted emotional speech databases, the emotional utterances are not associated with a context, since it requires a longer recording time, larger budget and a greater acting expertise; an exception is the GEMEP corpus, featuring different contexts for each emotion. (Banziger, Pirker, & Scherer, 2006)

We don’t briefly mention any acted emotional speech database in this section, for we do that already in section 2.3.1.

### 1.5.2. Descriptors

An emotional speech database being a set of utterances, each labelled with an emotion, one of the steps when creating one is to label each utterance to its respective emotion. There are two paradigms on emotional descriptors: they are either discrete or continuous. (Douglas-Cowie et al., 2003) Though it is easier to label emotions discretely, continuous descriptors offer more detail. For natural and induced emotional speech databases, featuring emotions that do not necessarily fall entirely into any of the discrete categories, instead falling ambiguously into more than one category, should be described continuously. For acted emotional speech databases, discrete emotional descriptors are the norm; however, the use of continuous descriptors might result in a more detailed emotion description.

# CHAPTER 2: Background and Literature Review

Speech emotion recognition is a product of four branches of knowledge: emotion expression theory, database creation methodology, low-level audio descriptors and statistical classification algorithms. This chapter will begin with an overview of the development in speech emotion recognition and then will discuss, in particular, the state-of-the-art of those four fields, in that order.

## 2.1. An Overview of Developments in Speech Emotion Recognition.

The first major work on emotion expression dates back to 1872, with Darwin's work "The Expression of the Emotions in Man and Animals". (Darwin, 1872) One century afterwards, Ekman's and Friesen's famous publication of the Facial Action Coding System manual in 1977 teaches the readers to interpret facial expression of emotions. (P. Ekman & Friesen, 1975) Interestingly, both Darwin and Ekman believed evolution played an important role on the expression of emotions, making them rather universal. More recently, since the 90s, with the developing of robust machine-learning systems, emotion recognition systems have been being developed.

Schuller et al. summarize the research development on speech emotion recognition, "We can sub-categorize the time-line of this field during the last 15 years into three phases: some spurious papers on recognition of emotion in speech during the second half of the 90s (less than 10 per year), a growing interest until 2004 (maybe some 30 per year), and then, a steep rise until today (>100 per year)." (Schuller, Batliner, Steidl, & Seppi, 2011)

In particular, in the 90s, emotion recognition research began with the study of both facial and speech emotion expressions individually. For example, Nicholson et al. obtained a 50% speech emotion recognition rate using eight different emotions. (Nicholson, Takahashi, & Nakatsu, 1999) But multimodal study of emotion expression, encompassing both facial and speech emotion expressions has rapidly caught attention, leading to improved emotion recognition performance; for example, Chen et al. obtained 97% recognition rate using an audiovisual database, whereas respectively 70% and 75% for visual and audio databases alone. (L. S. Chen, Tao, Huang, Miyasato, & Nakatsu, 1998) Since the 2000s, multimodality has been being strongly encouraged, for most of the emotion recognitions systems developed still performed on an uni-modality perspective. (Douglas-Cowie et al., 2003) Multimodality embraces studying simultaneously facial expression, speech, gestures and any other possible modalities. The encouraged idea was to create richer, bigger and better emotion databases: thus, not only were researchers encouraged to approach multimodality, they were also encouraged to feature more languages, dialects, cultures, emotions, instances and to be more realistic. Another common concern, aside the development of richer, bigger and more realistic databases, was to focus on an increasing number of features and the exploration of different classifiers. All these concerns that accompanied the development of emotion recognition were emphasized by multiple challenges that have oriented researcher's work during the last decade:

Since the first speech emotion challenge appeared in 2009 (Schuller, Steidl, & Batliner, 2009), many challenges have been created. These tend not only report the previous studies on emotion recognition, but also to reflect the contemporary needs on this field and to challenge researchers on these. For example, the Interspeech 2010 Paralinguistic Challenge challenged researchers on paralinguistic studies such as the retrieval of age and gender through speech. (Schuller, Steidl, Batliner, Burkhardt, et al., 2013) Since then, Interspeech has been realizing a challenge on speech information retrieval every year. (Schuller et al., 2014, 2015; Schuller, Batliner, Burgoon, & Coutinho, 2016; Schuller, Steidl, Batliner, Vinciarelli, et al., 2013; Schuller, Steidl, et al., 2012; Schuller, Steidl, Batliner, Schiel, & Krajewski, 2011) Furthermore, ACM has been challenging researchers to work on emotion and depression recognition through facial and speech modalities since 2011, releasing thus the Audio/Visual Emotion Challenges (AVEC); their aim is to investigate interaction between different modalities and to compare audiovisual signal processing and machine learning methods to advance recognition systems. (Ringeval, Schuller, Valstar, Cowie, & Pantic, 2015; Schuller, Valstar, et al., 2011; Schuller, Valstar, Eyben, Cowie, & Pantic, 2012; Valstar et al., 2013, 2016; Valstar, Schuller, Jarek, Cowie, & Pantic, 2014) Finally, ICMI has also been challenging emotion recognition researchers since 2013 to work on real world observations, so as not to be limited to studio situations, releasing thus the Emotion Recognition in the Wild (EmotiW) challenges. (Dhall, Goecke, Joshi, & Gedeon, 2014, 2015; Dhall, Goecke, Joshi, Hoey, & Gedeon, 2016; Dhall, Goecke, Joshi, & Wagner, 2013) Last but not least, these challenges have been encouraging standardization in research and avoiding overlapping research, a problem that had previously been reported.

## 2.2. A review of emotion theories

Our main interest is the recognition of emotions in human speech. Therefore, we need a certain number of different emotions to recognize. And a basic understanding of each emotion.

Some researchers ventured into listing emotion-related words. Cowie reviews that in English, 196 emotion words were found; in German 235 emotion words were found; in Italian, 153 emotion words were found. Cowie argues that generally languages do not multiply terms, unless needed. His example is that we usually struggle on grasping an emotion term to express how a certain work of art makes us feel. (Cowie & Cornelius, 2003) Therefore, one expects that a complete emotion theory should embrace a huge number of emotions. However, for emotion recognition purposes, a dramatically smaller number of recognizable emotions would be enough for a first approximation on this task.

There are many emotion theories, each featuring its own emotions and relations between them. We cannot cover them all, however, they can be classified through their similarities and differences. We will review discrete emotion theories and dimensional emotion theories.

### 2.2.1. Discrete emotions theories

Discrete emotion theories are emotion models in which the emotions do not share fundamental characteristics in common. (Paul Ekman & Cordaro, 2011) There are multiple theories of discrete emotions. Russell and Barrett review 7 different types of discrete emotion theories. (Russell & Barrett, 1999)

Often these models feature a concept reminiscent of the discrete emotion, the so-called basic emotion (or primary emotion or pure emotion). (Plutchik, 1991) Much as basic emotions are understood differently by different researchers, they always represent a subset of discrete emotions, developing thus further the structure of emotions. The general idea is that the different possible combinations of the basic emotions would create all the discrete emotions. We will review three different discrete emotion theories.

Ekman developed his own discrete theory: he supposes there is a small number of families of emotions, each family being a vast collection of emotions very similar to each other. Five of his families of emotions are anger, sadness, enjoyment, fear and disgust. He develops that anger, for example, includes berserk, indignation, vengeance, rage, etc. (Paul Ekman, 1993) He argues that those variations are social, in contrast to the families themselves, which are phylogenetic. Additionally, he defines basic emotions as discrete emotions that are evolutionary adaptations to the environment, in origin. For instance, he does not consider mood as emotion, considering moods as long-term phenomena that do not share the adaptive evolutionary quality of basic emotions. (Paul Ekman & Cordaro, 2011) His argument for the families of emotions arises from his study of human facial expressions, in which facial expressions of emotions within one family are relatively similar, whereas facial expressions of emotions from others families drastically differ. (Paul Ekman, 1993) Furthermore, Ekman is particularly known for his work on the human facial expression of the following 6 basic emotions (or Ekman's Big 6 Emotions): anger, sadness, fear, happiness, disgust and surprise. (P. Ekman & Friesen, 1975)

Tomkins' discrete model features eight basic emotions. These basic emotions are experienced with a variable intensity level. Tomkins labeled each of his eight basic emotions with two different names, one labeling a weaker intensity form of the same basic emotion. Writing the weaker form of the emotion first, these are: shame/humiliation, anger/rage, distress/anguish, contempt/disgust, fear/terror, surprise/startle, enjoyment/joy and finally, interest/excitement. (Lövheim, 2012; Tomkins, 1981)

Plutchick's model also features eight basic emotions, these being fear, anger, sadness, acceptance, disgust, joy, expectation and surprise. However, each basic emotion has its own opposite basic one too: surprise contrasts with expectation, disgust contrasts with acceptance, sadness contrasts with joy and anger contrasts with fear. And, in contrast with the previous model, their intensity is labelled with three words instead of only two, as it was the case in Tomkins' model. Moreover, Plutchick goes even further by representing each of the eight emotion on a circle, turning the model circumplex: this representation allows not only to express that each basic emotion is opposite to another one, but it also allows to express that each basic emotion is close to two others. Then, for each two adjacent basic emotions, Plutchick, combining these two, adds their combinations between these. Plutchick, therefore, is able to classifying 32 emotions. (Plutchik, 2001)

### 2.2.2. Dimensional emotion theories

Dimensional emotion theories are those that represent all emotion in a vector space. There are many dimensional emotion theories, each having its own set of vectors. Russell and Barrett reviewed several dimensional models, ranging from one to multiple dimensions. (Russell & Barrett, 1999)

This classification can be dated back to Wilhelm Wundt, when he proposed a three-dimensional emotion theory, his dimensions being pleasure versus displeasure, arousal versus subduing, and strain versus relaxation. (Wundt, 1897)

Schlosberg proposed a three-dimensional model featuring circumplexity, half a century later; his dimensions being pleasantness versus unpleasantness, attention versus rejection and activation. These dimensions are similar to Wundt's dimensions, the circumplexity being thus the difference between these. In contrast to arousal, activation implies not only activity, but also reactivity, as Schlosberg argues expresses, in order to explain his choice of the word "activation". (Schlosberg, 1954)

Half a century later, Lövheim developed a model uniting Tomkins' eight basic emotions with three monoamine neurotransmitters: dopamine (DA), serotonin (5-HT), and noradrenaline (NE). Representing the quantity of these three neurotransmitters respectively with three axes and assuming that each neurotransmitter has a maximum quantity, one obtains eight extremes states. So, basically, Lövheim argues that to each of these eight states corresponds one of Tomkins' eight basic emotions. One of Lövheim arguments to support his theory was especially interesting, because it focused on the effects of each of these three neurotransmitters on our personality. Explaining that Dopamine is responsible for pleasure and addiction, whereas serotonin is responsible for a sense of superiority and, finally, noradrenaline is responsible for arousal, attentiveness and activity, Lövheim argues the following: (Lövheim, 2012)

**Table 1: Relation between emotions and neurotransmitters in Lövheim's emotion model**

| | **Noradrenaline** | **Dopamine** | **Serotonin** |
|---|---|---|---|
| **Shame** | Low | Low | Low |
| **Contempt** | High | Low | Low |
| **Fear** | Low | High | Low |
| **Anger** | High | High | Low |
| **Distress** | Low | Low | High |
| **Surprise** | High | Low | High |
| **Enjoyment** | Low | High | High |
| **Interest** | High | High | High |

Whereas Tomkins had only listed eight basic discrete emotions, Lövheim transforms this model into a dimensional emotion classification, interconnecting them. Again, these dimensions are similar either with Wundt's or Schlosberg's dimensions. Because this model unites neurology, a set of basic emotions and a dimensional theory, we will use this one for our own research.

Finally, it is worth noting that the concept of basic emotions, previously defined in a discrete emotion context, can be brought into a dimensional emotion context: assuming that each dimension has a limit, then a basic emotion would be one that lies on one of the many vertices of a model. Therefore, in a two-dimensional model, there are four basic emotions; and, in a three-dimensional model, there are eight basic emotions.

### 2.2.3. Appraisal emotion theories

More recently, a new perspective on emotions has been gaining attention; this perspective, known as appraisal theories, developed the concept of emotion as a process, instead of emotion as a state. The first papers introducing this theory date back to the 60s with M. B. Arnold and R. S. Lazarus being the pioneers. Appraisals cannot be categorized in any of the above mentioned theories, for they can be both discrete or dimensional, depending on the author. (Moors, Ellsworth, Scherer, & Frijda, 2013) Within a dimensional perspective, appraisals have many dimensions, these being novelty, valence, goal significance, cause, and norms/values legitimacy; and, in contrast with previous dimensional theories, this new perspective not only affords a description of the subjective experience of emotion, but also explains its cause. Furthermore, in an appraisal context, emotions are understood as action tendencies or intuitive actions, an organism's more flexible approach to problem solving, in contrast to a deterministic approach. Moreover, in an evolution context, emotions, understood as action tendencies, can be considered adaptive. (Ellsworth, P. C., & Scherer, Ellsworth, & Scherer, 2003)

## 2.3. A review on Emotional Speech Databases

We have decided to create an acted emotional speech database. Therefore, for more information on these, for example, on acting technique, on the choice of actors, on the choice of the text material and other important questions, read subsection 3.1 in the methodology section. We will now begin by discussing the scope and, in the following section, we will discuss the context.

### 2.3.1. Emotional acted speech databases scope

For the purpose of reviewing acted speech emotional databases and studying their particular scope, we produced table 2. Furthermore, tables 3 to 8 review their respective evaluation.

**Table 2: Acted speech emotion databases review**

| Reference | Speakers | Utterances | Language | Emotions | Evaluation | Name |
|---|---|---|---|---|---|---|
| (Burkhardt et al., 2005) | 5 actors and 5 actresses | 5 short sentences and 5 long sentences | German | 7 emotions | By 20 subjects | emoDB |
| (Engberg & Hansen, 1996) | 2 actors and two actresses | 9 sentences, 2 single words and 2 passages | Danish | 5 emotions | By 10 subjects | DES |
| (Jovi, Ka, & Rajkovi, 2004) | 3 actors and 3 actresses | 30 short sentences, 30 long sentences, 32 single words and 1 passage | Serbian | 5 emotions | By 30 subjects | GEES |
| (Staroniewicz & Majewski, 2009) | 7 amateur actors and 6 amateur actresses | 10 sentences | Polish | 7 emotions | By both genders; and by musicians and non-musicians | |
| (Banse & Scherer, 1996) | 6 actors and six actresses | 2 sentences | German | 14 emotions | | |
| (Lima, Castro, & Scott, 2013) | 2 males and 2 females | Nonverbal vocal expression | Does not apply | 8 emotions | By 20 subjects | |
| (Castro & Lima, 2010) | 2 women | 16 short sentences and 16 short pseudo-sentences | Portuguese | 7 emotions | By 20 subjects | |

As it can be remarked, acted emotional speech databases usually feature less than a dozen of actors, both genders equally represented. And, aside from Jovi et al. database, others usually feature no more than a dozen of utterances to be uttered by all actors under, aside from Banse and Scherer's database, a maximum of 8 emotions. Furthermore, their evaluation is usually performed by 20 subjects. Large acted databases, featuring dozens of utterances to be uttered by dozens of actors, remain uncharted territory.

**Table 3: emoDB evaluation results**

| Emotion | anger | neutral | fear | boredom | happiness | sadness | disgust |
|---|---|---|---|---|---|---|---|
| Recognition rate | 96.9% | 88.2% | 87.3% | 86.2% | 83.7% | 80.7% | 79.6% |

Source: (Burkhardt et al., 2005)

**Table 4: DES evaluation results**

| Emotion | neutral | surprise | happiness | sadness | anger |
|---|---|---|---|---|---|
| Recognition rate | 60.8% | 59.1% | 56.4% | 85.2% | 75.1% |

Source: (Engberg & Hansen, 1996)

**Table 5: GEES evaluation results**

| Emotion | anger | neutral | happiness | fear | sadness |
|---|---|---|---|---|---|
| Recognition rate | 96.1% | 94.7% | 94.7% | 93.3% | 96.0% |

Source: (Jovi, Ka, & Rajkovi, 2004)

**Table 6: Staroniewicz et al. evaluation results**

| Emotion | happiness | anger | fear | sadness | surprise | disgust | neutral |
|---|---|---|---|---|---|---|---|
| Recognition rate | 68.2% | 71.1% | 40.5% | 44.7% | 72.5% | 30.4% | 73.4% |

Source: (Staroniewicz & Majewski, 2009)

**Table 7: Banse et al. evaluation results**

| Emotion | hot anger | cold anger | panic fear | anxiety | despair | sadness | disgust |
|---|---|---|---|---|---|---|---|
| Recognition rate | 78% | 34% | 36% | 42% | 47% | 52% | 15% |
| Emotion | elation | happiness | interest | boredom | shame | pride | contempt |
| Recognition rate | 38% | 52% | 75% | 76% | 22% | 43% | 60% |

Source: (Banse & Scherer, 1996)

**Table 8: Lima et al. evaluation results**

| Emotion | achievement | amusement | pleasure | relief |
|---|---|---|---|---|
| Recognition rate | 77,7% | 95,9% | 85,9% | 86,3% |
| Emotion | anger | disgust | fear | sadness |
| Recognition rate | 78,3% | 96,7% | 70% | 89,7% |

Source: (Lima et al., 2013)

**Table 9: Lima et al. evaluation results**

| Emotion | anger | neutral | fear | surprise | happiness | sadness | disgust |
|---|---|---|---|---|---|---|---|
| Sentence recognition rate | 77% | 88% | 75% | 87% | 75% | 84% | 50% |
| Pseudo-sentence recognition rate | 74% | 83% | 56% | 85% | 59% | 82% | 60% |

Source: (Castro & Lima, 2010)

Highest human recognition rates were obtained by Burkhardt et al. and Jovi et al, the former results being particularly remarkable since the emoDB features 7 emotions. On average anger is the highest recognizable emotion, while disgust is the least recognizable emotion. It is worth noting that Banse and Scherer's database is particularly interesting for featuring as many as 14 different emotions, the least recognizable emotion still being two times higher than random guess (7%). Finally, Lima et al. study reveals that pseudo-sentences (nonsensical sentences that are highly resemble language) emotions were less recognizable. This is maybe due to the fact that these are harder to portray, since they have no meaning and sounding thus bizarre.

### 2.3.2. Emotional speech database context

An elegant way to define context is to define it as the variable that makes the same emotion being expressed differently, by the same person. For example, one does not express his emotions on a business reunion as one does on a family dinner, even if subject to the same emotion. Furthermore, we can remark that the effect of the context on the expression of emotions can be understood as the degree of concealment of emotions from other people. In fact, the idea of Ekman's family of emotions is, on the one hand, a result of the influence of context—emotions may be being concealed, discerning thus hot anger from cold anger, for example (K. R. Scherer, 2003)—and, on the another hand a result of the influence of intensity. Now, in another paper, Cowie reviews evidence that "listeners use context to determine the emotional significance of vocal features". (Douglas-Cowie et al., 2003) Therefore, if developing an emotional database, one ought to always specify the context. Unfortunately, much as this has been done when creating natural emotional databases, the context has been ignored when dealing with acted emotional speech databases.

Cowie identifies four types of context: semantic context, structural context, intermodal context and temporal context. The semantic context relates refers to the way certain words have an intrinsic emotional meaning; the author states that there is room for semantic and vocal signs interaction. Secondly, the structural context refers to the way syntax and prosody acts as a medium for emotional expression, through repetitions, interruptions, long or short sentences, intonation, stress patterns, etc. Thirdly, the intermodal context refers to the influence of modalities being used in communication: for instance, as Cowie points out, telephone conversations carry all emotional content through audio. Therefore, audio alone is able to convey a rich emotional meaning. Finally, the temporal context refers to the way emotion "ebbs and flows over time", studying, for example, the emotional build-up occurrences. (Douglas-Cowie et al., 2003)

These four types of contexts are methods that we, humans, may consciously or unconsciously use to mask or pretend emotions. What causes us to pretend to be feeling other emotions? There are three causes, each related to the speaker's culture: the speakers and spectators, the local and the time of the day.

(a) Depending on whom is the speaker engaging a conversation with, he may adapt his emotional expression: some examples of causes are teacher-student dialogues, friendship conversations, family conversations and conversations between strangers. Moreover, the presence of other people, though not speaking, might influence the emotional expression. Hot anger is more easily bluntly expressed if there are no nearby listeners, whereas, should there be any listener, the speaker would rather certainly opt to a cold anger.

(b) Depending on the location the conversation is taking place, the conversation may conceal more or less their emotions. For example, in a church, at the parliament, at work, speakers may feel the need to repress their emotions.

(c) Finally, at night, so as not to wake up whoever may be asleep, emotions might be concealed.

Then, the emotional speech database state-of-the art methodology having been discussed, we will now turn our attention into feature extraction and pattern recognition, the final processes in speech emotion recognition.

## 2.4. Feature Extraction

The process of feature extraction can be divided into three parts: pre-processing, feature extraction and post-processing. The first part, pre-processing, corresponds to the process of de-reverberation (room reverberation reducing) and de-noising, should they be needed; we will not cover these. Instead, we will cover feature extraction and post-processing, starting with the difference between local and global features, moving then to the different types of features, then feature normalization and finally feature selection. Once having understood feature selection, we will cover the field of pattern recognition in section 2.4.

### 2.4.1. Local Features and Global Features

Features can be extracted either locally or globally. Local features are extracted by dividing the whole clip into small intervals, called frames, and then extracting a feature vector from each one, while global features are extracted globally from the whole clip. Therefore, global features are fewer in number and they do not include temporal information. In terms of accuracy, the use of global features hinder classification models that require a large number of features, such as the hidden Markov model (HMM) or the support vector machine (SVM), from working properly. Furthermore, global features are only efficient in distinguishing high activation/arousal emotions from low activation/ arousal ones. (El Ayadi et al., 2011)

Special cases of local feature extraction are when each frame corresponds to a phoneme or to a voiced speech segment. Local feature extraction per phoneme requires robust phoneme segmentation algorithms and enable to compare phonemes under different emotions. In contrast, global feature extraction is easier to implement. (El Ayadi et al., 2011)

### 2.4.2. Types of features

Features are categorized into different types; however, each author presents a distinct set of categories. Gangamohan et al. categorize features into prosodic features, voice quality features and spectral features (Gangamohan, Kadiri, & Yegnanarayana, 2016); Koolagudi and Rao categorize features into prosodic features, vocal tract features and excitation source features (Koolagudi & Rao, 2012); Ayadi et al. categorize features into continuous speech features, voice quality features, spectral-based speech features and nonlinear TEO-based features (El Ayadi et al., 2011); finally, Schuller et al. makes the distinguish between acoustical and linguistic features. (Schuller et al., 2011) We will cover prosodic features, voice-quality features, spectral-based, cepstral-based and TEO-based features and finally, linguistic features.

#### *2.4.2.1. Prosodic Features*

This set of features consist on the patterns of intonation, intensity and duration. Prosody is necessary to make human speech natural (Koolagudi & Rao, 2012) and comprehensible; in fact, prosody influences semantics. (Wennerstrom,

2001) Prosody can also express attitudes, such as doubt and assertiveness, while at the same time it can take the role of punctuation at structuring the discourse. (Wennerstrom, 2001) Prosodic features include the fundamental frequency (F0), intensity and timing features such as speaking rate, pause duration, average duration of voiced speech, syllables per second, etc. (Gangamohan et al., 2016)

#### *2.4.2.2. Voice Quality Features*

This type of feature "refers to the characteristic auditory coloring of an individual's speech", each speaker having its own "voice-quality signature". This characteristic is expressed in terms of laryngeal and supralaryngeal settings. Emotions are reported to shape the voice quality; for example, fear induces a harsh voice, angriness and happiness induce a breathy voice, etc. (Gangamohan et al., 2016)

The adjectives used to portray voice-quality are all quite subjective, such as harsh, tense, modal, breathy, whisper, creaky and lax-creaky, hoarse, quavering, ingressive, falsetto, rough, etc. (Douglas-Cowie et al., 2003; Ramakrishnan, 2012; Schuller, Steidl, Batliner, Burkhardt, et al., 2013)

This type of features includes the following features: shimmer, jitter and harmonic-to-noise ratio (HNR). (El Ayadi et al., 2011; Schuller et al., 2009) Shimmer is a value that reflects the "changes in amplitude of the waveform between successive cycles", whereas Jitter corresponds to the "changes in the frequency of the waveform between successive cycles". (Sbattella et al., 2014) Amir et al. calculated the shimmer and jitter respectively by calculating the number of changes in sign of the intensity derivate and of the pitch derivative. Shimmer and jitter are stated to be shaped by age. (Schuller, Steidl, Batliner, Burkhardt, et al., 2013) Finally, the HNR is stated to be a "potential discriminator for the breathy voice".

#### *2.4.2.3. Spectral-based, Cepstral-based and TEO-based Features*

Spectral features, notably, the distribution of the spectral energy across the speech range frequency is shaped by the emotional content. For example, happiness correlates with high energy at the high frequency range, whereas sadness correlates with lower energy at that same range. This type of features, in contrast with previous features, are less tangible, in the sense that they are often expressed with complex mathematical expressions. But, one of the examples of spectral-based features is fairly understandable: "formant frequencies and their respective bandwidths": F1, F2, F3, etc. (Gangamohan et al., 2016)

The other spectral-based features can be extracted applying linear predictor coefficients (LPC), one-sided autocorrelation coefficients (OSALPC), short-time coherence method (SMC), and least-squares modified Yule-Walker equations (LSMYWE). (El Ayadi et al., 2011)

Cepstral-based features can be derived from the corresponding linear features: linear predictor cepstral coefficients (LPCC) are derived from LPCs and OSALPCC are derived from OSALPC. Their efficiency in emotion recognition is still being discussed. (El Ayadi et al., 2011) Other cepstral-based features include mel-frequency cepstral coefficients (MFCC). (El Ayadi et al., 2011)

It is argued that non-linear features are needed to model speech. The Teager-energy-operator (TEO), a non-linear operator, was introduced by Teager and Kaiser. Some of the TEO-based features are TEO-decomposed FM variation (TEP-FM-Var), normalized TEO autocorrelation envelope area (TEO-Auto-Env), and critical band-based TEO autocorrelation envelope area (TEO-CB-Auto-Env). These type of features are reported to outperform others in speech stress recognition. (El Ayadi et al., 2011)

#### 2.4.2.4. *Linguistic Features*

Linguistic features are needed because certain words, sentence constructions and para-linguistic phenomenons such as laughing, crying, sighs, yawns, hesitations, coughs, etc. express emotional content. (Schuller et al., 2011)

### 2.4.3. Feature Combination

It is reported that combining different features enhances the statistical classification performance. (Koolagudi & Rao, 2012) Therefore, it is highly recommended to extract the the maximum number of features possible. In fact, emotion challenges have increased their number of features. (Schuller, Steidl, Batliner, Burkhardt, et al., 2013; Schuller et al., 2009; Schuller, Valstar, et al., 2012; Valstar et al., 2013)

### 2.4.4. Feature Normalization

Feature normalization is a necessary step if the extracted features have different units. The most common method for feature normalization reported to be through z-score normalization (El Ayadi et al., 2011):

$$\bar{x} = \frac{x - \mu}{\sigma}$$

In which $x$ is the value to be normalized, $\bar{x}$ is the value normalized, $\mu$ is the mean and $\sigma$ is the standard deviation.

### 2.4.5. Feature Reduction

Feature reduction is a valuable step: it reduces storage, computational requirement (El Ayadi et al., 2011) and it has been reported that a filtered set of features may enhance the performance of statistical classification. (Schuller et al., 2011) Though initially, feature reduction was designed heuristically, feature reduction algorithms are now commonly used. (Schuller et al., 2011) There are two approaches to feature reduction: feature selection and feature extraction (or feature transformation). In feature selection, a subset of the features is chosen, whereas in feature extraction, the initial features are mapped into a smaller set of features, while preserving as much relevant classification information as possible. (El Ayadi et al., 2011)

Reported to probably be the most common algorithm applied, the sequential forward search starts with an empty set and sequentially adds best features, at each step one or more features being deleted and others being chosen if suited. Another type of feature reduction algorithms is the hierarchical one: instead of optimizing the feature globally, for all emotion classes, it tries to optimize for groups of them. (Schuller et al., 2011)

Linear or Heteroscedastic Discriminative Analysis (LDA) and Principal Component Analysis (PCA) and are the most popular feature transformation methods. On the one hand, LDA is a supervised algorithm and it is limited by its demanding of a least some degree of Gaussian distribution and linear distribution of the input space (Schuller et al., 2011), as well as  by the demanding that the "reduced dimensionality must be less than the number of classes". (El Ayadi et al., 2011) On the other hand, PCA is an unsupervised algorithm and it has the disadvantage of requiring the guess of the dimensionality of the target space. It has been stated that it is not clear whether in fact it performs better than other feature reduction techniques.  (El Ayadi et al., 2011; Schuller et al., 2011)

Finally, also reported to be worth mentioning in a speech emotion recognition perspective, there is the Independent Component Analysis (ICA) and the Non-negative Matrix Factorization (NMF). The former, a feature transformation algorithm, maps the feature space onto an orthogonal space, the target features having the attractive property of being statistically independent. The latter is reported to be mainly applied in large linguistic feature sets. (Schuller et al., 2011)

## 2.5. Pattern Recognition

As our goal is speech emotion recognition, we must label each observation with its respective emotion. In machine-learning, data-mining or pattern recognition, learning can be categorized into different distinct types, such as supervised learning, unsupervised learning, reinforcement learning, etc. In supervised learning, we do know a priori each instance's labels, whereas, in unsupervised learning, we do not know a priori each instance's labels. (Kotsiantis, 2007) Speech emotion recognition is, therefore, supervised learning, for we set a priori the possible emotions; furthermore, because our labels are not numbers, but categories (emotions), this is a statistical classification problem.

Within statistical classification, there are several different steps: the first one is classifier selection, the second one is parameter selection, in which we modify parameters within the classifier; this step is followed by model learning, the one in which the classifier is trained; then, the final step is classification/regression: the moment in which our model is predicting new unlabeled observations. It is worth noting that some models are designed to predict multiple labels: this is multi-tasking learning—because we are interested only in emotion recognition, we will not cover this. (Schuller, Steidl, Batliner, Burkhardt, et al., 2013) Indeed, we will only overview the most used or appropriate classifiers in speech emotion recognition, specially SVMs, Random Forests, HMMs and Artificial Neural Networks (ANNs).

Many different classifiers have been used in speech emotion recognition, for example: hidden Markov models (HMM), Gaussian mixture models (GMM), support vector machines (SVM), artificial neural networks (ANN), decision trees, fuzzy classifiers and k-nearest neighbors (k-NN). However, there is no agreement on which is the optimal classifier, each having its own advantages and disadvantages; and their performance depending on the database and feature selection. Moreover, researchers have combined different classifiers, also known as ensemble learning, so as to take advantage of all their merits. (El Ayadi et al., 2011) Schuller et al. state that an appropriate classifier is one that tolerates high dimensionality, missing data, small data-sets and skewed classes, that solves non-linear problems and that is efficient computationally and on memory costs. But the problem of a high dimensional feature set, leading to regions of the feature space where data is too sparse, also known as "the curse of dimensionality", is reported to be usually better addressed by feature reduction. For example, though the k-NN has been used since the very first studies and turned out to be quite successful for non-acted emotional speech as well, it suffers from "the curse of dimensionality". (Schuller et al., 2011)

HMMs have been widely used in isolated word recognition and speech segmentation. GMMs can be considered as special continuous HMMs which contain only one state; these are more appropriate for emotion recognition in speech when only global features are extracted. In fact, GMMs cannot model the temporal structure of training data. (El Ayadi et al., 2011)

Generally, discriminative classifiers handle better small data sets; therefore, these are optimal for acted speech emotional datasets. Examples of popular discriminative classifiers are ANNs, SVMs and decision trees. (Schuller et al., 2011) We will now review each one of these. ANNs can be categorized into the three following types: multilayer perceptron (MLP), recurrent neural networks (RNN), and radial basis functions (RBF). MLP are relatively common in speech emotion recognition, whereas the latter is rarely being used in this application. However, ANNs are less robust to over fitting and require greater amounts of data, being therefore rarely used in speech emotion recognition (Schuller et al., 2011); in fact, ANNs performance has been reported to be "fairly low in comparison with other classifiers". (El Ayadi et al., 2011)

A SVM can be a linear or a non-linear classification model; its goal is to establish a linear partition of the feature space into two categories. And, if not possible to linearly separate the feature space, the model applies the "kernel trick", mapping the feature space into a higher dimensional one in which a linear separating hyperplane exists. Since only a subset of the training points—those close to the hyperplane—are used to train the classifier, this is a memory and computation efficient one. (Schölkopf, Burges, & Smola, 1998; Vapnik, 1998) Unfortunately, there is not a systematic method to find a desirable such a kernel function, leaving this method to heuristics. Moreover, kernels might induce overfitting, making this process more complicated. (El Ayadi et al., 2011) Therefore, usually a hyperplane that does not perfectly separate the classes is preferable, leading to a greater robustness to the addition of new observations (Schölkopf et al., 1998; Vapnik, 1998). Still, SVMs are widely used in speech emotion recognition and their performance is reported to be familiar (El Ayadi et al., 2011).

Decision trees are also a non-linear classification model. In contrast to SVMs and ANNs, they have the advantage of having easily understandable logical decisions, especially if the trees have been pruned. (Schuller et al., 2011)

Finally, Ayadi concludes that "the GMM achieve the best compromise between the classification performance and the computational requirements for training and testing." (El Ayadi et al., 2011) And interestingly, Koolagudi et al. remarks that few are the studies that choose the classification model based on experimentation. (Koolagudi & Rao, 2012)

Having overviewed individually the major classifiers in speech emotion training, we will address the idea of ensemble learning or multiple classification systems (MCS). There are three methods to combine different classifiers: hierarchical, parallel and serial. The idea of a hierarchy is to build a tree, in which, as we go more and more in depth, the number of available classes is diminished. The serial approach is a special case of that tree, in which each all classifiers are placed in a queue. Finally, in the parallel approach, classifiers work independently. (El Ayadi et al., 2011) A popular ensemble classifier is the Random Forests (RF), an ensemble of trees; it is practically immune to the "curse of dimensionality", while still providing all the advantages of classification trees. (Schuller et al., 2011)

Finally, we will now review results from speech emotion recognition tasks. As we will remark, these statistical classification results are overall considerably higher than human emotion recognition results. Luengo's automatic speech emotion recognition results, based on 6 features (mean pitch, mean energy, pitch variance, skew of logarithmic pitch, range of logarithmic pitch and range of logarithmic energy) and on a Basque acted speech emotional database, are reported in table 10. Furthermore, these results were overall slightly better than those based on 86 features, implying that a smaller number of features may be advantageous.

**Table 10: Luengo et al. speech emotion recognition using prosodic features**

| | Anger | Fear | Surprise | Distress | Joy | Sadness | Neutral |
|---|---|---|---|---|---|---|---|
| **SVM on 6 features** | 94.9% | 96.9% | 90.5% | 82.5% | 90.7% | 95.9% | 94.9% |
| **GMM on 6 features** | 91.8% | 92.8% | 87.4% | 75.3% | 84.5% | 85.6% | 89.7% |
| **GMM on 86 features** | 90,7% | 91,8% | 82,1% | 78,4% | 70,1% | 91,8% | 88,7% |

Source: (Luengo, Navas, Hernáez, & Sánchez, 2005)

Chen's automatic speech emotion recognition results, based on a mandarin acted speech emotional database, are reported in table 11. These results are particularly interesting for they compare two feature reduction methods, revealing that LDA tends to perform better on their database.

**Table 11: Chen et al. speech emotion recognition**

| | Anger | Fear | Surprise | Distress | Happiness | Sadness |
|---|---|---|---|---|---|---|
| **LDA+SVM** | 72.2% | 37.2% | 43.3% | 42.8% | 52.8% | 53.3% |
| **PCA+SVM** | 56.7% | 32.2% | 35.0% | 39.4% | 42.8% | 52.8% |
| **LDA+ANN** | 90.8% | 44.2% | 13.3% | 9.6% | 35.0% | 51.7% |
| **PCA+ANN** | 92.5% | 42.1% | 7.1% | 7.9% | 29.6% | 55.8% |

Source: (L. Chen, Mao, Xue, & Cheng, 2012)

Lin's speech emotion recognition results, using the Danish Emotional Speech (DES) database, are reported in table 12. These results are interesting because they feature high speech recognition rates; unexpectedly, gender independent results reveal higher recognition rates: this might be a result of employing a small database.

**Table 12: Lin et al. speech emotion recognition**

| | **Anger** | **Surprise** | **Happiness** | **Sadness** | **Neutral** |
|---|---|---|---|---|---|
| **HMM (gender independent)** | 100% | 100% | 94.7% | 100% | 100% |
| **SVM (female)** | 91.9% | 91.7% | 92.2% | 97.5% | 95% |
| **SVM (male)** | 86.9% | 89.2% | 86.7% | 100% | 84.2% |

Source: (Lin & Wei, 2005)

Hu's speech emotion recognition results, based on a mandarin acted speech emotional database, are reported in table 13.

**Table 13: Hu et al. speech emotion recognition**

| | **Anger** | **Fear** | **Happiness** | **Sadness** | **Neutral** |
|---|---|---|---|---|---|
| **GMM (Female)** | 97.9% | 78.4% | 90.8% | 98.6% | 100% |
| **GMM (Male)** | 96.5% | 91.1% | 80.0% | 96.3% | 88.2% |

Source: (Hu, Xu, & Wu, 2007)

Before proceeding to the next chapter, it's quite remarkable that Hu et al. results and Lin et al. results reveal that generally recognition rates on male utterances are higher than recognition rates on female utterances. Though little data is presented to draw any conclusion, such inquiry may reveal whether one gender reveals more emotional cues in speech.

# CHAPTER 3: Methodology

Our methodology is divided in two acts. The first one is the methodology used to design the European Portuguese Emotional Discourse Database (EPEDD). The second one is the methodology used to have a machine recognize emotions, based on the EPEDD.

## 3.1. Portuguese Emotional Speech Database Design

Databases used for pattern recognition must be high-quality ones; otherwise, wrong prediction or undesired action might be performed by the machine. On the following sections, we will explain the method we followed so as to create the EPEDD, while reviewing research advice on emotional speech database design; we will begin by discussing the scope and the utterances. We will then move to the discussion of the recording conditions. Then, we will discuss the validation. Finally, we will briefly comment on the modalities and on the database availability.

### 3.1.1. Scope

Cowie defines the emotional speech database issue as the variable that cover emotions, speakers, genders, language, dialect and social setting. (Douglas-Cowie et al., 2003) Aside from discussing these issues, we will discuss acting experience and technique.

#### *3.1.1.1. Number of speakers*

A reasonable number is needed to create a speaker independent emotion recognition system, for each speaker has its own voice quality and might have its own individual emotional expression. For instance, 10 speakers, as was the case for the German Database (Burkhardt et al., 2005), is considered insufficient: in fact, new databases have been created featuring as much as 50 speakers. (Schuller et al., 2009) Though, because we do not have the means for such a big number, we chose 8 speakers.

#### 3.1.1.2. Age and Sex of the speakers

The speakers are between 18 to 26 years old—this is a result of choosing to work with almost only acting school students. Both sexes are equally represented.

#### 3.1.1.3. Language, Dialect, Culture and Social Setting

Our database is based on the continental Portuguese culture and language. There is not a single Portuguese acting emotional speech database, therefore it is important to create one. Both the south and the north of Portugal are represented.

Culture is also important to keep in mind, because it may shape the way we express emotions. (El Ayadi et al., 2011) For example, Japanese society considers an open emotion display anti-social, furthermore, it is normal to smile when angry or embarrassed. (Staroniewicz & Majewski, 2009) Also, the social setting is by no means less important, because, for example, business type actions might be irritable in a social context. (Douglas-Cowie et al., 2003) As we were recording the actors, we did not tell them to imagine a specific social setting; such a task is too specific and would not only require a larger budget but probably also a specific acting training. Therefore, such a parameter is more suitable for a natural or induced speech database.

However, we did tell them to express openly the emotions. This openness is a crucial parameter. If we would tell the actors to express anger while trying to hide it, instead of sounding like the traditional hot-anger, it would sound more like the cold-anger. There is the hypothesis that the social setting—probably combined with the culture—is the same parameter as the emotional openness.

Basically, much as the culture and social setting can be ignored in an acting speech database, they must be considered in a natural or induced speech emotional database.

#### 3.1.1.4. *Acting Experience*

Researchers have been experimenting with actors with different experience background. Burkhardt's emoDB was recorded with amateur actors (Burkhardt et al., 2005), the Polish Emotional Speech Database was recorded with professional, semi-professional and amateur actors (Staroniewicz & Majewski, 2009) and the Danish Emotional Speech database was recorded only with semi-professional actors so as to avoid exaggeration. (Engberg & Hansen, 1996) (El Ayadi et al., 2011) We have chosen to work with acting school students for budget reasons; actors being usually criticized for overacting the emotions, amateur actors usually having difficulty enacting emotions on command, we strongly believe that semi-professional actors or acting students hold the best results.

#### 3.1.1.5. *Acting Method*

The actors would hear a brief description of the emotion requested; this step is crucial because emotion-related words are usually subjective. For example, actors would ask if it was a good or a bad surprise. And whether "interest" was related to "self-interest" or "romantic interest". Also, the distinction between "sadness" and "apathy" needed to be clear for the actors.

Then, applying the Stanislavsky method, they would utter each sentence in that emotion; i.e. the actor would reminisce a moment when they had felt the requested emotion in order to feel it again and express the utterance more genuinely. (Burkhardt et al., 2005) Because reading speech displays distinctive characteristics, actors were not allowed to read the sentences while uttering them. (Douglas-Cowie et al., 2003) However, they were allowed and encouraged to utter as many times as they desired each sentence. The time needed per emotion was around between 10 to 15 minutes. If an actor would feel exhausted, we would also make a 15-minute pause. The average time per session was 2 hours. Finally, actors would be recorded independently so as not to have them interact to each other.

Because we are interested in the analysis of prosody in speech, whispers and shouts were not allowed. Moreover, shouts disturb evaluation processes, because an evaluator, listening to a shout will immediately associate it to anger, without paying attention to other subtler prosodic, spectral and voice quality details.

When recording actors, a first evaluation is performed, therefore this work is usually done by more than one person in order to pay attention to all the details (Engberg & Hansen, 1996) (Jovi et al., 2004); we strongly recommended such an approach because it will reduce the recording time. However, only half of the time were we able to have two evaluators during the recordings: myself and an actor director with a bachelor degree on psychology.

Finally, a problem reported by Burkhardt (Burkhardt et al., 2005) is that the accent was not performed on the same words from emotion to emotion. Therefore, comparison between different sentences is harder. But we did not even suggest actors to try to maintain the same stress pattern, for fear that such an effort should compromise the actor's overall performance.

#### *3.1.1.6. Emotions*

We recorded eight unambiguous emotional states: anger, sadness, joy, interest, fear, disgust, apathy and surprise, plus neutral as a control. These are the eight emotions in Lövheim's model, the one we decided to base our research in, which is inspired by Tomkins' model. (Lövheim, 2012) Though researchers might prefer that emotion frequency in a database reflects their real distribution, we chose to have a uniform distribution, not being particularly interested in any specific emotion.

One should also refer that moods (long-lasting emotions), can coexist temporarily with short-lasting emotions. For example, someone in a depressive mood, might suddenly get surprised or interested (Lövheim, 2012). The way such a coexistence reflects in speech remains to be studied. And, because we are asking actors to act momentarily emotions, such a question is important to be answered, in order to design optimal emotional speech databases.

### 3.1.2. Utterances

#### *3.1.2.1. Sentences and Words*

Our corpus consists of two negative declarative sentences, four positive declarative sentences, two interrogative sentences and two imperative sentences. Half of these are short sentences; the other half are long ones. These are all as approximatively emotionally neutral as can be. Finally, these sentences could be used in everyday conversation and were inspired by the ones used in the emoDB (Burkhardt et al., 2005). In the following paragraphs, an explanation of our sentences is presented.

Sentences too short are harder to be pronounced emotionally and, the longer the sentence is, the harder it is for the actor to continually be in the expected emotional state. (Amir, Ron, & Laor, 2000) Therefore, we constrained ourselves only to short and long sentences; so, phenomena like emotional build-up and paragraph pausing and others requiring longer stretches of speech are to be studied ideally on natural or induced emotional speech databases. In fact, during the recordings, we also remarked that longer sentences were easier to be pronounced emotionally. Contrastingly, isolated words (for example, “Yes” and “No”) and paragraphs are found in many databases. (Jovi et al., 2004; Ververidis,

Ververidis, Kotropoulos, & Kotropoulos, 2003) For example, the Serbian Emotional Speech Database features 32 isolated words, 30 short words sentences, 30 long sentences, one passage and full phonetic balance. (Jovi et al., 2004) So, we decided to experiment and add the words “Yes” and “No” to the database, in an attempt to design a richer and more flexible database: the recognition of the emotional color of “yes” or “no” answers are particularly interesting. Actors did however struggle uttering these. However, big data validation requiring a considerable budget, we decided not to include single words within the dataset evaluation, because our database was already big enough, considering the validation effort it already would require.

Dialogues being the most common form of speech, they ought to be a priority in studying (Saratxaga & Navas, 2006); however, for budget and practical reasons, our corpus consists in only monologue speech. Because, if monologue speech already presents a challenge to actors, dialogue speech would probably be even more difficult. Thus, dialogues should only be studied for natural or induced emotional speech databases.

Some emotional speech databases do consist of non-sentences; much as these are certainly emotionally neutral, actors struggle uttering these, causing the naturalness of speech to decrease even further. (Burkhardt et al., 2005)

The designers of the Basque emotional speech database certified that it would contain the most common words, creating thus a large acting emotional speech database. (Saratxaga & Navas, 2006) We are not sure of the importance of such a word rich database, moreover such a task was too time-consuming.

Finally, these were the sentences we created and recorded (and their translation to English):

### (1) Short Sentence No.1

O António está dentro do elevador.

António is inside the elevator.

*u ã.t’o.niu iʃ.tˈa d’ẽ.tɾu du i.lɨ.vɐ.d’oɾ*

### (2) Short Sentence No.2

Ela não me devolveu isso ontem.

She didn’t give it back to me.

*‘ɛ.lɐ nɐ̃w mɨ dɨ.voɫ.vˈew ‘i.su ‘õ.tɐ̃j*

### (3) Short Sentence No.3

Conta-me isso, sim.

Tell me that, yes.

*k’õ.tɐ mɨ isu s’ĩ*

### (4) Short Sentence No.4

Daqui por sete horas, ele já terá entrado.

Within seven hours, he will already have arrived.

*dɐ.kˈi puɾ sˈɛ.tɨ ‘ɔ.ɾɐʃ ele ʒa tɨɾˈa ẽ.tɾˈa.du*

### (5) Short Sentence No.5

Foi este o cesto que me deram?

Was this the basket that they gave me?

*foi eʃtɨ u seʃtu kɨ mɨ dɛ.ɾɐ̃w*

### (6) Long Sentence No.1

A gente está a comer nas rochas, ao lado do moinho do norte.

We are eating on the rocks, next to the mill of the north.

*ɐ ʒˈẽ.tɨ ɨʃ.t’a ɐ k’u.meɾ nɐʃ ʀˈɔ.ʃɐʃ, ‘aw l’a.du du mw.ˈi.ɲu du nˈɔɾ.tɨ.*

### (7) Long Sentence No.2

Porque é que as mochilas estão ali, debaixo da mesa?

Why are the backpacks there, beneath the table?

*pˈoɾ.kɨ ɛ kɨ ɐʃ mu.ʃ’i.lɐʃ iʃ.tˈɐ̃w ɐ.l’i dɨ.bˈaj.ʃu dɐ m’e.zɐ*

### (8) Long Sentence No.3

Acaba de carregar isso para cima e põe-te de novo a ir para baixo.

Finnish carrying that upstairs and come back downstairs again.

*ɐ.k’a.bɐ dɨ kɐ.ʀɨ.gˈaɾ ‘i.su p’ɐ.ɾɐ s’i.mɐ i poĩ tɨ dɨ n’ov.u ɐ iɾ p’ɐ.ɾɐ b’aj.ʃu*

### (9) Long Sentence No.4

Aos fins-de-semana, eu ia sempre a casa e comprava húngaros.

At the weekend, I would always go home and buy cakes.

*awʃ fˈĩʃ dɨ sɨ.m’ɐ.nɐ ew i.ɐ s’ẽm.pɾɨ ɐ k’a.zɐ i kõ.p’ɾa.vɐ ‘ũ.gɐ.ɾuʃ*

### (10) Long Sentence No.5

Eu não quero comer fora, quero é tomar um copo com o Francisco.

I don’t want to eat at a restaurant, but I want to have a drink with Francisco.

*ew nɐ̃w kˈɛ.ɾu ku.m’eɾ foɾɐ kˈɛ.ɾu ɛ tu.m’aɾ ũ k’ɔ.pu kõ u fr’ã.siʃ.ku*

#### *3.1.2.2. Phonemes*

A representation of all phonemes is essential: this is a necessity for concatenative synthesis (Douglas-Cowie et al., 2003) and for classification algorithms applied on spectral features. Furthermore, certain phoneme combinations are considered necessary by the Basque Emotional Speech Database. (Saratxaga & Navas, 2006) Our completed corpus contains all the Portuguese phonemes, except for the phoneme "ʎ". Furthermore, all the vowels of the Portuguese language were represented in the final words of the sentences. Our sources include the web page of the Instituto Camões ("Convenções e Transcrição Fonética," n.d.) and the phonetic dictionary web page of the Instituto de Língua Teória E Computational (ILTEC). ("Dicionário Fonético," n.d.)

#### *3.1.2.3. Other Sounds*

Finally, some databases include the study of sounds such as laughter moaning, screams and sobbing. (Lima et al., 2013) Because we are interested in novelty, there was no interest in including these types of sounds.

### 3.1.3. Recording Conditions

Much as emotion recognition in speech ought to perform equally well with noise and under different acoustic conditions, an acted emotional speech database recorded in studio conditions (high-quality microphone and an isolated room) are enough, because noise and reverberations can be artificially added for further examination.

We recorded all actors in the same acoustically-treated and isolated room at the Universidade Católica do Porto. That room was large enough to accommodate the actors and the listeners together comfortably. Comfort was a priority because we wanted actors to be relaxed, for fear that anxiety should contaminate the recordings. We used a high-quality microphone: DPA 4041-Sp (omnidirectional). We used it to minimize self-noise and correctly retrieve the spectral content. It was put 15 cm away from the mouth of the actor, in order to have a minimum pre-amplifier gain, and thus minimal extraneous noise. However, the shorter this distance, the higher the amplitude variation, should the actor move his head from the ideal position, and the more freely the actor can gesticulate; therefore, ideally, a 30 cm is advisable for future approaches, something we only concluded halfway across the test.

### 3.1.4. Modalities

Though our database consists only of audio recordings, emotional speech databases are moving from one modality towards multimodality, including facial expressions, gestures, text and physiologic parameters (Schuller et al., 2009), among others. However, multimodal databases do not render monomodal ones without value, because monomodality might have specific characteristics. For example, a phone conversation might deliver a richer emotional content than a face to face conversation.

### 3.1.5. Availability

Emotional speech databases are moving from seldom public to public available. (Schuller et al., 2009) Ours will be a public available one, as we feel peer-validation to be a cornerstone of research.

### 3.1.6. Evaluation and validation

#### *3.1.6.1. Database evaluation platform development*

After recording the actors, we must evaluate how accurate each of the emotional utterances is. Typically, this process is omitted, researchers assuming that actors performed accurately. In contrast, researchers usually conduct a human evaluation of the database. (Burkhardt et al., 2005; Jovi et al., 2004; Staroniewicz & Majewski, 2009) For that purpose, we will have the utterances evaluated by judges. Ideally, judges ought to be expert; examples of expert judges may include acting directors, sound engineers and woman musicians. On the one hand, acting directors and sound engineers are considered expert listeners because they professionally work with actors respectively in film production and in Automated Dialog Replacement (ADR). On the other hand, woman and musicians achieved a recognition rate 10% higher than respectively man and non-musicians. (Staroniewicz & Majewski, 2009) Though experienced judges are better, we worked with anonymous Portuguese voluntary judges, our budget being limited. Most of them were students from different Portuguese institutions of higher education. We divided the 718 utterances into 37 evaluations, designed with Google Forms, each featuring a maximum of 20 utterances to evaluate. The estimated time was 10 minutes per evaluation.

Our objective being to acknowledge how accurate each of the emotional utterance is, we must not tell the experienced judges which emotion the actor was representing. Would the judges know a priori the emotions the actors had to represent, they would maybe unfortunately overestimate acoustic details of the emotion the actor tried to portray, while underestimating acoustic details of other existent mixed emotions. For each utterance, the judge must then decide the attempted emotion and its accuracy. For greater detail, the intensity of the emotion could also be required; however, to keep the evaluation simple, our database being a small one, and our main objective being emotion retrieval, we did not request the intensity.

The better a judge knows a speaker, the better he/she may retrieve his/her emotions. Consequently, so as not to have biased judges, the emotional utterances must be evaluated one at a time in a specific order, having judges listening to an utterance per every actor before repeating any actor. And once an utterance is evaluated, it cannot be evaluated again. Finally, the experienced judges will be allowed to listen repeatedly the same utterance.

Last but not least, development of platforms for database validations or evaluations are necessary for avoiding overlapping work.

#### *3.1.6.2. Validation criteria*

An evaluated utterance to be considered validated had to meet the following criteria:

(a) The utterance's most voted emotion is the emotion the actor had pretended to express and the most voted emotion must be at least 15% higher than the second most voted emotion, unless:

(b) If the utterance's most voted emotion is not the emotion the actor had pretended to express, it is not neutral and it has been voted by at least 68% of the evaluators, then the validated emotion is the one 68% of the evaluators voted for.

(c) If the actor had pretended to express disgust, but evaluators have not voted it as disgust and at least 20% of the evaluators must have voted disgust, then the validated emotion is disgust.

The first criteria certify stability and confidence in the validated emotion. However, to have a larger database, while trusting evaluators, criteria (b) was used. Also, few being the disgust utterances identified as disgust and this emotion being particularly hard to identify, we decided to allow emotions being classified as disgust with low recognition rates.

#### *3.1.6.3. Database Evaluation Analysis*

We extracted the following features on the validated database:

(a) Gender independent emotion recognition rate (overall and per emotion)
(b) Gender independent overall emotion recognition rate on short and long sentences
(c) Gender dependent (male and female) emotion recognition rate (overall and per emotion)
(d) Gender dependent (male and female) emotion recognition rate (overall and per emotion) on male actors
(e) Gender dependent (male and female) emotion recognition rate (overall and per emotion) on female actresses.

And we extracted the following features on the full database:

(a) Percentage of equivalence between pretended emotion and evaluated emotion
(b) Percentage of utterances validated
(c) Percentage of utterances validated with a recognition superior to 50%
(d) Percentage of utterances validated with a recognition superior to 70%
(e) Average acting quality
(f) Percentage of utterances with an acting quality superior to 3.1
(g) Average acting quality on utterances with and acting quality superior to 3.1

We decided to extract the percentage of utterances with an acting quality superior to 3.1 because we wanted to have an idea of the number of utterances with an acting quality above average.

There is little research on comparing male overall emotion recognition rates with female emotion recognition rates and there is no research on comparing male emotion recognition per emotion with female emotion recognition per emotion. (Staroniewicz & Majewski, 2009) Furthermore, there has also never been research into the difference between women emotion recognition on women, women emotion recognition on men, men emotion recognition on women and men emotion recognition on men. Thus, we decided pursue this line of analysis.

## 3.2. Speech Emotion Statistical Classification Method

### 3.2.1. Speech Emotion Feature Extraction

We used open source Opensmile software to extract audio features from each utterance.[1] Opensmile has multiple available configuration settings, some having been designed to extract features for Interspeech emotion challenges. (Schuller, Steidl, Batliner, Burkhardt, et al., 2013; Schuller, Steidl, Batliner, Vinciarelli, et al., 2013; Schuller, Steidl, et al., 2012, 2011; Schuller et al., 2009) In fact, because many speech emotion recognition researchers were using this software for speech emotion feature extraction, we decided to use it too for practical reasons, while ensuring a better comparability of results. The number of features to be extracted is suggested to be somewhere between 1000 and 50000: we, therefore, applied the Interspeech 2013 configuration settings, extracting 6373 features. This set includes prosodic features, voice-quality features (jitter, shimmer, HNR), and spectral and cepstral features (MFCCs). (Schuller, Steidl, Batliner, Vinciarelli, et al., 2013) So as to improve recognition, it is reported that,

[1] Should the reader be interested in more information on Opensmile, please visit their official website: http://audeering.com/technology/opensmile/

as soon as speech recognition systems will be robust enough, linguistic features will be commonplace within speech emotion feature extraction. (Schuller et al., 2009)

### 3.2.2. Speech Emotion Statistical Classification

We used open source Weka software to statically classify the utterances per emotions.[2] We began by applying Weka's PCA algorithm on the full database and on the validated one, respectively, reducing the number of features from 6373 to 457 and from 6373 to 53. Then we applied SVMs, Random Forests and ANNs, exploring multiple available parameters on both databases.

The recall equals to the rate of instances labeled with a specific emotion that the algorithm was able to identify, also usually referred as emotion recognition rate, while the precision equals to the rate of instances the algorithm correctly labeled. Weka computes both recall and precision; additionally, it also computes F-Measure. F-Measure is a harmonic mean between recall and precision.

$$Recall = \frac{TP}{TP + FN}$$

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

$$F - Measure = 2 \cdot \frac{Recall \cdot Precision}{Recall + Precision}$$

Usually speech emotion recognition researchers are only interested in recall values from statistical classification algorithms. However, having not defined the speech emotion recognition goal, whether we are more interested in achieving a high precision, detecting emotions very accurately, but at the expense of missing some, or we are interested in finding a huge quantity of emotions, but

[2] Should the reader be interested in more detailed information on Weka, please visit their official website: http://www.cs.waikato.ac.nz/ml/weka/

at the expense of some false positives, we will provide Recall and F-Measure values too, so as to be able to inspect whether some algorithms perform a higher recall or a higher precision. It was decided to use F-measure instead of precision, because this expression being a harmonic mean between precision and recall, it gives a better in insight to the overall algorithm performance.

# CHAPTER 4: RESULTS AND RESULTS DISCUSSION

## 4.1. Validation and Evaluation

Though our intention was to have 37 different Google Forms evaluated, we ended up with having only 15 evaluated (40% of the database) achieving 116 validated utterances. Our minimum requirement for an evaluation to be completed was 12 evaluators; however, the mean number of evaluators per utterance was 16. The clear majority of evaluators were female, leading to an underrepresentation of men, affecting thus any statistic concerned with emotion recognition by men. Moreover, not all emotions were equally validated in quantity, anger and neutral being drastically more numerous than disgust and sadness. Table 14 shows the number of utterances per emotion:

**Table 14: Number of instances per emotion**

| Emotion | anger | joy | excitement | neutral | fear | sadness | disgust | apathy | surprise |
|---|---|---|---|---|---|---|---|---|---|
| Number | 24 | 9 | 13 | 22 | 11 | 10 | 7 | 9 | 11 |

As it can be seen, not only does the emotion representation vary, but disgust, apathy, joy, fear, sadness and surprise have a very limited number of instances. Should the validation be complete, we would probably have on average 2,5 times more instances per emotion. The following tables contain statistical results based retrieved from the performed evaluations:

**Table 15: Gender independent emotion recognition rates in validated utterances**

| | |
|---|---|
| Overall recognition | 69,6% |
| Overall recognition on short sentences | 67,3% |
| Overall recognition on long sentences | 70,9% |
| Sadness recognition | 60,6% |
| Disgust recognition | 40,5% |

| | |
|---|---|
| Surprise recognition | 73,1% |
| Excitement recognition | 70,0% |
| Neutral recognition | 71,2% |
| Anger recognition | 79,7% |
| Apathy recognition | 54,7% |
| Joy recognition | 74,9% |
| Fear recognition | 69,9% |

**Table 16: Gender dependent emotion recognition rates in validated utterances**

| | |
|---|---|
| Overall recognition by men | 68,3% |
| Overall recognition by women | 69,8% |
| Sadness recognition by women | 66,0% |
| Disgust recognition by women | 51,9% |
| Surprise recognition by women | 76,4% |
| Excitement recognition by women | 71,8% |
| Neutral recognition by women | 70,0% |
| Anger recognition by women | 83,2% |
| Apathy recognition by women | 55,7% |
| Joy recognition by women | 77,5% |
| Fear recognition by women | 67,5% |

**Table 17: Database validation statistics**

| | |
|---|---|
| Percentage of equivalence between pretended emotion and evaluated emotion | 51,5% |
| Percentage of utterances with an acting quality superior to 3.1 | 62,8% |
| Percentage of utterances validated | 40,3% |
| Percentage of utterances validated with a recognition superior to 50% | 32,9% |
| Percentage of utterances validated with a recognition superior to 70% | 18,8% |

**Table 18: Average acting quality**

| | |
|---|---|
| Average acting quality | 2,2 |
| Average acting quality on utterances with and acting quality superior to 3.1 | 3,5 |
| Average male acting quality | 1,84 |
| Average female acting quality | 2,49 |

Firstly, from the 290 evaluated utterances, only 40,3% were considered. Emotions recognition was on average 69.6%, ranging from 40,5% (disgust) to

79,7% (anger). These are quite good results, considering that random guessing would result on a recognition rate of 11%, evaluators having 9 hypotheses. It is conspicuous that disgust seems to be on average the hardest emotion to recognize, should anyone try to recognize it. The easiest emotion to be recognized appears to be anger; however, Staroniewicz et al. highest recognized emotion was surprise, a not common emotion in most of the databases. (Staroniewicz & Majewski, 2009)

Additionally, short sentences had a 67,3% rate, in contrast to long sentences, which had a 70,9% recognition rate. This 3,5% difference is probably due to the fact that long sentences feature more emotion acoustic cues, since they are longer.

In contrast with our expectations, emotion recognition by men was 1,5% better than by women. However, because there was a too small number of men evaluating the database, this data has little meaning. And, our evaluators being voluntaries, it is interesting that more women than men took part in the validation process: whether women are more solidary or they are more interested in an emotion evaluation inquiry.

Only 51,5% of the sentences were recognized as the emotion the actors had pretended to express. Only 62,8% of the sentences were evaluated as at least 3,1 in acting quality, on a scale from 1 to 5, the acting quality being on average only 2,1 and the acting quality on sentences with a quality higher than 3,1 being 3,5. The acting was in fact a big limiter in our speech emotion database. We believe there are at least three reasons for this; the first is due to having worked with actors, the second is due to not having worked with the best actors. And the last reason is probably because, during the recordings, at least one expert in emotion recognition should have been present.

Finally, only 32,9% and 18,8% of the sentences, respectively, had at least 50% and 70% of the evaluators agreeing on the emotion.

## 4.2. Statistical Classification

We applied multiple statistical classification algorithms to the full database (both validated and non-validated instances) and to the validated database using the software Weka. Respectively, tables 19 to 24 contain data obtained applying SVMs, Random Forests and ANNs on the full database. Then, tables 25 to 30 contain data obtained applying SVMs and Random Forests on the full database. The data contained in the tables concerns only F-Measure and recognition rates (recall); for detailed data information on the results and on the implementation of these algorithms, please see the annexes.

**Table 19: SVM F-Measure values on the full database**

| | | | | |
|---|---|---|---|---|
| total | anger | joy | excitement | neutral |
| 48.9% | 63.4% | 32.9% | 32.7% | 60.0% |
| fear | sadness | disgust | apathy | surprise |
| 49.1% | 41.0% | 39.7% | 79.0% | 41.8% |

**Table 20: SVM recognition rates on the full database**

| | | | | |
|---|---|---|---|---|
| total | anger | joy | excitement | neutral |
| 48.7% | 58,2% | 33.8% | 32.9% | 63.8% |
| fear | sadness | disgust | apathy | surprise |
| 52.5% | 42.5% | 36.3% | 77.5% | 41.3% |

**Table 21: Random Forests F-Measure values on the full database**

| | | | | |
|---|---|---|---|---|
| total | anger | joy | excitement | neutral |
| 34.7% | 52.0% | 19.6% | 13.5% | 47.5% |
| fear | sadness | disgust | apathy | surprise |
| 43.0% | 23.0% | 28.8% | 56.4% | 28.1% |

**Table 22: Random Forests recognition rates on the full database**

| total | anger | joy | excitement | neutral |
|---|---|---|---|---|
| 36.8% | 57.0% | 17.5% | 11.4% | 65.0% |
| fear | sadness | disgust | apathy | surprise |
| 46.3% | 21.3% | 25.0% | 66.3% | 21.3% |

**Table 23: ANN F-Measure values on the full database**

| total | anger | joy | excitement | neutral |
|---|---|---|---|---|
| 26.6% | 34.4% | 22.8% | 18.6% | 24.8% |
| fear | sadness | disgust | apathy | surprise |
| 26.7% | 24.0% | 25.2% | 42.0% | 20.5% |

**Table 24: ANN Recognition rates on the full database**

| total | anger | joy | excitement | neutral |
|---|---|---|---|---|
| 26.2% | 27.8% | 26.3% | 16.5% | 25.0% |
| fear | sadness | disgust | apathy | surprise |
| 28.8% | 27.5% | 25.0% | 37.5% | 21.3% |

**Table 25: SVM F-Measure values on the validated database**

| total | anger | joy | excitement | neutral |
|---|---|---|---|---|
| 44.1% | 72.0% | 37.5% | 45.5% | 69.6% |
| fear | sadness | disgust | apathy | surprise |
| 18.2% | 8.3% | 33.3% | 40.0% | 25.0% |

**Table 26: SVM recognition rates on the validated database**

| total | anger | joy | excitement | neutral |
|---|---|---|---|---|
| 44.0% | 75,0% | 33.3% | 38.5% | 63.6% |
| fear | sadness | disgust | apathy | surprise |
| 18.2% | 10.0% | 28.6% | 33.3% | 27.3% |

**Table 27: Random Forests F-Measure values on the validated database**

| total | anger | joy | excitement | neutral |
|---|---|---|---|---|
| 31.9% | 55.3% | 30.8% | 0% | 46.9% |
| fear | sadness | disgust | apathy | surprise |
| 0% | 26.7% | 44.4% | 33.3% | 16.7% |

**Table 28: Random Forests recognition rates on the validated database**

| | | | | |
|---|---|---|---|---|
| total | anger | joy | excitement | neutral |
| 38.8% | 87.5% | 22.2% | 0% | 68.2% |
| fear | sadness | disgust | apathy | surprise |
| 0% | 20.0% | 28.6% | 22.2% | 9.1% |

**Table 29: ANN F-Measure values on the validated database**

| | | | | |
|---|---|---|---|---|
| total | anger | joy | excitement | neutral |
| 35.7% | 47.1% | 50.0% | 15.4% | 48.9% |
| fear | sadness | disgust | apathy | surprise |
| 26.1% | 10.5% | 33.3% | 26.7% | 38.1% |

**Table 30: ANN Recognition rates on the validated database**

| | | | | |
|---|---|---|---|---|
| total | anger | joy | excitement | neutral |
| 36.2% | 50.0% | 55,6% | 15.4% | 50.0% |
| fear | sadness | disgust | apathy | surprise |
| 27.3% | 10.0% | 28.6% | 22.2% | 36.4% |

This paragraph is concerned with the discussion of the results obtained from the full database, while the next one is more concerned with those obtained from the validated one. SVM proved to be the best algorithm for our full database with an overall 48,7% emotion recognition rate, the Random Forests and the ANN being mediocre in comparison, respectively with an overall 26,2% and 36,8% emotion recognition rate: we will therefore turn our attention exclusively to the results from this algorithm. Our SVM results are similar to those from Chen et al. (L. Chen, Mao, Xue, & Cheng, 2012), but well below other reported results above. However, considering that we are working with 9 emotions, instead of the typical 5 or 7 emotions, our results are quite good. Apathy was the more recognizable emotion, which is very attractive, especially because rarely this emotion is found in speech emotion recognition tasks. In contrast, the confusion matrix reveals that sadness recognition rate would have been significantly higher, if we had not included apathy in our emotions to be recognized. Moreover, surprise, excitement and joy were often misinterpreted between them; this misinterpretation might be the result of bad acting.

The validated database exhibited quite unexpected results, thus interesting ones. SVM proved to be the best algorithm for our validated database with an overall 44,0% emotion recognition rate, the Random Forests and the ANN being lower in comparison, respectively with an overall 38,8% and 36,2% emotion recognition rate. We will turn for now our attention to the SVM results and then we will discuss the other two. The highest recognized emotion was anger, with a 75% recognition rate, while the lowest recognized emotion was sadness, with a 10,0% recognition rate. In contrast, the gender independent sadness recognition was 60,6%. This reflects that either the validation was poorly performed or the statistical algorithm requires larger data. In fact, there were only 10 sadness-labeled instances, disgust was the poorest recognized emotion in the full database by the SVM and there were only 7 disgust-labeled instances in the validated database. Certainly, should one of the validation requirements be that any validated utterance's emotion was equal to its pretended emotion, then sadness results would have been higher. Random Forests exhibited three unexpected results. The first one is the high recognition rate for anger-labeled instances. The second one are the zero-recognition rate for fear-labeled instances and for excitement-labeled instances. While the increased anger recognition implies that anger validation has been performed correctly, the decreased fear and excitement recognition is also probably a consequence of a lack of data or of a poorly validated database. Finally, ANN revealed three interesting results: a 55,6% joy recognition rate, the highest joy recognition rate and a 36,4% surprise recognition rate, a quite high recognition rate for surprise. The third result is that, though the database was smaller, its recognition greatly improved, implying that ANN might be the ideal algorithm on a carefully validated database.

On the one hand, we can deduce that, though more work must be done to develop a larger and better validated database, so as to create a high-quality emotional speech dataset, each algorithm has its own set of best recognized emotions. Ensemble statistical classification algorithms that attribute specific algorithms to recognize specific emotions might achieve great results. For example, one idea would be to have Random Forests recognize anger from all other emotions, then moving to SVMs to recognize neutral from the 7 other emotions and so on until ANNs recognize joy. In contrast to a serial ensemble classifier, another paradigm would be to apply a parallel ensemble classifier, in which one algorithms recognizes sums of emotions. For example, one algorithm could recognize the sum of joy and excitement, from the other emotions, and then, another algorithms (ANN) would distinguish joy from excitement.

On the other hand, having noticed that certain algorithms deliver significantly different F-measure values from recall ones, for certain emotions, for instance Random Forest anger recall is 32,2% higher that its F-Measure, speech emotion recognition systems can be developed so as to be directed at delivering either a high recall or a high precision, depending on the user goal.

# CHAPTER 5: CONCLUSION

The goal we proposed was to create a validated Portuguese speech emotion database and to achieve speech emotion recognition through statistical classification algorithms based on it, while discussing database development methodology and, in particular, acted emotional speech database one.

Based on state-of-the-art acted speech emotional database methodology, we built the European Portuguese Emotional Discourse Database (EPEDD), limited however by budget. We evaluated 40% of the database, enabling us to produce a validated database, filtering 60% of the evaluated utterances. The validated database is small, containing only 116 instances for a total of 9 different emotion-labels. Therefore, for further investigation on acted Portuguese speech emotion recognition, it is recommended to completely evaluate this future public available database.

The average acting quality of the original database was evaluated, in a scale from 1 to 5, as 2,3, indicating that had we at least one expert listener or an experienced acting director present during the recordings, more utterances would have been validated, leading to an optimized process. And, should expert listeners be responsible for the validation, then an additional non-experienced evaluation would lead to evaluate non-experienced human emotion recognition ability or to generate a database containing only popularly recognizable emotional speech. For instance, the validated database that we created has its utterances recognized at a 69,6% rate, by unexperienced judges. Databases featuring popularly recognizable emotions can have pragmatic interest, should the ability to mimic human emotion recognition rates be a goal.

Moreover, the validated databases features anger as the highest recognizable emotion at a 79,7% rate, while disgust as the lowest recognized emotion at a 40,5% rate.  A better recognition rate could have been obtained, had we created a more rigorous validation requirement; however, such a validation would unfortunately make our database even smaller.

In the scope of automatic speech emotion recognition, applying Opensmile and Weka software, we had respectively features extracted and statistical classification algorithms trained on the full database and on the validated one, SVMs proving to be the optimal algorithm for both. Respectively, the emotion recognition rates were 48,7% and 44,0%. Apathy had the highest recognition rate at 79.0%, while excitement had the lowest emotion recognition rate at 32.9%. Interestingly, SVM performance was lower in the validated database, probably due to a lack of data and maybe even due to a poorly performed validation. However, it is worth noting that Random Forests and ANN achieved relatively great emotion recognition rates on the validated database, respectively for anger (87,5%) and joy (55,6%). Finally, certain algorithms delivered significantly different F-measure values from recall ones, for specific emotions, for example, Random Forest anger recall was 32,2% higher that its respective F-measure.

To optimize speech emotion recognition systems, more research is encouraged on ensemble statistical classification, directing specific algorithms to recognize specific emotions or groups of emotions. Furthermore, more research is also encouraged in precision oriented algorithms, since speech emotion recognition researchers tend to only retrieve recall values.

Lövheim's three-dimensional emotion model has been applied, including, aside from Ekman's big 6 emotions, apathy and excitement. (Lövheim, 2012) These are unusual emotions to be recognized in emotion recognition systems; unexpectedly, apathy recognition rates were the highest using SVMs, whereas excitement were the lowest ones. We encourage more research based on Lövheim model, since more emotions are covered and it unites discrete and dimensional models, and neurology. Moreover, apathy being an emotion that reflects mental struggle, more research in this emotion is particularly encouraged.

Last but not least, the development of large acted emotion databases, featuring a larger number of utterances uttered by a larger set of actors and under a larger set of emotions, for instance, based on Lövheim's model, is important so to have a statistically large enough database for strong conclusions to be drawn.

# REFERENCES

Amir, N., Ron, S., & Laor, N. (2000). Analysis of an emotional speech corpus in Hebrew based on objective criteria. *Proceedings of the ISCA Workshop on Speech and Emotion*, 19–33.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614–636.

Banziger, T., Pirker, H., & Scherer, K. R. (2006). GEMEP - GEneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions. *Proceedings of the Fifth Conference on Language Resources and Evaluation*, *6*(May), 15–19.

Batliner, A. (2004). ’ You stupid tin box ’ - Children interacting with the AIBO robot : A cross-linguistic emotional speech corpus.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A Database of German Emotional Speech. In *INTERSPEECH* (pp. 1517–1520).

Castro, S. L., & Lima, C. F. (2010). Recognizing emotions in spoken language: A validated set of Portuguese sentences and pseudosentences for research on emotional prosody. *Behavior Research Methods*, *42*(1), 74–81.

Chen, L., Mao, X., Xue, Y., & Cheng, L. L. (2012). Speech emotion recognition: Features and classification models. *Digital Signal Processing*, *22*(6), 1154–

1160.

Chen, L. S., Tao, H., Huang, T. S., Miyasato, T., & Nakatsu, R. (1998). Emotion recognition from audiovisual information. In *IEEE International Workshop on Multimedia Signal Processing* (pp. 83–88).

Convenções e Transcrição Fonética. (n.d.). Retrieved from http://cvc.instituto-camoes.pt/cpp/acessibilidade/capitulo2_1.html

Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, *40*(1–2), 5–32.

Crumpton, J., & Bethel, C. L. (2016). A Survey of Using Vocal Prosody to Convey Emotion in Robot Speech. *International Journal of Social Robotics*, *8*(2), 271–285.

Darwin, C. (1872). *The expression of the emotions in man and animals*. London: Murray, John.

Dhall, A., Goecke, R., Joshi, J., & Gedeon, T. (2014). Emotion Recognition In The Wild Challenge 2014 : Baseline , Data and Protocol. In *ICMI* (pp. 461–466).

Dhall, A., Goecke, R., Joshi, J., & Gedeon, T. (2015). Video and Image based Emotion Recognition Challenges in the Wild : EmotiW 2015. In *ICMI* (pp. 423–426).

Dhall, A., Goecke, R., Joshi, J., Hoey, J., & Gedeon, T. (2016). EmotiW 2016 : Video and Group-Level Emotion Recognition Challenges. In *ICMI* (pp. 427–432).

Dhall, A., Goecke, R., Joshi, J., & Wagner, M. (2013). Emotion Recognition In The Wild Challenge 2013. In *ICMI* (pp. 509–515).

Dicionário Fonético. (n.d.). Retrieved December 10, 2016, from http://portaldalinguaportuguesa.org/index.php?action=fonetica&act=list&region=lbx

Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication*, *40*, 33–60.

Douglas-cowie, E., Cowie, R., & Schröder, M. (2000). A New Emotion Database: Considerations, Sources and Scope. In *ISCA Workshop on Speech and Emotion* (pp. 39–44).

Ekman, P. (1993). Facial expression and emotion. *The American Psychologist*.

Ekman, P., & Cordaro, D. (2011). What is Meant by Calling Emotions Basic. *Emotion Review*, *3*(4), 364–370.

Ekman, P., & Friesen, W. V. (1975). *Unmasking the face*. Palo Alto: Consulting Psychologists Press.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*(3), 572–587. http://doi.org/10.1016/j.patcog.2010.09.020

Ellsworth, P. C., & Scherer, K. R. (2003), Ellsworth, P., & Scherer, K. (2003). Appraisal processes in emotion. *Handbook of Affective Sciences*.

Engberg, I. S., & Hansen, A. V. (1996). Documentation of the danish emotional speech database des. *Internal AAU Report, Center for Person Kommunikation, Denmark.*

France, D. J., Shiavi, R. G., Member, S., Silverman, S., Silverman, M., & Wilkes, D. M. (2000). Acoustical Properties of Speech as Indicators of Depression and Suicidal Risk. *IEEE Transactions on Bio-Medical Engineering*, *47*(7), 829–837.

Gangamohan, P., Kadiri, S. R., & Yegnanarayana, B. (2016). Chapter 11: Analysis of Emotional Speech—A Review. In *Toward Robotic Socially Believable Behaving Systems - Volume I* (pp. 205–238).

Jovi, S. T., Ka, Z., & Rajkovi, M. (2004). Serbian emotional speech database: design, processing and evaluation. In SPIIRAS (Ed.), *SPECOM'2004: 9th Conference "Speech and Computer."* Saint-Petersburg.

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: A review. *International Journal of Speech Technology*, *15*(2), 99–117.

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*.

Lima, C. F., Castro, S. L., & Scott, S. K. (2013). When voices get emotional : A corpus of nonverbal vocalizations for research on emotion processing. *Behavior Research Methods*, *45*(4), 1234–1245.

Lövheim, H. (2012). A new three-dimensional model for emotions and monoamine neurotransmitters. *Medical Hypotheses*, *78*(2), 341–348. http://doi.org/10.1016/j.mehy.2011.11.016

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal theories of emotion: State of the art and future development. *Emotion Review*, *5*(2), 119–124.

Nicholson, J., Takahashi, K., & Nakatsu, R. (1999). Emotion recognition in speech using neural networks. In *Neural Information Processing, 1999. Proceedings. ICONIP '99. 6th International Conference on* (Vol. 2, pp. 495–501 vol.2).

Plutchik, R. (1991). *The Emotions*. Lanham, Maryland: University Press of America.

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots. *American Scientist*. http://doi.org/10.1511/2001.4.344

Ramakrishnan, S. (2012). Recognition of Emotion from Speech: A Review. In *Speech Enhancement, Modeling And Recognition – Algorithms And Applications* (pp. 121–138).

Ringeval, F., Schuller, B., Valstar, M., Cowie, R., & Pantic, M. (2015). AVEC 2015 – The 5th International Audio / Visual Emotion Challenge and Workshop. In *ACM Multimedia* (pp. 1335–1336).

Russell, J. (University of B. C., & Barrett, L. F. (Boston C. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of Personality and Social Psychology*, *76*(5), 808–819.

Saratxaga, I., & Navas, E. (2006). Designing and recording an emotional speech database for corpus based synthesis in Basque. In *Proc. of fifth international conference on Language Resources and Evaluation (LREC)* (pp. 2126–2129).

Sawoski, P. (2006). The Stanislavski System Growth and Methodology.

Sbattella, L., Colombo, L., Rinaldia, C., Tedesco, R., Matteucci, M., & Trivilini, A. (2014). Extracting Emotions and Communication Styles rom Prosody. *LNCS*, *8908*, 21–42.

Scherer, K. (2000). A cross-cultural investigation of emotion inferences from voice and speech : Implications for speech technology. *INTERSPEECH*, *1*, 379–382.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1), 227–256.

Schölkopf, B., Burges, C. J. C., & Smola, A. J. (1998). Introduction to Support Vector Learning. In B. Schölkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in Kernel Methods - Support Vector Advances in Kernel Methods* (pp. 1–16).

Schuller, B., Batliner, A., Burgoon, J. K., & Coutinho, E. (2016). The INTERSPEECH 2016 Computational Paralinguistics Challenge : Deception , Sincerity and Native Language. In *INTERSPEECH*.

Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, *53*(9), 1062–1087.

Schuller, B., Steidl, S., & Batliner, A. (2009). The INTERSPEECH 2009 Emotion Challenge. In *INTERSPEECH*.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. (2013). Paralinguistics in speech and language - State-of-the-art and the challenge. *Computer Speech and Language*, *27*(1), 4–39.

Schuller, B., Steidl, S., Batliner, A., Epps, J., Eyben, F., Ringeval, F., … Zhang, Y. (2014). The INTERSPEECH 2014 Computational Paralinguistics Challenge : Cognitive & Physical Load. In *INTERSPEECH* (pp. 427–431).

Schuller, B., Steidl, S., Batliner, A., Hantke, S., Florian, H., Orozco-arroyave, J. R., … Weninger, F. (2015). The INTERSPEECH 2015 Computational Paralinguistics Challenge: Nativeness, Parkinson's & Eating Condition. *INTERSPEECH*, 478–482.

Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., … Weiss, B. (2012). The INTERSPEECH 2012 Speaker Trait Challenge. In *INTERSPEECH* (pp. 254–257).

Schuller, B., Steidl, S., Batliner, A., Schiel, F., & Krajewski, J. (2011). The INTERSPEECH 2011 Speaker State Challenge. In *INTERSPEECH* (pp. 2–5).

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., … Kim, S. (2013). The INTERSPEECH 2013 Computational Paralinguistics Challenge : Social Signals , Conflict , Emotion , Autism. *INTERSPEECH*, 1–5.

Schuller, B., Valstar, M., Eyben, F., Cowie, R., & Pantic, M. (2012). AVEC 2012 – The Continuous Audio / Visual Emotion Challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction. ACM*.

Schuller, B., Valstar, M., Eyben, F., Mckeown, G., Cowie, R., & Pantic, M. (2011). AVEC 2011 – The First International Audio / Visual Emotion Challenge. In *Affective Computing and Intelligent Interaction* (pp. 415–424).

Staroniewicz, P., & Majewski, W. (2009). Polish Emotional Speech Database – Recording and Preliminary Validation. In *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions.* (pp. 42–49).

Tomkins, S. S. (1981). The quest for primary motives: Biography and autobiography of an idea. *Journal of Personality and Social Psychology*, *41*(2), 306–329. http://doi.org/10.1037//0022-3514.41.2.306

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres, M. T., … Pantic, M. (2016). AVEC 2016 – Depression , Mood , and Emotion Recognition Workshop and Challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge* (pp. 3–10).

Valstar, M., Schuller, B., Jarek, K., Cowie, R., & Pantic, M. (2014). AVEC 2014 : the 4th International Audio / Visual Emotion Challenge and Workshop. In *Proceedings of the 22nd ACM international conference on Multimedia* (pp. 1243–1244).

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., … Pantic, M. (2013). AVEC 2013 – The Continuous Audio / Visual Emotion and Depression Recognition Challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge* (pp. 3–10).

Vapnik, V. (1998). Three Remarks on the Support Vecotr Method of Function Estimation. In B. Schölkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in Kernel Methods - Support Vector Advances in Kernel Methods* (pp. 25–

42).

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, *48*(9), 1162–1181.

Ververidis, D., Ververidis, D., Kotropoulos, C., & Kotropoulos, C. (2003). A Review of Emotional Speech Databases. In *Proceedings of the 9th Panhellenic Conference on Informatics (PCI)* (pp. 560–574).

Wennerstrom, A. (2001). *The Music of Everyday Speech*.

Wundt, W. (1897). Outlines of psychology (C.H. Judd, Trans.).

# Appendix A

The following pages contain Weka's statistical classification algorithms output performed on the full database and on the validated one. The first three pages concern the full database, whereas the following ones concern the validated database. We begin by presenting SVM results, then we present Random Forest results and finally we present ANN results.

```
=== Run information ===

Scheme:       weka.classifiers.functions.LibSVM -S 0 -K 1 -D 3 -G 1.0E-5 -R 10.0 -N 0.5
-M 40.0 -C 100000.0 -E 0.01 -P 0.1 -model -seed 1
Instances:    718
Attributes:   458
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         350               48.7465 %
Incorrectly Classified Instances       368               51.2535 %
Kappa statistic                          0.4234
Mean absolute error                      0.1139
Root mean squared error                  0.3375
Relative absolute error                 57.6601 %
Root relative squared error            107.387  %
Total Number of Instances              718

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC
Area  Class
                 0,582    0,031    0,697      0,582    0,634      0,597    0,775
0,452     anger
                 0,338    0,089    0,321      0,338    0,329      0,243    0,624
0,182     joy
                 0,329    0,085    0,325      0,329    0,327      0,243    0,622
0,181     excitement
                 0,638    0,061    0,567      0,638    0,600      0,548    0,788
0,402     neutral
                 0,525    0,077    0,462      0,525    0,491      0,424    0,724
0,295     fear
                 0,425    0,082    0,395      0,425    0,410      0,333    0,672
0,232     sadness
                 0,363    0,058    0,439      0,363    0,397      0,332    0,652
0,230     disgust
                 0,775    0,024    0,805      0,775    0,790      0,764    0,876
0,649     apathy
                 0,413    0,071    0,423      0,413    0,418      0,346    0,671
0,240     surprise
```

```
Weighted Avg.    0,487    0,064    0,493     0,487    0,489     0,425    0,712
0,318

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i   <-- classified as
 46  3 16  1  3  0  1  0  9 |  a = anger
  3 27 15 11  6  1 10  1  6 |  b = joy
 11 14 26  1  7  1  5  0 14 |  c = excitement
  0 10  0 51  2 10  3  1  3 |  d = neutral
  0  4  4  6 42 13  3  1  7 |  e = fear
  0  2  0  8 17 34  9 10  0 |  f = sadness
  2 13  5  6  8 10 29  1  6 |  g = disgust
  0  0  0  3  0 15  0 62  0 |  h = apathy
  4 11 14  3  6  2  6  1 33 |  i = surprise
Options: -P 100 -I 20000 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

RandomForest

Bagging with 20000 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

=== Error on training data ===

Correctly Classified Instances         718               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0717
Root mean squared error                  0.1141
Relative absolute error                 36.2863 %
Root relative squared error             36.3167 %
Total Number of Instances              718


=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC
Area  Class
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000
1,000     anger
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000
1,000     joy
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000
1,000     excitement
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000
1,000     neutral
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000
1,000     fear
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000
1,000     sadness
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000
1,000     disgust
```

```
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000
1,000     apathy
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000
1,000     surprise
Weighted Avg.    1,000    0,000    1,000      1,000    1,000      1,000    1,000
1,000


=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i   <-- classified as
 79  0  0  0  0  0  0  0  0 |  a = anger
  0 80  0  0  0  0  0  0  0 |  b = joy
  0  0 79  0  0  0  0  0  0 |  c = excitement
  0  0  0 80  0  0  0  0  0 |  d = neutral
  0  0  0  0 80  0  0  0  0 |  e = fear
  0  0  0  0  0 80  0  0  0 |  f = sadness
  0  0  0  0  0  0 80  0  0 |  g = disgust
  0  0  0  0  0  0  0 80  0 |  h = apathy
  0  0  0  0  0  0  0  0 80 |  i = surprise


=== Run information ===

Scheme:       weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0
-S 0 -E 20 -H a
Instances:    718
Attributes:   458

=== Stratified cross-validation ===

Correctly Classified Instances         188               26.1838 %
Incorrectly Classified Instances       530               73.8162 %
Kappa statistic                          0.1695
Mean absolute error                      0.1678
Root mean squared error                  0.3586
Relative absolute error                 84.9676 %
Root relative squared error            114.1095 %
Total Number of Instances              718


=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC
Area  Class
                 0.278    0.042    0.449      0.278    0.344      0.293    0.800
    0.368     anger
                 0.263    0.130    0.202      0.263    0.228      0.118    0.585
    0.150     joy
                 0.165    0.075    0.213      0.165    0.186      0.100    0.655
    0.167     excitement
                 0.250    0.096    0.247      0.250    0.248      0.154    0.709
    0.255     neutral
```

```
                0.288    0.108    0.250      0.288    0.267      0.169    0.674
    0.221     fear
                0.275    0.127    0.214      0.275    0.240      0.133    0.673
    0.212     sadness
                0.250    0.092    0.253      0.250    0.252      0.158    0.652
    0.200     disgust
                0.375    0.052    0.476      0.375    0.420      0.360    0.802
    0.396     apathy
                0.213    0.108    0.198      0.213    0.205      0.101    0.646
    0.201     surprise
Weighted Avg.    0.262    0.092    0.278      0.262    0.266      0.176    0.688
    0.241


=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i   <-- classified as
 22  5 12  6  6  6  9  1 12 |  a = anger
  4 21  8 13  4 12  4  4 10 |  b = joy
  6 14 13  4  9  9 10  2 12 |  c = excitement
  3 16  3 20 13 11  3  7  4 |  d = neutral
  3 11  8  9 23 10  6  3  7 |  e = fear
  4 10  3 10 11 22  7  3 10 |  f = sadness
  2 15  3  5  8  8 20  8 11 |  g = disgust
  2  5  6  4  8 11 11 30  3 |  h = apathy
  3  7  5 10 10 14  9  5 17 |  i = surprise
```

```
=== Run information ===

Scheme:       weka.classifiers.functions.LibSVM -S 0 -K 1 -D 3 -G 1.0E-7 -R 10.0 -N 0.05
-M 40.0 -C 1.0E7 -E 0.001 -P 0.1 -model -seed 1
Instances:    116
Attributes:   53
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          51               43.9655 %
Incorrectly Classified Instances        65               56.0345 %
Kappa statistic                          0.3505
Mean absolute error                      0.1245
Root mean squared error                  0.3529
Relative absolute error                 64.3981 %
Root relative squared error            113.5292 %
Total Number of Instances              116
```

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC
Area  Class
                 0.750    0.087    0.692      0.750    0.720      0.644    0.832
0.571     anger
                 0.636    0.117    0.560      0.636    0.596      0.495    0.760
0.425     neutral
                 0.385    0.039    0.556      0.385    0.455      0.408    0.673
0.283     excitement
                 0.182    0.086    0.182      0.182    0.182      0.096    0.548
0.111     fear
                 0.333    0.028    0.500      0.333    0.400      0.369    0.653
0.218     apathy
                 0.100    0.123    0.071      0.100    0.083      -0.020   0.489
0.085     sadness
                 0.273    0.095    0.231      0.273    0.250      0.165    0.589
0.132     surprise
                 0.333    0.037    0.429      0.333    0.375      0.332    0.648
0.195     joy
                 0.286    0.028    0.400      0.286    0.333      0.303    0.629
0.157     disgust
Weighted Avg.    0.440    0.079    0.453      0.440    0.441      0.369    0.680
0.302

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i   <-- classified as
 18  1  1  2  0  0  2  0  0 |  a = anger
  0 14  0  0  0  2  4  0  2 |  b = neutral
  3  1  5  1  0  1  0  2  0 |  c = excitement
  2  1  0  2  0  4  1  0  1 |  d = fear
  0  0  0  1  3  5  0  0  0 |  e = apathy
  0  2  1  2  3  1  0  1  0 |  f = sadness
  2  3  1  1  0  1  3  0  0 |  g = surprise
  1  1  1  1  0  0  2  3  0 |  h = joy
  0   2  0  1  0  0  1  1  2 |  i = disgust
=== Run information ===

Scheme:       weka.classifiers.trees.RandomForest -P 100 -I 20000 -num-slots 1 -K 0 -M
1.0 -V 0.001 -S 1
Instances:    116
Attributes:   53

Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

RandomForest

Bagging with 20000 iterations and base learner
```

```
weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          45               38.7931 %
Incorrectly Classified Instances        71               61.2069 %
Kappa statistic                          0.2558
Mean absolute error                      0.1809
Root mean squared error                  0.2943
Relative absolute error                 93.5315 %
Root relative squared error             94.6869 %
Total Number of Instances              116

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC
Area  Class
                 0.875    0.337    0.404      0.875    0.553      0.438    0.861
0.616     anger
                 0.682    0.287    0.357      0.682    0.469      0.322    0.806
0.577     neutral
                 0.000    0.010    0.000      0.000    0.000      -0.033   0.709
0.245     excitement
                 0.000    0.057    0.000      0.000    0.000      -0.076   0.646
0.137     fear
                 0.222    0.009    0.667      0.222    0.333      0.359    0.919
0.620     apathy
                 0.200    0.028    0.400      0.200    0.267      0.237    0.822
0.274     sadness
                 0.091    0.000    1.000      0.091    0.167      0.288    0.730
0.368     surprise
                 0.222    0.019    0.500      0.222    0.308      0.298    0.925
0.557     joy
                 0.286    0.000    1.000      0.286    0.444      0.523    0.649
0.299     disgust
Weighted Avg.    0.388    0.135    0.431      0.388    0.319      0.271    0.794
0.445





=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i   <-- classified as
 21  3  0  0  0  0  0  0  0 |  a = anger
  5 15  0  1  0  1  0  0  0 |  b = neutral
 10  2  0  0  0  0  0  1  0 |  c = excitement
  5  4  0  0  0  2  0  0  0 |  d = fear
  2  3  0  2  2  0  0  0  0 |  e = apathy
```

```
1  4  0  2  1  2  0  0  0 |  f = sadness
3  6  0  0  0  0  1  1  0 |  g = surprise
3  2  1  1  0  0  0  2  0 |  h = joy
2  3  0  0  0  0  0  0  2 |  i = disgust
```

```
=== Run information ===

Scheme:       weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0
-S 0 -E 20 -H a
Instances:    116
Attributes:   53

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          42               36.2069 %
Incorrectly Classified Instances        74               63.7931 %
Kappa statistic                          0.2608
Mean absolute error                      0.1428
Root mean squared error                  0.3227
Relative absolute error                 73.8536 %
Root relative squared error            103.8129 %
Total Number of Instances              116

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC
Area  Class
                 0.500    0.163    0.444      0.500    0.471      0.323    0.771
0.435     anger
                 0.500    0.128    0.478      0.500    0.489      0.366    0.832
0.644     neutral
                 0.154    0.107    0.154      0.154    0.154      0.047    0.670
0.212     excitement
                 0.273    0.086    0.250      0.273    0.261      0.180    0.724
0.231     fear
                 0.222    0.037    0.333      0.222    0.267      0.223    0.822
0.387     apathy
                 0.100    0.075    0.111      0.100    0.105      0.026    0.650
0.179     sadness
                 0.364    0.057    0.400      0.364    0.381      0.320    0.685
0.324     surprise
                 0.556    0.056    0.455      0.556    0.500      0.456    0.892
0.397     joy
                 0.286    0.028    0.400      0.286    0.333      0.303    0.814
0.308     disgust
Weighted Avg.    0.362    0.099    0.356      0.362    0.357      0.262    0.764
0.383

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i   <-- classified as
 12  2  7  2  0  0  1  0  0 |  a = anger
  3 11  1  1  0  1  2  1  2 |  b = neutral
  5  1  2  1  0  1  0  3  0 |  c = excitement
  3  1  1  3  1  2  0  0  0 |  d = fear
  2  1  0  1  2  2  1  0  0 |  e = apathy
 0  2  0  2  3  1  0  2  0 |  f = sadness
 1  2  1  2  0  1  4  0  0 |  g = surprise
 0  1  1  0  0  0  1  5  1 |  h = joy
 1  2  0  0  0  1  1  0  2 |  i = disgust
```