UNIVERSITAT DE BARCELONA

FUNDAMENTALS OF DATA SCIENCE MASTER'S THESIS

Non-acted Multi-view Audio-Visual Dyadic Interactions Project

Non-verbal Emotion Recognition in Dyadic Scenarios and Speaker Segmentation

Author:
Pablo LÁZARO HERRASTI

Supervisor:
Dr. Sergio ESCALERA
Co-Supervisor:
Cristina PALMERO

A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamentals of Data Science

in the

Facultat de Matemàtiques i Informàtica

September 2, 2019

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

Non-acted Multi-view Audio-Visual Dyadic Interactions Project

Non-verbal Emotion Recognition in Dyadic Scenarios and Speaker Segmentation

by Pablo LÁZARO HERRASTI

Abstract of the Project

Socially-intelligent systems have to be capable of accurately perceiving and inferring the personality and other particularities of different individuals, so as to provide a more effective, empathic, and natural tailored communication. To embody this human likeness into such systems, it is imperative to have a deeper understanding of real human-human interactions first, to computationally model both individual behavior and interpersonal interinfluence. However, there is a lack of publicly-available audiovisual databases of non-acted face-to-face dyadic interactions, which cover the richness and complexity of social communications in real life. In this project, we collected the first of its kind non-acted audio-vidual multi-view dataset of dyadic interactions. The main goals of this dataset and associated research is to analyze human communication from a multidisciplinary perspective (i.e. technological, sociological and psychological) and to research and implement new paradigms and technologies of interpersonal behavior understanding. It is expected to move beyond automatic individual behavior detection and focus on the development of automatic approaches to study and understand the mechanisms of perception of and adaptation to verbal and non-verbal social signals in dyadic interactions, taking into account individual and dyad characteristics. In addition to the collection of more than 80 hours of dyadic interactions including 150 participants performing cognitive tasks designed by the psychologists, this project performed a proof of concept analysis of different technical challenges included in the database:

- Setup design, calibration and synchronization of 6 HD cameras, 2 HD egocentric cameras, 2 wrist heart rate monitors, 2 lapel microphones and 1 ambient microphone.
- Multi-view joint optimization of hand and body skeleton poses for enhanced hand and body pose recovery.
- Speaker audio segmentation.
- Audio-visual spatio-temporal modeling of human emotions.
- Multi-task face attributes analysis

The different contributions are presented and justified in the context of their respective state-of-the-art, evaluated on proper public datasets, and finally tested as a proof of concept evaluation on the recently designed dyadic dataset. Detailed discussion of the implemented work and its associated future research is provided.

Abstract of this Master Thesis

In particular, this Master Thesis is focused on the development of baseline **Emotion Recognition System** in a dyadic environment using raw and handcraft audio features and cropped faces from the videos. This system is analyzed at frame and utterance level without temporal information. As well, a baseline **Speaker Segmentation System** has been developed to facilitate the annotation task. For this reason, an exhaustive study of the state-of-the-art on emotion recognition and speaker segmentation techniques has been conducted, paying particular attention on Deep Learning techniques for emotion recognition and clustering for speaker segmentation.

While studying the state-of-the-art from the theoretical point of view, a dataset consisting of videos of sessions of dyadic interactions between individuals in different scenarios has been recorded. Different attributes were captured and labelled from these videos: body pose, hand pose, emotion, age, gender, etc. Once the architectures for emotion recognition have been trained with other dataset, a proof of concept is done with this new database in order to extract conclusions. In addition, this database can help future systems to achieve better results.

A large number of experiments with audio and video are performed to create the emotion recognition system. The IEMOCAP database is used to perform the training and evaluation experiments of the emotion recognition system. Once the audio and video are trained separately with two different architectures, a fusion of both methods is done. In this work, the importance of preprocessing data (face detection, windows analysis length, handcrafted features, etc.) and choosing the correct parameters for the architectures (network depth, fusion, etc.) has been demonstrated and studied.

On the other hand, the experiments for the speaker segmentation system are performed with a piece of audio from IEMOCAP database. In this work, the prerprocessing steps, the problems of an unsupervised system such as clustering and the feature representation are studied and discussed.

Finally, the conclusions drawn throughout this work are exposed, as well as the possible lines of future work including new systems for emotion recognition and the experiments with the database recorded in this work.

Master Thesis Student Contributions

This project has been accomplished by a group of 4 Master students. The contribution of this master thesis within the whole project is explained in the following lines.

Study of the state-of-the-art of the emotion recognition problem using audiovisual sources. Deliver a emotion recognition system using Deep Learning techniques based on unimodal audio features, raw audio and faces, and their possible fusion. On the other hand, study of state-of-the-art of the speaker segmentation problem using audio sources and unsupervised learning techniques such as **Spectral Clustering**. Help and participate during the recordings of the different sessions of the

Face-to-face Dyadic Interaction Dataset placing and collecting the setup and attending the participants. Also annotate this database labeling the utterances of the videos.

Acknowledgements

First of all, I would like to express my special thanks of gratitude to my supervisor Sergio Escalera and my Cosupervisor Cristina Palmero for their able guidance and support in completing my project. Withouth their persistent help this project would not have been possible. I would like to express my thanks as well to Javier Selva for his patience with the recordings and the hours spent solving questions. Last, I would like to thanks to all the professor that I have been having throughout the master, thanks to them I have been able to face different problems and apply everything learned.

On the other hand, I would like to thanks my project partners Rubén Barco, Andreu Masdeu y Aleix Casellas. We have been working hard during the last months to finish this project and I couldn't have better partners than you. We decided to develop our master thesis in group and I can say now that it was a success. Despite the distance, I am sure that we are going to keep in touch and I feel very proud to have met you.

Last, I would like to thanks my family and girlfriend. It has been a very hard year with many difficult moments, but in the end it has come to an end in the best way. When I decided to do my master in Barcelona, you offered me your support from the first moment and I have worked very hard to give it back to you. Thank you very much for always being by my side.

Contents

Chapter 1

Introduction

In the computer vision and machine learning communities, one ongoing line of research is modelling human interaction and behaviour. Social signal processing and affective computing are two research fields that aim at understanding these interactions by extracting audio-visual features and combining them to build bigger constructs such as personality or dominance. In particular, **dyadic interactions** between individuals is a crucial aspect in studying how human beings react to the environment and with each other. Exploring and analyzing dyadic sessions may result in a better understanding of human interactions and behaviour. But as for every purpose or plan intended in machine learning there is always one mandatory thing to have: **data**. In particular, data for dyadic interactions between humans is quite rare to find, for instance, there is the **IEMOCAP** dataset. However, data for this sort of interaction analyzed from a psychological perspective does not exist.

One goal of this master thesis is precisely to assist in the creation of a **dataset** consisting of videos of sessions of dyadic interactions between individuals in different scenarios. Different attributes were captured and labelled from these videos: body pose, hand pose, emotion, age, gender, etc. Furthermore, all participants of the dataset were given a personality test survey, which means that each individual personality was labelled. This represented a very ambitious challenge and makes of this dataset a very unique object which may be very helpful for the computer vision and machine learning communities in the future.

Although many attributes could be extracted from the data, one of the main goals in this group master thesis is to develop a non-verbal **emotion recognition** system using audio and frames sources from videos at frame level. This two sources are proposed because facial expressions can help the audio to aim better the real emotion of the subject. Deep learning techniques have been used to be able to model the implicit emotions within a dyadic conversation between two subjects. Due to the state of the art, one can affirm that predicting emotions using deep learning is very challenging nowadays, due to different problems that are described in the following chapters.

When two subjects are talking between them, the emotion can be modeled in every piece of time without taking into account the course of the conversation. Instead, this work proposes to divide the conversation in utterances and predict the emotion of each utterance. From that idea came the second goal of my part in this group master thesis: use deep learning to **segment the audio in utterances**. Again, this is a very challenging task, mainly because the overlapping parts that a conversation may have.

This master thesis includes a discussion on the creation and contents of the dataset, a chapter discussing the current state-of-the-art in Emotion Recognition and Speaker

Segmentation, a chapter describing the methods chosen, a chapter analyzing the results obtained and finally a chapter containing the conclusions and future work.

Chapter 2

Emotion Recognition

Humans verbally communicate by speech and language. This enables faster sharing of messages, conveying of ideas and spreading of inventions. Communication between humans is actually not just what humans say, but also how they say it. Furthermore, facial expressions, as a part of non-verbal communication, are responsible for about 55%, voice intonation for about 38% and actual words for 7% of the message perception [4]. Emotion recognition is a very actively growing field of research. The **goal of human emotion recognition** is to automatically classify user's temporal emotional state basing on some input data.

As it was mentioned, emotions play a major role in a Human life. At different kind of moments or time Human face reflects that how he/she feels or in which mood he/she is. Humans are capable of producing thousands of *facial actions* during communication that vary in complexity, intensity, and meaning. Emotion or intention is often communicated by subtle changes in one or several discrete features [23]. Moreover, *speech communication* contains paralinguistic information of the speaker such as tone or pitch of voice, what may be helpful to predict an emotion during a conversation. Although enormous efforts are invested in recognizing the emotions from speech, still much research is needed. For these reasons, **voice and facial expressions** are the ones that have been exploited during the last years to develop a robust emotion recognition system.

Notably, most approaches try to obtain these emotions from *posed facial expression*. In these approaches, is easier to achieve higher performance because it only focuses on facial expression without being immersed in a conversation. Nowadays, there exist different methods that model these posed facial expressions, but there is a lack of methods trying to model facial expressions during a conversation between two subjects. Considering the inter-personal influences that thrive in the emotional dynamics of dialogues becomes more challenging the paradigm of emotion recognition.

Definition 1 ***Emotion recognition*** *is the process of identifying human emotion, most typically from facial expressions as well as from verbal expressions.*

[16] said that *an emotion is a reaction to stimuli that lasts for seconds or minutes*. As it is described, there are different categories and measurements of emotion, depending on the final purpose. In the last years, **deep learning** techniques have became more important for researchers to improve traditional methods such as hand-crafted features. Later, it will be discussed the state of the art and the evolution that the main emotions recognition techniques have undergone over the last years.

2.1 Applications

Although all the effort is over developing new systems and new methods to recognize an emotion in different scenarios, it is very complicated to imagine the benefits that these systems can obtain with in different industries. This section briefly describes some real applications and discuss how useful are for the society [16][20].

- **Software usability**: There is a lot of evidence, that human emotions influence interactions with software products. There is also a record of investigation on how products can influence human feelings and those feelings make people buy or not. Therefore investigating emotions induced by products is an object of interest of designers, investors, producers and customers, as well. Software usability depends on multiple quality factors, such as functionality, reliability, interface design, performance and so on.
- **Education**: Some emotional states support learning processes and other suppress them. The distinction of the two groups of emotional states in some cases is not obvious, for example such positive mood as hilarity is not good for learning processes, while slightly negative emotional states foster critical thinking and are appropriate for analytical tasks. Automatic emotion recognition algorithms can help to explore this phenomena by making assessments of learner emotional states more objective than typical questionnaire-based investigations.
- **Enhanced websites customization**: With the grow of the Internet, service providers collect more and more information about their users. Based on these data, content, layout and ads are displayed according to the user's profile. Adding information about the emotions of users could provide more accurate personality models of the users.
- **Video games**: There are a lot of reasons that can influence upon human's behavior during playing the game. They could be divided into factors connected with the game, such as increasing monotony or becoming accustomed player, and to game independent factors which are connected with current physical and mental condition of the player. The first group of reasons may be in some extent predicted or estimated by the game designer but that is impossible to the reasons of the other group. That is why the real-time recognition of player's affect may become such important for video games industry in the nearest future. Video games that are able to dynamically react to the recognized current player's emotions are called truly affect aware video games.
- **Others**: public services (government, airports security), satisfaction in call centres, determining patients feeling and comfort level about the treatment, determining fatigue in the case of driving and alerting in advance, facial emotion detection in interviews, etc.

2.2 Affective Computing and Emotion Classification

Affective computing is the study and development of systems and devices that can recognize, interpret, process, and simulate human affects. Then, it can be said that Emotion recognition is a sub-field of Affective computing. In Affective computing,

researchers usually compute emotions using different sources, such as facial expressions or speech.

In speech recognition as well as emotion recognition from speech, data material is an important resource. First, it has to be recorded, then it is necessary to reprocess the data, i.e. to transcribe and annotate the recordings. One of the main challenge of emotion recognition is how to annotate accurately the ground truth emotion of a speaker or subject. To achieve a level of confidence in emotions ground truth, there are several ways to label an emotion. While there is ongoing debate regarding what the defining features of an emotion are, emotions can be broadly conceptualized as "*adaptive action tendencies that occur in response to changes inside or outside the organism, specifically changes that challenge states and systems necessary for survival*". Emotions can be measured across at least three different modes of response: physiological, cognitive and behavioral. Within this broad approach to defining emotion, there are at least two major perspectives adopted for more specifically understanding emotions. The first approach can be referred to as a *discrete emotions* perspective. Within this perspective, emotions are generally thought to be specific, cross-cultural, innate, systemic responses. For example, fear can be thought of as a discrete emotion that across cultures is characterized by elevated physiological arousal, perceived threat, and escape behavior. In contrast to the discrete emotions perspective, the dimensional perspective suggests that emotions are responses to environmental stimuli that vary along dimensions of key features or characteristics. Within the dimensional perspective on emotion, theory and empirical work has converged to suggest at least three core features to an emotional response (i.e., valence/pleasure, arousal, and dominance/control). These dimensions are theorized to respond somewhat independently resulting in dimensions that can differentially respond across time to emotion eliciting events. Now, the next list describes three important methods for annotating emotions [34]:

- **Basic Emotions**:

  Basic Emotions is a Discrete method. Although the view that some emotions are more "basic" than others is widely accepted by emotion theorists, there is little agreement on which emotions should be included in the list of the basic ones. Their number varies depending on the theory. The most popular list, sometimes referred to as "The Big Six," was used by Ekman et al. (1969) in their research on universal recognition of emotion from facial expression. The list included happiness, sadness, fear, surprise, anger, and disgust, which are still the most commonly accepted candidates for basic emotions. Over the years some theorists, including Ekman, have shortened or expanded the list. For instance, Plutchnik (1980) added acceptance and anticipation, Ekman (1999; Ekman and Cordaro 2011) added contempt, and Levenson (2011) added interest, relief, and love. More recently, other candidates for basic emotions have been proposed, e.g., love or jealousy (Sabini and Silver 2005). Some authors have also used their own terminology. For instance, Panksepp (2007; Panksepp and Watt 2011) listed play, panic/grief, rage, seeking, fear, lust, and care as the basic ("primary-process") emotions. As well, some researchers extended it with the categories neutral, since certain parts of communication can also be non-emotional, and other to give the annotators the opportunity to rate emotions which do not fit the given labels. The decision whether a particular emotion qualifies as basic is based on a set of criteria. Although these criteria vary across theories, many authors agree that a basic emotion should be associated

with distinctive universal nonverbal expression, distinctive neural and physiological components, distinctive subjective experience, and distinctive regulatory and motivational properties [28]. As these Basic Emotions represent primary emotions, mixed or blended emotions cannot be represented.

- **Geneva Emotion Wheel (GEW)**:

  The Geneva Emotion Wheel (GEW) is a dimensional method and consists of discrete emotion terms corresponding to emotion families that are systematically aligned in a circle. Underlying the alignment of the emotion terms are the two dimensions valence (negative to positive) and control (low to high), separating the emotions in four quadrants: Negative/low control, negative/high control, positive/low control, and positive/high control. As it is shown in Fig. 2.1, the response options are "spikes" in the wheel that correspond to different levels of intensity for each emotion family from low intensity (towards the center of wheel) to high intensity (toward the circumference of the wheel). Also, in the very center of the wheel, the response options "no emotion" and "other emotion" is offered. As the labellers can assign at most three labels to each utterance, mixed emotional states can also be labelled by this method. The GEW has previously been used in a variety of contexts, ranging from managers' affect during decision making (Tran, 2004) to the evaluation of body movements and consumer experiences. These studies show that the GEW is a particularly useful measurement instrument under time pressure and with repeated measurements [33].

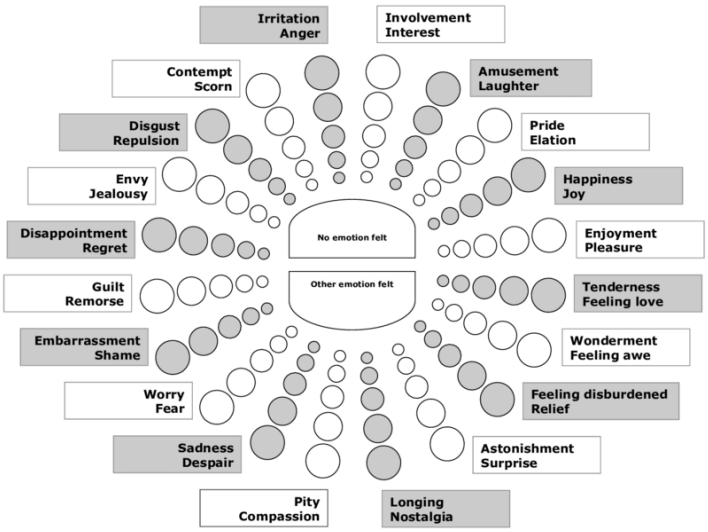

FIGURE 2.1: GEW with 40 emotion terms arranged in 20 emotion families.

- **Self Assessment Manikins (SAM)**:

  The Self-Assessment Manikin (SAM) is a dimensional method and a picture-oriented questionnaire developed to measure an emotional response. It contains five images for each of the three affective dimensions that the participant rates on either a 9- or 21-point scale. The annotators can label the utterance on three different axes, valence-arousal-dominance (VAD), without needing to cope with emotional categories. The advantage of this method is the absence of discrete emotional categories, so that different perceptions cannot influence

the assigned emotion label. The VAD-space is divided into octants, to allow a comparison to categorial emotion theories, and a neutral centroid is placed in the centre of the space. In the typical implementation it is used axes with the following classification: +V and V for positive and negative valence, +A and A for aroused and unaroused, and +D and D for dominant and submissive. Additionally, emotions can be identified located on a boundary area between octants as "mixed emotions" and count them proportionally for all corresponding octants. The SAM is an imagery-based measure that therefore can be thought of as language-free. Thus, use of the SAM is not circumscribed to any one culture, and it can be easily understood and appropriate for use in different countries. Another feature of the SAM that makes it widely applicable is that it is brief. Due to its brevity, it can be used to capture emotional responses to a wide array of emotion elicitation methods [7].

2.3 Audio and Video Features

Development of machines with emotional intelligence has been a long-standing goal of AI. With the increasing infusion of interactive systems in our lives, dialogue videos have proliferated across the internet through platforms like movies, webinars and video chats. This project tries to detect emotions in videos of **dyadic conversations**. A dyadic conversation is a form of a dialogue between two entities. This kind of videos convey information through three channels: audio, video, and text. This master thesis is focused on developing a real time emotion recognition system. That is why only audio an video sources are going to be used because text does not allow to implement a real time system in this scenario.

One of the challenges in the emotion recognition problem is how to represent the data. An important part of the literature focuses on the ways to represent audio/video contents by features that can be subsequently used by classifiers.

2.3.1 Audio

Speech is considered to be a very complex signal, since apart from the meaning it carries information regarding the speaker's identity and language and his/her emotion. When the type of the information is considered, it is common to divide the methods that adopt speech into two distinct categories [27]: **explicit** or linguistic information, which concerns articulated patterns by the speaker; and **implicit** or paralinguistic information, which concerns the variation in pronunciation of the linguistic patterns. This last one approach ignore the content of speech and instead focus on associating low-level features to emotions. Extracted features may either be low-level descriptors or statistics extracted on these descriptors. The main disadvantage of the linguistic models is that they do not typically provide a language-independent model. Each language has its own specifics and is subject to cultural differences. As such, there might exist a plethora of different sentences, speakers, speaking styles and rates.

There are some different methods for treating the audio data, but all of them have a common part: the audio is first divided into audio-segments using windows (binary temporal mask). Normally, these audio windows have a length between 30-200 ms and are processed with techniques of voice normalization and intensity thresholding.

Once the audio is segmented using sliding windows, there are three principal methods for obtaining the audio features:

1. **Hand-crafted features**: The most common software used for this task is **openSMILE**, which is a toolkit that unites feature extraction algorithms from the speech processing and the Music Information Retrieval communities. Audio low-level descriptors such as CHROMA and CENS features, loudness, Mel-frequency cepstral coefficients, perceptual linear predictive cepstral coefficients, linear predictive coefficients, line spectral frequencies, fundamental frequency and formant frequencies are supported. Delta regression and various statistical functionals can also be applied to the low-level descriptors [11]. Literature shows that there exists a high correlation between many statistical measures of speech with speakers' emotion. For example, high pitch and fast speaking rate often denote anger while sadness associates low standard deviation of pitch and slow speech rate.

2. **1D-Convolution**: Instead of using a hand-crafted method, each raw audio window is passed through a 1D-Convolutional Neural Network to obtain a feature vector.

3. **Spectrogram + 2D-Convolution**: A **spectrogram** is a visual representation of the spectrum of frequencies of a signal as it varies with time. The process for obtaining a spectogram is the following one: first, the audio signal is divided in windows; then, the frequency representation of each window is computed applying the STFT; and finally, each frequency representation is plotted in time [Fig 2.2]. They are used extensively in the fields of music, sonar, radar, and speech processing. Spectrograms of audio can be used to identify spoken words phonetically and they are highly used in the studies of *phonetics* and *speech synthesis*. The reason of using spectograms is because they join the temporal and frequency information in one image. The objective is to obtain the spectrogram of each audio window and pass them through a 2D-Convolutional Neural Network to obtain a feature vector.

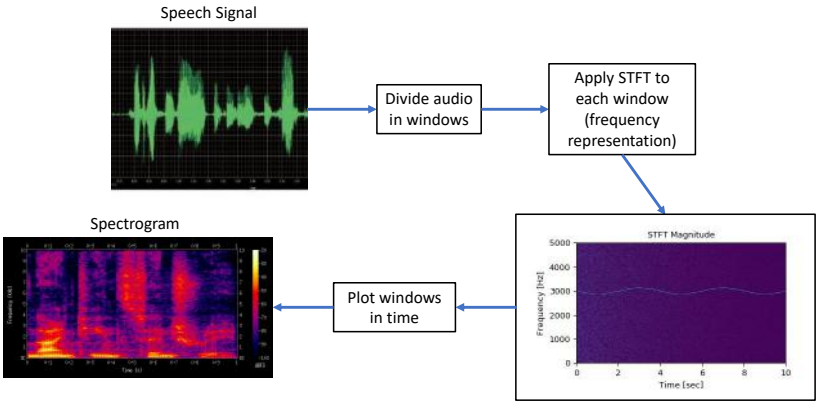

FIGURE 2.2: Computation process of a spectrogram.

2.3.2 Videos and Images

Visual indicators such as facial expressions are key to understand emotions. But image and video classification and the process of preprocessing and extracting features remains challenging. Visual features can be broadly categorized into three classes:

1. **Local handcrafted approaches**. They include handcrafted features and their corresponding encoding methods. There are lots of methods for this purpose, but the following three are the most common:

   (a) **SIFT** (Scale Invariant Feature Transform): This method is invariant at scale and rotation and works in two steps: detection of feature point as the first step and feature description as the second one. At the beginning of the procedure, gradients and orientations of pixels are computed. This is done for each key point and its neighborhood. The feature vector is a combination of the orientation of histograms within the sub-regions around the feature point.

   (b) **SURF** (Speeded-Up Robust Features): It is similar to SIFT, but it is faster and detects the key points by using the determinant of the Hessian matrix.

   (c) **HOG** (Histogram of Oriented Gradient): At each pixel, the image gradient vector is calculated and converted to an angle, voting into the corresponding orientation bin with a vote weighted by the gradient magnitude. Votes are accumulated over the pixels of each cell. The cells are grouped into blocks and a robust normalization process is run on each block to provide strong illumination invariance. The normalized histograms of all of the blocks are concatenated to give the window level visual descriptor vector for learning.

2. **Learning based approaches**. They are mainly represented by CNNs (Convolutional Neural Networks) for image recognition. There are two different approaches. The first one consists of using **2D-CNNs** to obtain the features for each frame of the video independently. This 2D networks can be useD either as feature extractors and return a feature vector for each input, or can be followed by a classifier to obtain an image classification system. The second approach aims to include the spatio-temporal information across frames or utterances by using **3D-CNNs** or adding **RNNs** (Recurrent Neural Networks) layers after the 2D-CNN. This architectures and some examples of these methods are explained in section 2.5.

3. **Facial Action Coding System (FACS)**: It is a comprehensive, anatomically based system for describing all visually discernible facial movement. It breaks down facial expressions into individual components of muscle movement, called Action Units (AUs). There are 49 different Action Units in total. Each of them represent a different movement of a part of the face, so it is easy to describe each emotion by the combination of these Action Units. In Fig. 2.3 there are some examples of the most common facial expressions and the Action Units that are involved in them.

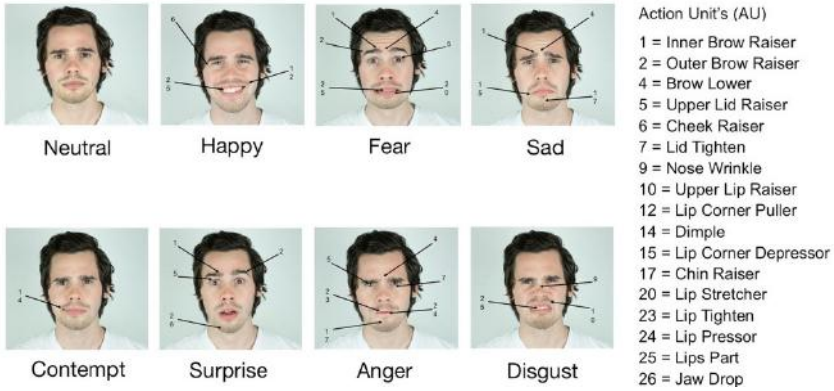

FIGURE 2.3: Example of emotion expressions and their respective Action Units.

2.4 State of the art - Databases

Since emotions are expressed through a combination of verbal and non-verbal channels, a joint analysis of speech and gestures is required to understand expressive human communication. In this context, one of the major limitations in the study of emotion expression is the **lack of databases** with genuine interaction that comprise integrated information from most of these channels. For this reason, in this section is described the most used databases for emotion expression, depending on the final objective of each one.

2.4.1 Posed Emotions

Emotions acted out based on conjecture or with the guidance from actors or professionals are called **posed expressions** [31]. Most facial emotion databases, especially the early ones i.e. Banse-Scherer, CK and Chen-Huang, consist purely of posed facial expressions, as it is the easiest to gather. However, they also are the least representative of real world authentic emotions as forced emotions are often over-exaggerated or missing subtle details. Due to this, human expression analysis models created through the use of posed databases often have very poor results with real world data [2]. To overcome the problems related to authenticity, professional theatre actors have been employed, e.g. for the GEMEP database.

In 2000, the **Cohn-Kanade (CK)** database was released for the purpose of promoting research into automatically detecting individual facial expressions. This initial release, includes 486 sequences from 97 posers. Each sequence begins with a neutral expression and proceeds to a peak expression. The peak expression for each sequence in fully FACS coded and given an emotion label. The emotion label refers to what expression was requested rather than what may actually have been performed. For the **CK+** distribution, they have augmented the dataset further to include 593 sequences from 123 subjects (an additional 107 sequences and 26 subjects) [19].

The CK [19] only consist of frontal portrait images taken with simple RGB cameras. Newer databases try to design collection methods that incorporate data, which is closer to real life scenarios by using different angles and occlusions (i.e. hats, glasses, etc.). Great examples are the MMI and Multi-PIE databases, which were some of the first well-known ones using multiple view angles. In order to increase

FIGURE 2.4: Examples of the CK+ database. The images on the top level are subsumed from the original CK database and those on the bottom are representative of the extended data.

the accuracy of the human expression analysis models, databases like the FABO have expanded the frame from a portrait to the entire upper body [2].

2.4.2 Emotion Under Speech

In this subsection, it is presented the most important databases that contains recordings with emotions under speech (conversations, phrases,etc), so it can be possible to analyze emotions from video and audio. In case of multimodal data, the audio component can provide a semantic context, which can have a larger bearing on the emotion than the facial expressions themselves. However, in case of solely audio data, like the Bank and Stock Service and ACC databases, the context of the speech plays a quintessential role in emotion recognition [2].

There are mainly two types of emotion databases that contain audio content: stand-alone audio databases and video databases that include spoken words or utterances. The information extracted from audio is called context and can be generally categorized into a multitude, wherein the three important context subdivisions for emotion recognition databases are the semantic, structural, and temporal ones [2]. Following, there is a description of the most important **video databases** of the literature.

- **IEMOCAP - Interactive Emotional Dyadic Motion Capture Database** [6]:

  This kind of databases are usually called **inducted** and the participants usually interact with other individuals or are subject to audiovisual media in order to invoke real emotions. Induced emotion databases have become more common in recent years due to the limitations of posed expressions. The performance of the models in real life is greatly improved, since they are not hindered by overemphasised and fake expressions, making them more natural.

  It was captured by the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC) in 2007. This database was recorded from ten actors in dyadic sessions with markers on the face, head, and hands, which provide detailed information about their facial expression and hand movements during scripted and spontaneous spoken communication scenarios. The actors performed selected emotional scripts and also improvised hypothetical scenarios designed to elicit specific types of emotions (happiness, anger, sadness, frustration and neutral state). The corpus contains approximately twelve hours of data (20 month capturing) and each video has a variable number of utterances, assigning only one emotion to each utterance

found in the conversation. The detailed motion capture information, the interactive setting to elicit authentic emotions, and the size of the database make this corpus a valuable addition to the existing databases in the community for the study and modeling of multimodal and expressive human communication.

For the recording, the subjects were seated all the recording in order to avoid having gestures outside. As well, the subjects had 3 meters of separation between them, just to avoid sound interferences and visual occlusions. They used two microphones (Schoeps CMIT 5U, 48KHz) to record both speakers and the final audio is provided merging both signal in unique audio file. Fig. 2.5 is an example of a full conversation between two actors.

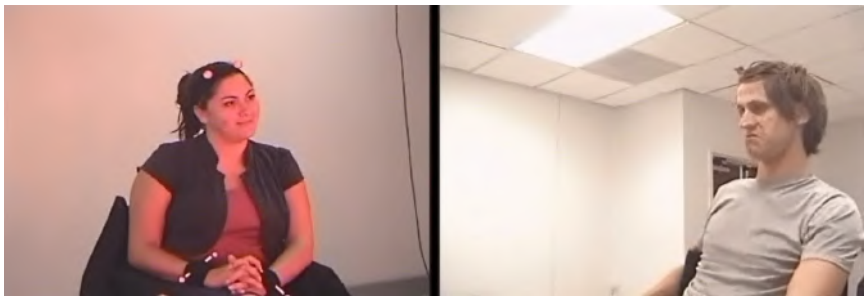

FIGURE 2.5: Example of the IEMOCAP database.

As commented before, there are 5 principal emotion to have into account (happiness, anger, sadness, frustration and neutral state), but this database has three methods to label the emotions and make it more robust to changes:

1. **Subjective evaluations**: Six human evaluators were asked to assess the emotional content of the database in terms of emotional categories. As mentioned before, the database was designed to target anger, sadness, happiness, frustration and neutral state. However, some of the sentences were not adequately described with only these emotion labels. Since the interactions were intended to be as natural as possible, the experimenters expected to observe utterances full of excitement, fear and other broad range of mixed emotions that are commonly seen during natural human interactions.
2. **Self-emotional evaluations**: In addition to the emotional assessments with naïve evaluators, they asked six of the actors who participated in the data collection to self-evaluate the emotional content of their sessions using categorical (i.e., sadness, happiness) and attribute (i.e., activation, valence) approaches.
3. **Continuous emotional descriptors**: An alternative approach to describe the emotional content of an utterance is to use primitive attributes such as valence, activation (or arousal), and dominance. The self-assessment manikins (SAMs) were used to evaluate the corpus in terms of the attributes valence [1-negative, 5-possitive], activation [1-calm, 5-excited], and dominance [1-weak, 5-strong]

IEMOCAP database contains 5498 utterance (normally distributed in 4290 for training and 1208 for testing) from 5 different sessions with 2 actors per session (10 actors in total). It sums up a total number of 151 different videos of variable length, representing different situation and acted or improvised conversations.

This **dyadic database** is the most used for emotion recognition since its creation. For this reason, the experiments of this project are made with this database in order to compare results and have a reference. This database has some similarities with the one that has been recorder in this project (scenario,number of participants,etc), so this is another reason for using this database to perform our experiments.

- **SEMAINE - Sustained Emotionally coloured Machine human Interaction using Nonverbal Expression** [25]:

  SEMAINE is based on a scenario known as the 'Sensitive Artificial Listener', or SAL for short. In 2007, they created a large audiovisual database as part of an iterative approach to building agents that can engage a person in a sustained, emotionally coloured conversation, using the Sensitive Artificial Listener (SAL) paradigm. Data used to build the system came from interactions between users and an 'operator' simulating a SAL agent, in different configurations: Solid SAL (designed so that operators displayed appropriate nonverbal behaviour) and Semiautomatic SAL (designed so that users' experience approximated interacting with a machine). Having built the system, they recorded user interactions with the most communicatively competent version and baseline versions with reduced nonverbal skills. High quality recording is provided by five high-resolution, high framerate cameras, and four microphones, recorded synchronously.

  Recordings have a total of 150 participants, for a total of 959 conversations with individual SAL characters, lasting approximately 5 minutes each. Solid SAL recordings are transcribed and extensively annotated: 6-8 raters per clip traced five affective dimensions and 27 associated categories. Other scenarios are labelled on the same pattern, but less fully. Additional information includes FACS annotation on selected extracts, identification of laughs, nods and shakes, and measures of user engagement with the automatic system.

FIGURE 2.6: Examples of the SEMAINE database.

The rating procedure involved full rating for five dimensions and then optional rating for instances of another 27 dimensions. The five fully rated dimensions are valence, activation, power, anticipation/expectation and intensity. The other traces dealt with more or less categorical descriptions, and were made after the five core dimensions have been annotated. The next dimensions are basic emotions: fear, anger, happiness, sadness, disgust, contempt and amusement. They include different ways of annotating such as epistemic states (certain, agreeing, interested) or others (shows solidarity, shows antagonism, shows tension).

It is clear that this database has got a lot of diversity. They try to measure different emotions, situations and activities, so this is not the best database for

the experiments of this project because the interaction is not made between two subject and they are measuring nonverbal expressions.

- **AFEW Database - Acted Facial Expressions In The Wild Database** [10]:

  This kind of databases are called **spontaneus** and are considered to be the closest to actual real-life scenarios. However, since true emotion can only be observed, when the person is not aware of being recorded [40], they are difficult to collect and label. The acquisition of data is usually in conflict with privacy or ethics, whereas the labelling has to be done manually and the true emotion has to be guessed by the analyser [31].

  In 2009, they present a labelled temporal facial expression database from movies. Human facial expression databases till now have been captured in controlled 'lab' environments. This database constitutes information on clips of both single and multiple subjects interacting with each other. To overcome the tedious and error-prone process of manual data collection, they use a semi-automatic method based on searching the Subtitles for Deaf and Hearing impaired (SDH) as well as Closed Caption (CC). The database covers unconstrained facial expressions, varied head poses and movements, vast age range, occlusions, varied focus, multiple people in the same scene and close-to-real world illumination. The range of age of subjects in the clips is large from 1 - 70 years. The information about the clips has been stored in an extensible XML schema and the subjects in the clips have been annotated with attributes such as Name, Age of Actor, Age of Character, Pose, Gender, Expression of Person and the overall Clip Expression.

  It contains 957 videos labelled with six basic expressions Angry, Happy, Disgust, Fear, Sad, Surprise and the Neutral expression. The subjects in the database exhibit natural (including out-of-plane) head poses and movements, which are largely missing in other current temporal facial expression databases. The subjects in the database exhibit natural (including out-of-plane) head poses and movements, which are largely missing in other current temporal facial expression databases. AFEW is currently the only facial expression database, which has multiple labelled subjects in the same frame. This enables an interesting study on the 'theme' expression of a scene with multiple subjects, which may or may not have the same expression at a given time. As well, The movies have been chosen covering a large set of actors. Many actors have appeared in multiple movies in the dataset, which will enable to research on how their expressions have evolved over the time, whether they differ for different genres, etc.

FIGURE 2.7: Examples of the AFEW database.

This database is widely used for emotion recognition. The difference with the database of this project is that the videos are all acted situations and they involve two or more subjects, so there could be some problems if it is used this database to train the methods proposed here.

2.5 State-of-the-art: Methods

Emotion recognition has attracted attention in various fields such as natural language processing, psychology or cognitive science. Ekman (1993) found correlation between emotion and facial cues. It was an important improvement on the field, but emotion recognition remained a really challenging task. In a dyadic conversation, which is a dialogue between two entities, the audio and visual information can be used separately or fused using different techniques, which are called unimodal or multimodal scenarios, respectively. Following, there is an explanation of some of the state-of-the-art techniques that have been used for the emotion recognition problem in the unimodal and multimodal scenarios. This work does not include the textual information because the aim of it is to train a real-time system. This is the reason why the next subsections are about audio, video and their fusion.

2.5.1 Audio

Important audio features for emotion recognition are pitch, formants, duration, spectral energy, and Mel frequency cepstral coefficients (MFCCs). These features has been studied both at utterance-level and at frame-level. In [13], a total of 106 utterance-level audio features are extracted related to fundamental frequency, energy, duration and spectral envelope. Then, a feature selector is applied to obtain the best 40 features and, after that, feature reduction techniques such as PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis) are applied. The classification experiments are performed with Gaussian classifiers. In [32], a hybrid model is proposed. In it, the MFCC and the LLD (Low-Level Descriptors) features are computed separately. The LLD features are normalized and the MFCC features are passed through a GMM-HMM (Gaussian Mixture Model-Hidden Markov Model) and both outputs are used to train the same SVM classifier.

But these methods are relatively old. Nowadays, most of the studies about emotion recognition using audio, make a fusion with other information such as the visual or the textual ones. But there are some studies that evaluate the IEMOCAP database (the one that is used in this work) using only the audio information. These studies used to use the **OpenSMILE** toolkit to obtain the vector of features for each audio file. In [23], the audio feature extraction process is performed at 30Hz frame rate with 100ms sliding window. OpenSMILE is used to obtain the audio features, but prior this feature extraction, the audio signals are processed with voice intensity thresholding and voice normalization. Finally, each audio segment is converted to a vector of 6392 features. Other works, such as [15][24], use a similar procedure to obtain a feature vector for each audio segments.

But there are other recent studies that use the raw audio as input of a deep learning network to obtain a feature vector or directly as a classification method. This is the case of [3], where the author uses the raw audios as the input of a CNN. The CNN is designed as a simple layer with a convolution window of size 200 and an overlapping step of 50. This convolutional layer is the responsible of the feature extraction. After that, there is a max-pooling layer followed by a fully-connected layer and the final softmax layer.

2.5.2 Video

The visual-based techniques in the emotion recognition literature are mainly based on facial expressions, since face plays a vital role in conveying emotions. In [18], the author describes the problem of emotion recognition in three steps:

1. The first one consists of locating the faces in the scene (*face detection*)
2. The second one aims the extraction of the facial features from the detected faces (*facial feature extraction*)
3. in the last one an analysis of the changes in these features is performed followed by a classification of this information into some facial-expression-interpretative categories.

This work also refers to the importance of making clearly the differences between acted and naturalistic behaviours.

The **face detection** process is simple: first, the video is divided into frames and then a **face detector** is applied over every frame to obtain all the faces in the video. There are different face detectors, but one of the most used in the state-of-the-art is the OpenFace toolkit. This is a state-of-the art tool intended for facial landmark detection, head pose estimation, facial action unit recognition and eye-gaze estimation.

Once the faces are obtained and cropped, the extraction of facial features and the classification processes are performed. Facial features can be subdivided into two broad categories: **geometric** and **appearance**. The geometric features are the shapes of the facial components (eyes, mouth, etc.) and the locations of facial fiducial points (corners of the eyes, mouth, etc.), while appearance features represent the texture of the facial skin in specific facial areas including wrinkles, bulges, and furrows. The contraction of the facial muscles produce facial expressions, induce movements of the facial skin and changes in the location and/or appearance of facial features. Such changes can be detected analyzing the optical flow, the changes in the action units of two consecutive faces, etc.

The most useful strategy for extracting audio or image features is using **Convolutional Neural Networks (CNNs)** (see Sec. 2.5.2). CNNs have two important variants for this scenario, 3D-CNNs or 2D-CNNs, depending on if the movement or temporal information across frames is wanted to be included into the model or not. In most of the recent studies about this problem, a **3D-CNN** is used on the video to focus not only on the feature extraction from each video but also on the temporal features across frames. Publications like [23][15][24] uses a similar scheme: they use a 3D-CNN to extract spatio-temporal features across frames. The working of a 3D-CNN is identical to its 2D counterpart with an input being a video v of dimension $(3, f, h, w)$. Here, 3 represents the RGB channels and f, h, w are the number of frames, height and width of each frame, respectively. For the convolution operation, a 3D filter f_l of dimension $(f_m, 3, f_d, f_h, f_w)$ is used where, $f_{[m/d/h/f]}$ represents the number of feature maps, depth, height and width of the filter, respectively. Max-pooling is applied to the output of this convolution across a 3D sliding window of dimension (m_p, m_p, m_p).

Convolutinal Neural Network

A Convolutional Neural Network, also known as **CNN** or **ConvNet**, is a class of neural networks that specializes in processing data that has a grid-like topology,

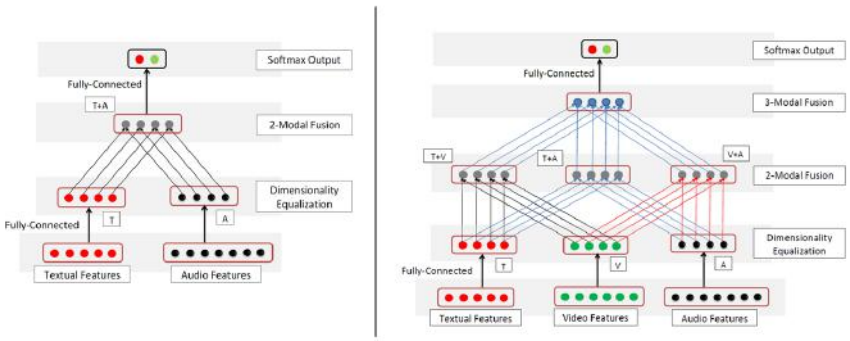

FIGURE 2.8: On the left: bimodal fusion. On the right: trimodal fusion. Source: [23]

such as an image. Each neuron in a CNN processes data only in its receptive field as it does the human brain. The layers on a CNN are arranged in such a way so that they detect simpler patterns first (lines, curves, etc.) and more complex patterns (faces, objects, etc.) further along. A CNN typically has three layers: a convolutional layer, pooling layer and fully connected layer.

1. **Convolutional Neural Layer**. This layer performs a dot product between two matrices, the kernel and the restricted portion of the receptive field. The kernel slides across the height and width of the image producing the image representation of that receptive region. This produces a two-dimensional representation of the image known as an activation map. The values of the kernel are the parameters learned during the training process.

2. **Pooling Layer**. The pooling layer replaces the output of the network at certain locations by deriving a summary statistic of the nearby outputs (normally taking the maximum value of a neighborhood). This helps in reducing the spatial size of the representation, which decreases the required amount of computation and weights.

3. **Fully Connected Layer**. Neurons in this layer have full connectivity with all neurons in the preceding and succeeding layer. This is why it can be computed by a matrix multiplication followed by a bias effect. The FC layer helps map the representation between the input and the output.

2.5.3 Multimodal Fusion

Multimodal fusion is the process of combining data collected from various modalities for analysis tasks. It has gained increasing attention from researchers in diverse fields due to its potential for lots of applications (i.e. sentiment analysis, emotion recognition, human tracking, image segmentation, video classification, etc.). The fusion of multimodal data can increase the accuracy in many scenarios. There are mainly two levels or types of fusion studied by researchers: feature-level fusion or early fusion, and decision-level fusion or late fusion [29].

1. **Feature-level** or **early fusion** fuses the features extracted from various modalities such as visual features, text features and audio features, to a general feature vector that is sent for analysis. The advantage of feature level fusion is that the correlation between various multimodal features at an early stage can potentially provide better task accomplishment. The disadvantage of this fusion process is time synchronization.

2. **Decision-level** or **late fusion**. In this fusion process, the features of each modality are examined and classified independently and the results are fused as a decision vector to obtain the final decision. The advantage of decision-level fusion is that the fusion of decisions obtained from various modalities becomes easy compared to feature-level fusion, since the decisions resulting from multiple modalities usually have the same form of data. Another advantage of this fusion process is that every modality can utilize its best suitable classifier or model to learn its features. But this is also a disadvantage because the learning process of all these classifiers becomes time consuming.

3. **Hybrid multimodal fusion**. This type of fusion is the combination of both feature-level and decision-level fusion methods in order to try to exploit the advantages of both strategies and overcome the disadvantages.

In the emotion recognition field, most of the researchers prefer to use the **feature-level** or **early fusion strategy**. In many works, the authors give more importance to the process of extracting the contextual information between utterances and speakers or to the feature extraction methods. This is the reason why a simple **concatenation** of features from different modalities is used as the fusion method.

But there are exceptions. In [14], the authors aims to classify emotions using audio, visual and textual information by **attaching probabilities** to each category based on automatically generated trees, with SVMs acting as nodes. In [23], the authors defend that the concatenation or early fusion has the problem of cannot be able of filtering out conflicting or redundant information obtained from different modalities. To compute their trimodal fusion (audio, video and textual features), they propose a **hierarchical approach** which proceeds from unimodal to bimodal vectors and then from bimodal to trimodal vectors. Fig. 2.8 shows how the fusion is done: first a fully connected layer is applied to equalize the dimensionality of the feature vector of the three modalities, then a bimodal fusion is applied by pairs and finally a fully connected layer with softmax as activation function performs the classification. The *bimodal fusion* consists of this: imagine this is the matrix of values obtained after equalizing the dimensionality

$$g_x = \begin{bmatrix} c^x_{11} & c^x_{21} & c^x_{31} & \dots & c^x_{D1} \\ c^x_{12} & c^x_{22} & c^x_{32} & \dots & c^x_{D2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c^x_{1N} & c^x_{2N} & c^x_{3N} & \dots & c^x_{DN} \end{bmatrix}$$

, where D is the dimension, N is the maximum number of utterances in a video and x is A, V or T depending on if it is audio, video or text. The bimodal combination is achieved using a fully-connected layer as it can be seen in the following expressions

$$i^{VA}_{lt} = tanh(w^{VA}_l \cdot [c^V_{lt}, c^A_{lt}]^T + b^{VA}_l) \tag{2.1}$$

$$i_{lt}^{AT} = tanh(w_l^{AT} \cdot [c_{lt}^A, c_{lt}^T]^T + b_l^{AT}) \tag{2.2}$$

$$i_{lt}^{VT} = tanh(w_l^{VT} \cdot [c_{lt}^V, c_{lt}^T]^T + b_l^{VT}) \tag{2.3}$$

, where the values of the previous matrices are concatenated by pairs. Learning these weights helps to achieve a better fusion of modalities. The *trimodal fusion* is similar to the bimodal one. First, they apply the dimensionality equalization to the feature vectors of the three modalities and then the bimodal fusion to obtain three matrices. After that, they apply a fully-connected layer to obtain the trimodal fusion. The only difference is that instead of concatenating two values inside the *tanh* function of the fully-connected layers, now they have the concatenation of three bimodal values.

2.5.4 Baseline

For this project, [8] is going to be used as baseline. In this work, the authors use the same database (i.e. IEMOCAP) and analyze the emotion recognition problem using the video, the audio and the text. Since this project does not use the textual information, the comparisons with the baseline are going to take into account the audio and video modalities and their fusion. One of the reasons of using [8] as baseline is because they propose also an emotion recognition system at frame level. The method they implement consists of a CNN for the textual features, OpenSMILE for the audio segments of 100ms and a CNN with only two convolutional layers for the faces, which is much simpler than the one used here (i.e. VGG). For the faces, they only take every tenth frame to reduce the amount of data for the training and then they fuse every t and $t + 1$ face into a single image. They use a simple concatenation for the fusion, and then a SVM for the final decision.

Chapter 3

Speaker Segmentation and Clustering

Nowadays, a rapid increase in the volume of recorded speech is manifested. Indeed, television and audio broadcasting, meeting recordings, and voice messages have become a commonplace [36]. However, the huge volume size hinders content organization, navigation, browsing, and retrieval. *Speaker segmentation* and *speaker clustering* are tools that alleviate the management of huge audio archives.

Is important to realize the time that is consumed when trying to segment speakers in a long duration audio file. Manually segmentation has always been a problem for researchers and is something that is starting to change in recent days. For example, if there is a meeting recording audio with two or more people talking or a supervised system has to be trained with all labels correctly assigned, there will be a lot of time wasting in the annotating part of the process. For those reason, is important to develop an automatic system that splits an audio in different speakers. There exist new techniques that use audio and visual features to develop this systems, but this project is going to focus only on audio features because is the most simple one to develop.

Definition 2 ***Speaker segmentation*** *aims at splitting an audio stream into acoustically homogeneous segments, so as every segment ideally contains only one speaker [35].*

Definition 3 ***Speaker clustering*** *refers to unsupervised classification of speech segments based on speaker voice characteristics [38]. That is, to identify all speech segments uttered by the same speaker in an audio recording and assign a unique label to them*

As it is defined in definition 2 and 3, there are some differences between speaker segmentation and speaker clustering. Speaker segmentation and clustering consists of identifying who is speaking and when, in a long meeting conversation. Ideally, a speaker segmentation and clustering system will discover how many people are involved in the meeting, and output clusters corresponding to each speaker. First of all, speaker segmentation divides the audio in utterance and detects where a speaker change has happened. Then, speaker clustering outputs clusters of this segments assigning one segment to one or more speakers.

3.1 Speaker Segmentation Methodology

3.1.1 Methodology

Speaker segmentation followed by speaker clustering is called **diarization**. Diarization is the process of automatically splitting the audio recording into speaker segments and determining which segments are uttered by the same speaker. Nowadays, this methods *cannot split audios with overlapping between speakers* and the researchers are working on it. The main steps that are followed when developing a diarization system are commented here:

1. **VAD (Voice Activity Detection**: Voice activity detection usually addresses a binary decision on the presence of speech for each frame of the noisy. The most simple VAD schemes are based on a energy detector. If the energy of the signal rises a threshold amount above the noise floor, then the increase in energy is assumed to be to associated with voice (figure 3.1). In waveform and spectral analysis, voice activity detection makes use of the known characteristics of the speech. Applying VAD in this method is more computational intensive than energy based solutions, but are better able to detect noise in non-stationary noise and low SNR scenarios. For example, voiced speech contains a strong fundamental frequency with it's harmonics. Thus, the analysis of cepstrum of a signal can reveal source of the signal energy.

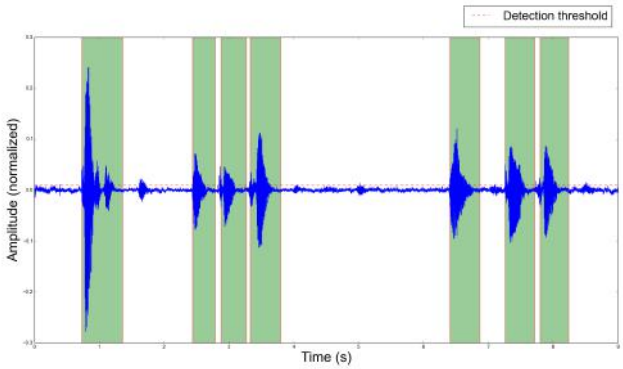

FIGURE 3.1: VAD example: energy detector.

2. **Feature Extraction**: Once non-speech segments are removed, the speech is divided in overlaped small windows to extract easily the features. The most used parametrization features in speaker diarization are Mel Frequency Cepstral Coefficients (MFCC), Linear frequency cepstral coefficients (LFCC), Perceptual Linear Predictors (PLP), Linear Predictive Coding (LPC) and others. It is common as well to use some Neural Networks to learn d-vectors to better learn the representation of each window. The disadvantage of this method is that it has to be supervised

3. **Classification**: After the feature extraction, is common to use clustering technique (unsupervised method) such as K-means or Hierarchical clustering when we face an unsupervised problem. When we have a supervised problem is common to use Neural Networks or GMM (Gaussian Mixture Models) to learn the characteristics.

3.1.2 Error Metrics

The main metric for diarization performance measurement is the **Diarization Error Rate (DER)**. The metric is computed in two steps [30]: the first step is to establish a mapping between the speaker tags provided by the system and the speaker identities found in the reference. The second step then computes the error rate using that mapping Computing an error rate requires defining what the errors can be. Three error types are defined in the diarization context [12]:

- The *confusion error*, when the system-provided speaker tag and the reference do not match through the mapping
- The *miss error*, when speech is present in the reference but no speaker is present in the hypothesis
- The *false alarm error* when speech has incorrectly been detected by the system.

That gives us the final DER definition as

$$DER = \frac{confusion + miss + false\ alarm}{total\ reference\ speech\ time}$$

A last point in building that metric is taking into account the intrinsic human imprecision. It is very hard to define a precise point in time when speech starts or stops, especially when background noise or overlapping speech is present.

There are more metrics such us **utterance accuracy**, where the correct utterance predicted is divided (without taking into account the exactly time window) between the total number of utterances of that speaker. This is a good metric if for not taking into account the human error when annotating the utterances.

3.2 State of the art - Speaker Segmentation

In this section some of the best methods for speaker segmentation/diarization that used only the audio source information are going to be commented.

In [1], for the feature extraction they divide the audio into 30 ms frames with a hop of 10 ms and to remove the redundancy they take collection of periodogram bins and sum them using Mel filterbank. Finally they take DCT of the log filterbank energies. Then, they apply a VAD extracting the MFCC features and for the speaker segmentation they apply Hypothesis Testing (Null and alternate) or Growing Window. For the Growing Window a small window is taken. If the feature vectors at the endpoints of window are better modeled by separate distributions, the midpoint is declared as speaker change point. In this case, the search is again started from the next segment. Otherwise the window is slightly increased and once again the above conditions are checked. For the speaker clustering they use a pretrained model of GMM from Universal Background model (UBM) because the duration of their audios is very short to train a new GMM.

In [37], they propose to explore the use of joint representation learning and similarity metric learning with triplet loss in speaker diarization, while entirely dispensing the need for i-vector extraction. For the first time, they leverage attention networks to model the temporal characteristics of speech segments.

The first paper that used RNNs (Recurrent Neural Networks) for speaker segmentation was [9]. They applied a recurrent neural network directly on the magnitude spectrograms (SFTF) to learn a set of high-level feature representations also referred to as speaker embeddings. This network consists of 11 learnable layers: 4 convolutional blocks followed by 2 recurrent layers with 1 fullyconnected at the end.

In a recent paper [41] from Google, they first extract embeddings from sliding windows of size 240ms and 50% overlap. Then, A simple voice activity detector (VAD) with only two full-covariance Gaussians is used to remove non-speech parts, and partition the utterance into nonoverlapping segments with max length of 400ms. Then they average window-level embeddings to segment-level d-vectors, and feed them into the clustering algorithm to produce final diarization results. The text-independent speaker recognition network for computing embeddings has three LSTM layers and one linear layer. The baseline architecture is shown in figure 3.2. In this paper, the important part is that they presented a speaker diarization system where the commonly used clustering module is replaced by a trainable unbounded interleaved-state RNN. Since all components of this system can be learned in a supervised manner, it is preferred over unsupervised systems in scenarios where training data with high quality time-stamped speaker labels are available. In this project, avoiding this kind of problems is important, because labelling is time consuming.

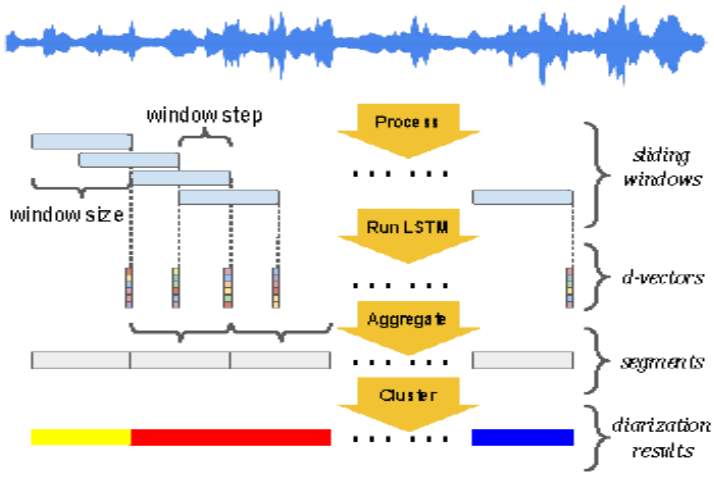

FIGURE 3.2: Google baseline system architecture.

After Google publication [41], [18] propose a new similar method. They apply the first steps (VAD and d-vectors) in the same way and then a similarity measurement algorithm computes the score between every embedding vector pair and form a square similarity matrix S. Finally, they perform clustering among all segments based on S. They use **Spectral Clustering** and it will be very important in our project and we will explain later. The main change with respect [41] is that they propose to predict each row of the similarity matrix S with a stacked Bi-LSTM using the binary cross entropy (BCE) loss function. Since [41] achieves a DER of 7.6%, with the same data [18] achieves a DER of 6.63%.

Chapter 4

Face-to-face Dyadic Interaction Dataset

4.1 Introduction

Technology providers must endow **socially-intelligent** machines with the capacity of understanding and adapting to different social contexts and individuals, so as to provide a more empathetic, inclusive, tailored communication. To do so, it is necessary to train such systems with computational models that capture the richness and complexity of natural human-human interactions. The DECODE project aims at **analyzing human communication, from a multidisciplinary perspective**, to research and implement new paradigms and technologies of **interpersonal behavior understanding**, by means of a novel annotated database of dyadic face-to-face spontaneous interactions. The database, which would be publicly available for the research community in compliance with GDPR, will consist of audio-visual recordings, personality profiling and sociodemographic data of diverse user populations. The project will entail the development of new automatic models of interpersonal influence understanding and personality traits regression from interaction social signals using novel deep learning techniques, with the ultimate goal of demonstrating the feasibility of developing socially-aware systems able to decode the real personality and characteristics of users from verbal and non-verbal social signals, and adapt to them accordingly.

4.1.1 Project

In this new era of ubiquitous intelligent systems becoming more and more seamlessly integrated into our daily life, the future looks daunting for part of the society. "*How should we adapt to this new technology? How can we, as users, learn how to interact with machines?*". Evidences of racially and gender biased artificial intelligence (AI) have also contributed to a skeptical vision of such technologies. While valid, such questions and fears must be tackled from a different perspective, that is, "*how can technology providers train these systems to interact in a more humane way, while accounting for possible society biases?*" The so-called task of humanizing AI deals precisely with these subjects, in order to produce intelligent systems capable of successfully interpreting and reacting to human factors. According to the psychology literature, the way some social signals are perceived, such as eye gaze and facial expressions, is affected by our personality (Ponari et al., 2013), health status (Surguladze et al., 2004) and cultural identities (Riviello and Esposito, 2016). Therefore, **socially-intelligent systems have to be capable of accurately perceiving and inferring the personality**

and other particularities of different individuals, so as to provide a more effective, empathetic, and natural tailored communication.

To embody this human likeness into such systems, it is imperative to have a deeper understanding of real human-human interactions first, to computationally model both individual behavior and interpersonal influence. Current literature in computer vision and machine learning for human behavior understanding has mainly focused on research and development of perception, analysis and synthesis methods for individual behavior; however, interpersonal-based tasks such as perception and modelling of the communication flow and the adaptation between communication partners have been largely unexplored from a technology point of view (Vinciarelli et al., 2015). To advance in such areas, the community is in need of **publicly-available annotated datasets of non-acted, spontaneous interactions** among dyads and small groups belonging to different population groups in terms of age, gender, and cultural background. While several acted datasets exist (Busso et al., 2008), natural interactions are preferred, as they cover the richness and complexity of social communications in real life.

As a result, the main goal of DECODE is to **analyze human-human communication, from a multidisciplinary perspective (i.e. sociological, psychological and technological)**, to research and implement new paradigms and technologies of interpersonal behavior understanding by means of a common database of dyadic face-to-face interactions. The purpose is to move beyond automatic individual behavior detection and focus on the development of automatic approaches to study and understand the mechanisms of perception of and adaptation to verbal and nonverbal social signals in dyadic interactions, taking into account individual and dyad characteristics. Our central research question revolves around the **feasibility of developing socially-aware systems able to decode the real personality and internal process of an individual by the social signals they convey, as well as understand how such interaction partners perceive and react to those cues directed to them**. To answer our hypothesis, the project will address the following sub-tasks:

1. To design and collect **highly-varied audio-visual, physiology, sociodemographic and personality profiling data from target population, comprising: audio-visual 360º recordings of face-to-face dyadic natural interactions** from third and first-person view cameras and ambient and lapel microphones; heart rate data from wearable monitors; real and apparent personality, temper and current mood profiling through self- and hetero-evaluation questionnaires (filled in by both interaction partners); sociodemographic individual metadata consisting in gender, age, ethnicity, country of origin, country of residence, occupation and education; and dyadic metadata such as relationship among participants (i.e., friends, family, unknown).

2. To manually **annotate the dataset** of audio-visual recordings, ranging from low-level behaviors such as facial expressions, eye gaze, gestures, body poses and utterances, to high-level behaviors such as attention, interest, engagement and overall quality of interaction. Transcription of dialogues would also be included. Further individual and dyadic ground truth would be extracted from the analysis of the profiling data, such as dominance, leadership, agency, communion, and high-order universal phenotypic personality traits (openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism),

using standardized questionnaires such as BFI-2 (Soto and John, 2017), HEXACO (Lee et al., 2004), CBQ-short (Putnam and Rothbart, 2006) and EATQ-R (Ellis and Rothbarth, 2001).

3. To explore relationships between ground truth annotations of social signals with behavioral constructs and sociodemographic and profiling data, using modern and flexible statistical multivariate procedures that allow testing complex relations between variables at a time.

4. To research and develop automatic **methods of interpersonal influence understanding and personality traits regression from interaction social signals using novel multimodal deep learning techniques**.

The database is planned to include a sample of 150 Spain-based individuals to better capture social interaction nuances of our society, who would participate in over 200 dyadic interactions. It would consist of several tasks that elicit particular social behaviors, such as joint attention, competition, collaboration, and free conversation, with different levels of cognitive load. Currently, **no annotated dataset of such dimensions and characteristics is publicly available for research purposes**. By releasing the annotated database to the research community, **it is expected a great data sharing and collaboration among different disciplines, reuse, and repurposing of new research questions, as well as foster international visibility and exposure of Spain-based research**. Data collection and sharing would be guaranteed by ethical committee approval and consent of participants in agreement with GDPR.

In a world that becomes more and more automated, the knowledge and novel techniques that this project would produce would enable **more representative and accurate interaction models**, which in turn would provide us with more empathic agents, tailored to the user needs, social context and characteristics (i.e. sociodemographic group, personality traits, etc.). The project would also serve as a proof of concept for more fair, socially-inclusive, explainable and interpretable automatic models, getting away from the "*black-box*" preconception of AI-based systems, and to automatically detect and **account for possible sources of bias in our society**. The applications are countless, ranging from virtual tutoring and therapy systems to assisting care for the elderly and people with disabilities (e.g. a personalized virtual assistant with compatible features with the elderly), as well as support in job interviews, among others, ultimately **improving our quality of life**.

4.2 Summary

The Face-to-face Dyadic Interaction Dataset contains a total of **194 interaction sessions by 150 different participants**. The recording procedure of each session consists of 5 task done by pairs. This tasks are completely different from each other with different behavior elicitation conditions and cognitive workload.

The duration of each session, on average, is 25 minutes. This means that the entire dataset is made up by 81 hours of interaction audiovisual content. Recruitment and organization of sessions followed a strict criteria. The recording session were programmed with respect to the participants availability, age, gender and relationship among participants, trying to record each participant more than one time and with known and unknown people. The maximum number of sessions per participant is 5 and the minimum allowed age is 4 years old. Fig. 4.1 shows an histogram

with the number of sessions done by participant, which indicates that the mean number of sessions is **2.59**.

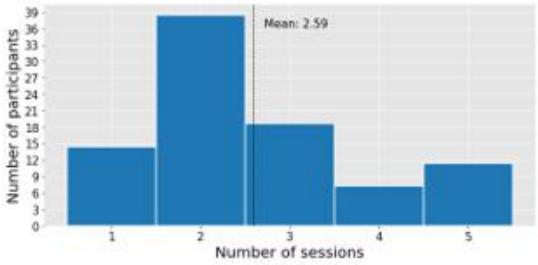

FIGURE 4.1: Histogram of number of sessions done by participant

4.3 Setup

In order to record all the sessions, we use a synchronized setup composed for the following elements:

- 6 HD 720p cameras:
  - 3 AXIS M1124 IP
  - 2 Revotech i712
  - 1 Revotech i706
- 2 HD egocentric cameras Victure AC800
- 2 wrist heart rate monitors FitBit Charge 3
- 2 label microphones Rode Smartlav+
- 1 ambient microphone Olympus ME-33

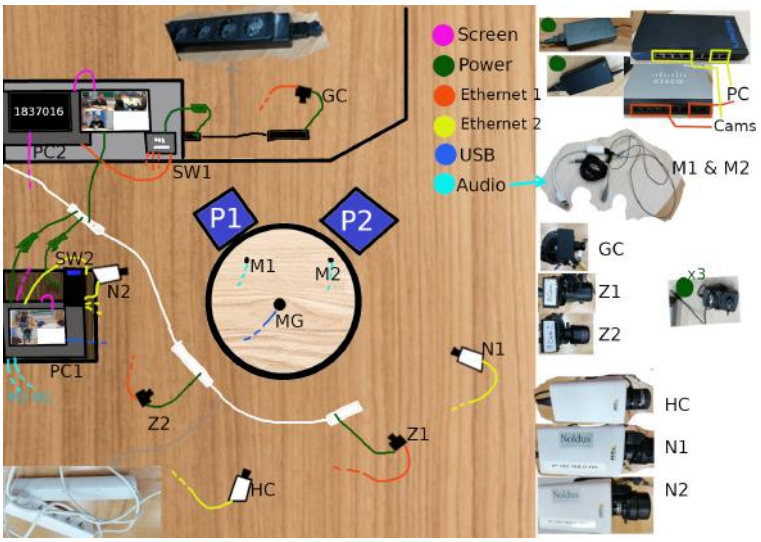

FIGURE 4.2: Picture of the setting of the project. The 6 HD cameras are named N1, N2, Z1, Z2, HC, GC. P1 and P2 are the subjects of the experiment, and M_i are the microphones

The 6 HD cameras are synchronized between them, three by three. We can see a picture of the setting in Figure 4.2. The cameras N1,Z1 and HC are synchronized between them, same for N2,Z2 and GC. With a clock visible for N1 and Z1, we synchronize the rest of them. The wrist heart monitors and egocentric cameras do not appear in the image, they are worn by the subjects: the wrist heart rate monitor is obviously in the wrist, and the egocentric cameras are hanging on the neck.

All cameras except GC are frontal. N1, N2, Z1, Z2 are focusing on of the others subjects, while HC and ZC are general, as we see in Figure 4.3. It also appear the clock in the views Z1 and N1 (two images in the top left), which we use to synchronize these two views and consequently the rest.

FIGURE 4.3: Comparison between the cameras. Frontal 1 correspond to Z1 and Z2 cameras; Frontal 2 to Z1 and Z2. General to HC and GC.

4.4 Tasks

The five tasks done by the subjects in the experiments are always the same, but the order of them is changing. It is another variable of the dataset. These tasks realized are the following ones:

- **Free conversation task** (around 5 mins.)
  - Talk about any subject (e.g how was your day? hobbies, tv series, etc.), avoiding private information.
  - Allows analysis of common conversation constructs, such as turn taking, synchrony, empathy and quality of interaction, among others.

- **'Who am I?' game** (around 7 mins.)
  - Participants use 10 YES or NO questions each to guess the identity of the animal they have on the forehead.
  - 3 difficulty levels (easy - e.g. penguin, medium - e.g. leopard, hard - e.g. albatross).
  - Analyze cognitive processes (e.g. thinking, gaze events).

- **Lego building** (around 5 mins.)
  - Participants build a lego together following a set of instructions.
  - 4 difficulty levels (super easy - for children, easy, medium, hard).
  - Fosters collaboration, cooperation and joint attention.
  - Elicits leader-follower behaviors and subject-object interaction.

- **Ghost blitz game** (around 5 mins.)
  - Participants have to select, among a set of 5 figurines, the figure whose color and shape is not shown in a selected card from a deck of cards.
  - Fosters competitive behavior and interaction with objects.
  - Allows to analyze cognitive processing speed and reflexes.

- **Eye gaze check** (around 1.5 mins.)
  - Participants follow instructions to look *'Elsewhere'*, *'at others face'* or *'at static/moving object'* while moving head and eyes.
  - Elicits different gaze behaviors, such as smooth pursuit.
  - Useful for automatic gaze estimation methods.

4.5 Data and details of participants

Together with the annotations that will be done in the dataset, there is other data which the participants give through some questionnaires. The first one is about personality and temperament. It is self-evaluated for every participant before the first session they do (real data) and hetereo-evaluated after each session between participants (apparent data). This questionnaire differ depending on the age of the participant:

- From 4 to 8 years old: Temperament. **CBQ-short** (Putnam and Rothbarth, 2006 [**Putnam**]).
- From 9 to 15 years old. Temperament and self-regulation. **EATQ-R** (Ellis and Rothbarth, 2001 [**Ellis**]).
- For 16+ years old: Personality. **Big-Five BFI-2** (Soto and John, 2017 [**Soto**]), **Honesty-Humility HEXACO** (Lee et al., 2004 [**Lee**]).

The other questionnaire to be answered is about current mood: **PEQPN** (Williams et al. 2000). This is self-evaluated before and after session, and hetero-evaluated after session.

Sociodemographic metadata of the participants such as age, gender, ethnicity, country of origin, country of residence, occupation, maximum level of studies and relationship among participants was already know before planning the sessions. The relation between them is done in order to have the maximum variability in the dataset. We see in Figure 4.4 the age of the participants of the study. It ranges from 4 to 84 years old, with mean of 31.47 ± 14.6 and 55% are male. Most part of people is from Young age, since a big amount of them were friends of the researchers of this project.

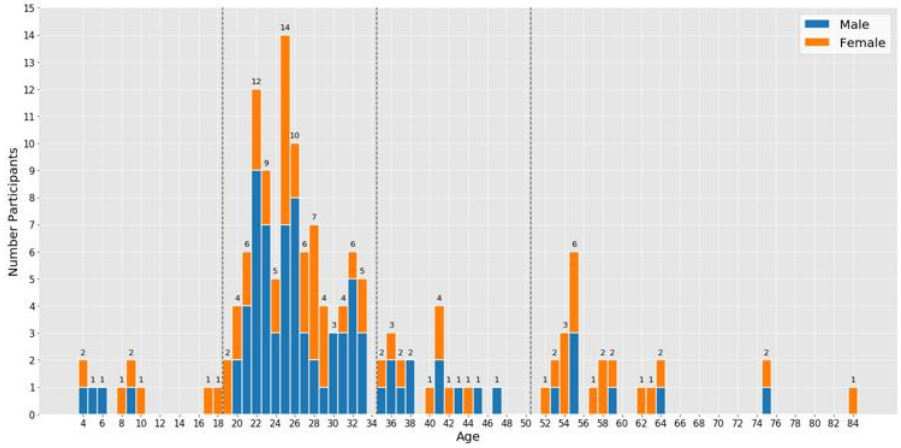

FIGURE 4.4: Histogram of number of participants wrt. age and gender.

In Figure 4.5 it is shown the distribution of the sessions wrt. age, gender and known/unknown people. Again, most of the interactions are done between young people. From the 194 sessions, 44% of them were done between known people (e.g. parent-child, friends, work/study colleagues, couples, etc).

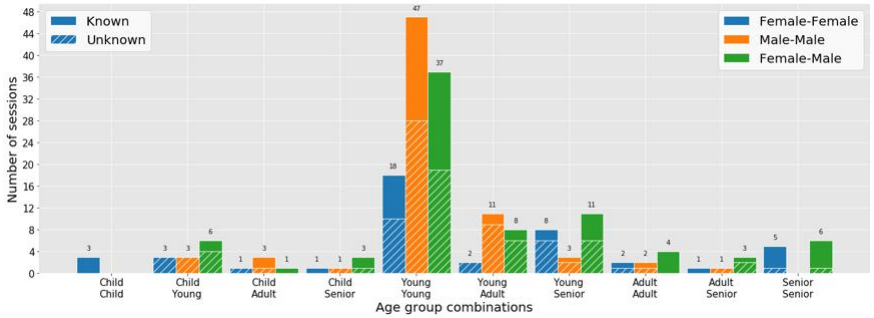

FIGURE 4.5: Distribution of interaction sessions wrt. age, gender and interaction among known/unknown people

The country of origin and sessions done between people from different countries is shown in Figure 4.6. From the 150 subjects, 69% had Spain as country of origin. It was also a considerable number of people from Venezuela (12) and Chile (5). From the 194 sessions, 50% of them were done with both participants from Spain as country of origin.

FIGURE 4.6: Country of origin of the participants and sessions done between them.

4.6 Annotations

As for any dataset, the more labels available for it, the more rich it will be. Some labels can be obtained using reliable automated techniques, for example for tasks such as body and hand pose estimation. However, for other labels a ground-truth is required, specially if the objective is to make this dataset a reference in dyadic human interaction. In particular, manual human annotation will be required for labelling emotion states and utterances. The protocol for these annotations has been defined and some pilots have been tested

4.6.1 Emotion Annotation

Emotions will be rated for short video segments (i.e. 3 second intervals), taking into account the whole scene, using all available modalities, as well as the history of the interaction. That is, information of the affective state of a participant can come from speech, language content, gestures, facial expressions, objects of the scene, or even the reaction of the other participant. Annotators will rate three dimensions of emotion, characterized by the Valence-Arousal-Dominance (VAD) model. Furthermore, the annotators will provide a verbal description of what parts of the image are the most important ones to perceive the felt emotion, as well as a verbal description of the emotions they are perceiving.

Videos will be annotated using the Valence-Arousal-Dominance model (VAD, or PAD, from Pleasure) (Russell and Mehrabian, 1977), which consists of a continuous 3-dimensional model of affective space characterized by three dimensions: valence (pleasant-unpleasant), arousal (active-calm), and dominance (in control-submissive). This dimensional model allows to characterize emotions on three dimensions, each of which spans an interval of real-valued numbers indicating the strength and orientation of each dimension. Dimensional approaches as VAD allows us to represent emotions in a more fine-grained way, in contrast to categorical approaches as Ekman's basic emotions (Anger, Disgust, Fear, Happiness, Sadness and Surprise). Ekman's can be mapped to the VAD model as in Figure 4.7. One can observe that

Ekman's categories are unevenly distributed in the VAD space (Buechel and Hahn, 2016).

Each dimension of the VAD model is characterized as follows:

- **Valence**: measures how pleasant or unpleasant one feels about something. It represents the positive or negative dimension of an emotion. For instance, joy and happiness are pleasant emotions, while fear and anger score in the unpleasant side.
- **Arousal**: measures how excited or apathetic one feels about something. It represents the degree or the strength of the emotion, ranging from calm, bored and sleepy to aroused and excited. It may also denote the mental activity, ranging from low engagement to ecstasy). Note that it is not the intensity of the emotion, as grief and depression can be low arousal intense feelings. For instance, while both anger and rage are unpleasant emotions, rage has a higher arousal state. However, boredom, which is also an unpleasant state, has a low arousal value.
- **Dominance**: it measures the extent to which the emotion makes the subject feel in control of the situation, ranging from being in control of one's emotions (in control, empowered, confident) to dominated by them (submissive, oppressed). Note that having a high level of dominance does not mean one is the dominant person of the interaction, as it is not an interpersonal attribute, but an individual one influenced by the situation and context. The authors of the model claimed that this is a necessary dimension to describe emotion as only dominance makes it possible to distinguish "angry" from "anxious," "alert" from "surprised," "relaxed" from "protected," and "disdainful" from "impotent". The first adjective of these pairs denotes a dominant emotion, while the latter denotes a feeling of being controlled by one's emotions.

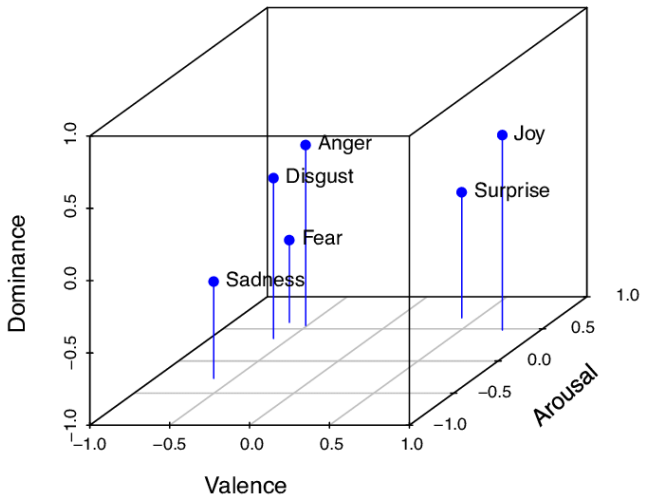

FIGURE 4.7: Mapping from VAD to Ekman's emotions

To assess the three dimensions of valence, arousal, and dominance, the Self-Assessment Manikin (SAM), an affective rating system devised by Lang (1980) will be used. In this system, a graphic figure depicting values along each of the 3 dimensions on a continuously varying scale is used to indicate emotional reactions. As can

be seen in Figure 4.8, SAM ranges from a smiling, happy figure to a frowning, unhappy figure when representing the valence dimension. For the arousal dimension, SAM ranges from an excited, wide-eyed figure to a relaxed, sleepy figure. For the dominance dimension, SAM ranges from a large figure (in control) to a small figure (dominated). Annotators can select any of the 9 figures comprising each scale, which results in a 9-point rating scale for each dimension. Ratings are scored such that 9 represents a high rating on each dimension (i.e., high pleasure, high arousal, high dominance), and 1 represents a low rating on each dimension (i.e., low pleasure, low arousal, low dominance).

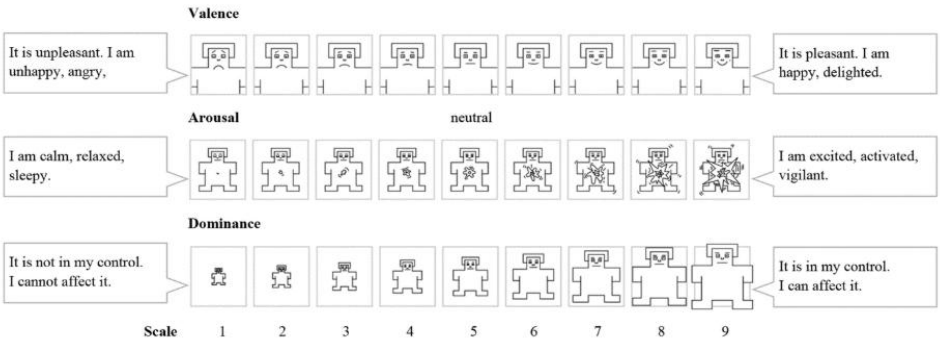

FIGURE 4.8: Self-Assessment Manikin visualization

4.6.2 Utterances

There is no common consensus regarding what an "utterance" is, and each research community uses it in a different way. There are many proposals, like the one from Traum and Heeman, 1997. In this work, instructions and conventions from [17] are followed to annotate utterances.

Determining where the boundaries fall between utterances is an important and tricky part when annotating a conversation. Lots of segments of speech qualify as utterances: a word, a short phrase, or a complex sentence with many embedded clauses. Unfortunately, people do not speak with periods, commas, and question marks to let you know when one utterance ends and another one begins.

Definition 4 *An **utterance** can be a word, a phrase, or an entire sentence. It is the smallest unit of speech.*

The following are all potential examples of utterances: (*ok?*), (*uhuhh*), (*the pink one*), (*yeah, well, I thought she was going to, but she never did.*), (*bugs lives outside, honey.*). For this work, non-verbal vocalizations are also labeled (e.g., laughing, sobbing, crying, mouth clicking, sighing) as utterances, as they have a communicative meaning. However, other non-communicative sounds, like sneezes and coughs, are not considered utterances. In the following lines, some rules that have been followed through the annotations are presented:

- **Pauses of 2 seconds or less**: If the speaker hesitates to find a word, it is treated as one utterance, unless the speaker pauses for more than two seconds. (e.g., *put it in the – toy box*).
- **Self-corrected speech**: If the speaker interrupts herself to correct herself, it is treated as one utterance. (e.g., *don't do that, Nathan – Jake!*).

- **Self-interrupted speech**: If the speaker interrupts herself to express an entirely different thought, it is treated as two separate utterances. (e.g., *why don't you put it – don't do that*).

- **False starts**: If the speaker has a false start in her speech, it is treated as one utterance. (e.g., *will you – will you go to your room?*).

- **Speaker stumbling over words**: If the speaker stumbles over his or her words while trying to formulate an utterance, still count this as a false start and transcribe the attempts on a single utterance line. To count as a false start, the words must be spoken very quickly or be otherwise clearly an attempt at verbalizing a single thought. (e.g., *no – don't – oh no!*).

- **Utterances never span more than one conversational turn**: When people are taking turns talking, having a conversation, a single utterance will never span over more than one person's turn.

- **Utterances are never more than one complete sentence**: Complete sentences include things like "the man is walking" and "I'm sleepy". If sentences are joined by conjunction words (like "and", "or", "but", "because", "after", "for", "so", "if", "when", etc.), then they are transcribed them together on a single utterance line, but if there are no conjunction words, the sentences are separated utterances.

- **Using semantic and syntactic cohesion to determine utterance boundary**: In general, when phrases are related semantically and grammatically, they should are transcribed together as one utterance if there is less than a two second pause between them. For instance, the phrases "I'll go first" and "and then you can have a turn" are related semantically because after "I'll go first" is the time when you can have a turn, and the conjunction words "and then" connect the phrases to each other grammatically.

- **Using intonational contours to help determine utterance boundary**: If two phrases are part of two totally separate intonational contours, they are transcibed as two separate utterances instead of a single utterance. Intonational contour refers to the pattern the pitch of your voice makes when you utter questions, propositions, and commands. An example is shown in figure 4.9.

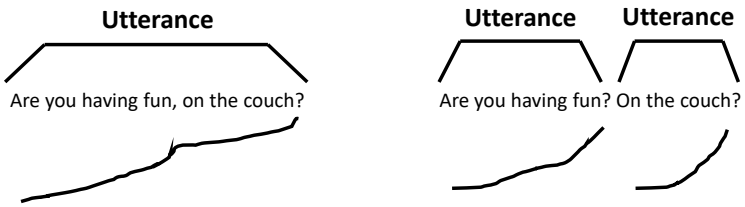

FIGURE 4.9: Examples of intonational contours.

- **Repeated words and phrases**: If a speaker keeps repeating a word or short phrase over and over again (e.g. "no no no, no no, no no no....no!" or "I won't I won't I won't I won't!", or is counting, saying the alphabet, or reciting a list, the intonation and 2 second pause rules are used to find the boundary.

Chapter 5

Method Description - Emotion Recognition System

This section describes in detail an **Emotion Recognition System** implemented in this project. A system based on **CNNs (Convolutional Neural Networks)** has been implemented for recognizing emotions, so audio and video has been used as sources for the system. As well in the first subsection it is described all the *preprocessing* followed before entering the network. As it is seen in the next section, a **public database (IEMOCAP)** has been used to develop our systems.

5.1 Data Preprocessing

As commented before, the **IEMOCAP database** is used for the experiments. This database is composed of videos, therefore images and audios are used as sources of information (see Sec. 2.4.2). The next chapter will be about the amount of information that were finally used and how structured all videos for training and testing are structured. This section is describing how to treat and preprocess all videos before entering the data to the network.

5.1.1 Audio Preprocessing

In the IEMOCAP database, the audio is apart from the video, but they are completely synchronyzed except for the final, where it is mandatory to deal with duration. The audio format is *.wav*, with 44100 Hz of sample rate and stereo channels. The main idea of the preprocessing step is obtaining small windows of the audio centered around an image and training the model with this windows (extracting features directly with the network or with an external feature extractor). The following list presents the main steps that are followed through all preprocessing:

1. **Stereo to mono**: The first step is converting stereo audio to mono. To do this, it is used a python library called *pydub*, after isolating and organizing the audio in a new folder.

2. **Synchronize duration**: In general, the audio has longer duration than video. It was not a problem of synchronization, but the video was cut before the audio finishes. Therefore, using a library called *moviepy.editor*, duration of each video was extracted and saved in a *.txt* file.

3. **Window segmentation**: Now, using again *moviepy.editor* and the *.txt* file duration to compute the real length of the audio, each audio is read and then

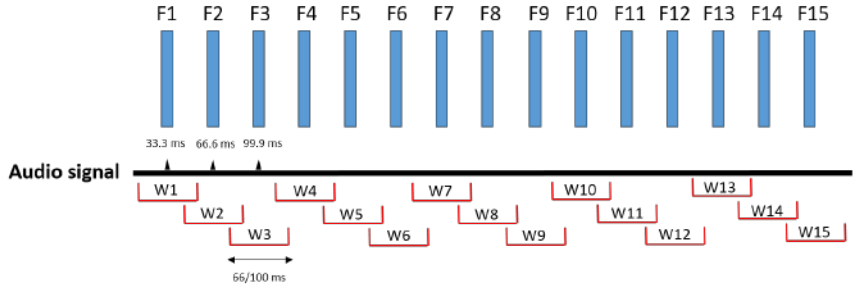

FIGURE 5.1: Window segmentation step. F: Frame. W: Window.

the number of divisions necessary for the window division is computed. In this case, because the video was 29.97 frames/s, it is advisable to use an audio window step of 33.3 miliseconds and centered around each frame, with length 66 ms or 100 ms. For example, for length 100, there will be 50 miliseconds protruding on each side. Fig. 5.1 shows graphically the process.

4. **Window selection**: Once all audio windows has been created for each frame, only those windows that fall inside an utterance with an emotion in any set are selected.

5.1.2 Video Preprocessing

All videos are in *.avi* extension, with frame rate of 29.97 frames/s and frame size of 720x480 pixels. This subsection is going to enumerate the steps that are followed to achieve the final face, that is the input of our system. First of all, folders that contains each video are reorganized. Then, the next steps are followed (Fig.5.2):

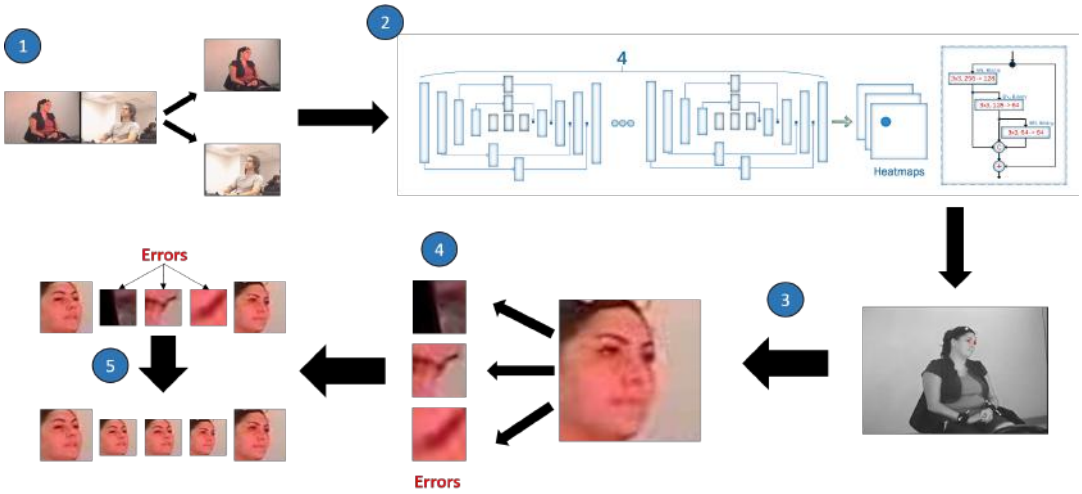

FIGURE 5.2: Video preprocessing steps. **Number 1** is the step of cropping videos and it is the input of the Face landmark detector, which is the step **Number 2** (Hourglass architecture). (**Number 3**) is the face alignment using the position of the eyes. **Number 4** is the error detector. The interpolation after the error correction is **Number 5**.

1. **Crop videos**: As it is illustrated in 5.2, in a single video there are two people talking. The authors of the database jointed both camera views in one video.

This is a problem for face detectors, because sometimes it is going to detect different speakers in consecutive frames. This problem was fixed just cropping the video in the middle (360x480) with *moviepy.editor* and silencing the audio, so two videos from one with this new size are obtained. Then, it happened that in some videos a person appears at the back, so the face detector starts missing some frames. To solve this problem, each video is cropped manually to avoid wrong faces.

2. **Face landmark detector**: For the face landmark detection a software called *OpenFace* was used at first, but it had some drawbacks when the face was not in front position. The *OpenCV library* was used to read each video and save all the frames in other folders. Once this new frame-database was obtained, a more successful landmark detector was applied to each image. The landmark detector predicts a set of 68 2D landmarks. The face landmark detector and alignment are from [5] library. They coin the network used for their experiments simply Face Alignment Network (FAN). They construct FAN based on one of the state-of-the-art architectures for human pose estimation, namely the HourGlass (HG) [26]. They go one step further and replace the bottleneck block with the recently introduced hierarchical, parallel and multi-scale block of [39].

3. **Face alignment**: The next step is aligning the image using the eyes. The implemented method first computes the center of mass of each eye landmark detected and then computes the angle between both centroids. Then, it determines the scale of the new resulting image by taking the ratio of the distance between eyes in the current image to the ratio of distance between eyes in the desired image. Angle, scale and median center are used to compute the rotation matrix and then the affine transformation is applied to obtain the face aligned. Then, the face is rescaled to the same size of the input of the network (224x224x3).

4. **Face error detection**: Because there are lot of faces in black (because there is not face detected) or in a wrong position (because of the wrong position of the detected landmarks), an error detector that detects when a face frame is incorrect was created. Some conditions that almost detect all error were found: if the angle between both eyes is bigger than 60 degrees and smaller than 318 degrees, it is a wrong face. It happens the same for the scale and the distance between eyes, but it depends on the video, because each person has different head size.

5. **Face error correction - interpolation**: Once all the wrong faces were detected, an interpolation was made between frames to fill the gaps of the wrong faces. The next conditions were followed: if an utterance losses more than 30 consecutive frames, this utterance is discarded because it has more than one second missing and this can cause some problems. If a sequence of less than 30 frames is found, an interpolation is made just filling the left part of the gaps with the left face and the right part with the right face. To clarify this correction, an example could be if there are missing 20 frames, it has to be repeated 10 times the frame just in the left and 10 times the frame just in the right. This method will discard some utterance (for example with face occlusion), but as well will recover a huge number of them.

5.2 Proposed System - Local Level

There are three important steps in a multimodal scenario for emotion recognition. The first one is the election of the architectures or methods to obtain the unimodal features. The second one is the fusion strategy. The last one is the implementation of a recurrent model or not. In the following sections, this three steps are going to be explained along with the architectures that have been used for extracting the unimodal features and for the fusion methods.

First of all, one of the objectives of this work is to know how accurate a model can be only at frame or local level, which means not using any context or temporal information. For this purpose, three different architectures are implemented: one for raw audio, one for audio features and one for faces images, which input is only a segment of audio, a feature vector or an image, respectively, and their output is the label of that input. For performing this, each audio segment, feature vector or image has its own label, which is the label of the utterance to which it corresponds. This means that each frame/face has associated its own audio segment and, for correspondence, its own audio feature vector. This results in a database with the same number of frames, raw audio segments and audio feature vectors.

One of the reasons of doing this is because after analyzing the unimodal systems at frame or local level, the benefits of fusing them by pairs or trios is studied. In Sec. 5.2.3 these fusion architectures are explained.

Another point to take into consideration is the optimizer and the loss function. For all this architectures the **Adam** optimizer with learning rate of 0.001, $beta_1$ of 0.9 and $beta_2$ of 0.999 were used. The same loss function for all the architectures was also used: the **Categorical Cross-Entropy** loss function because it is a multi-class classification problem. The expression of this function is

$$L(y, y^*) = -\sum_{j=0}^{M}\sum_{i=0}^{N}(y_{ij} * log(y_{ij}^*)) \tag{5.1}$$

, where y^* is the predicted value.

5.2.1 Audio Architecture

Following the state of the art (see Sec. 2.3.1), the most implemented method for obtaining the audio features is using the OpenSMILE toolkit. In this work, two different strategies were implemented: 1) **OpenSMILE** with audio segments of 66 and 100ms, and 2) a deep learning network using **1D convolutional layers** to obtain the features of audio segments of 66 and 100ms and classify them directly.

Raw Audio Architecture

Fig. 5.3 shows the architecture that was used for classifying the raw audio segments. Different architectures were tried but this is the one with best performing. The input of this architecture is a raw audio segment of 66 or 100ms. This input is passed through four consecutive 1D Convolutional layers with 16, 32, 64 and 128 filters, respectively. All these convolutional layers have a filter size of 9, ReLU as the activation function and are followed by a Pooling layer: the first three ones by a MaxPooling layer with *pool_size* of 2 and the last one by a GlobalAveragePooling

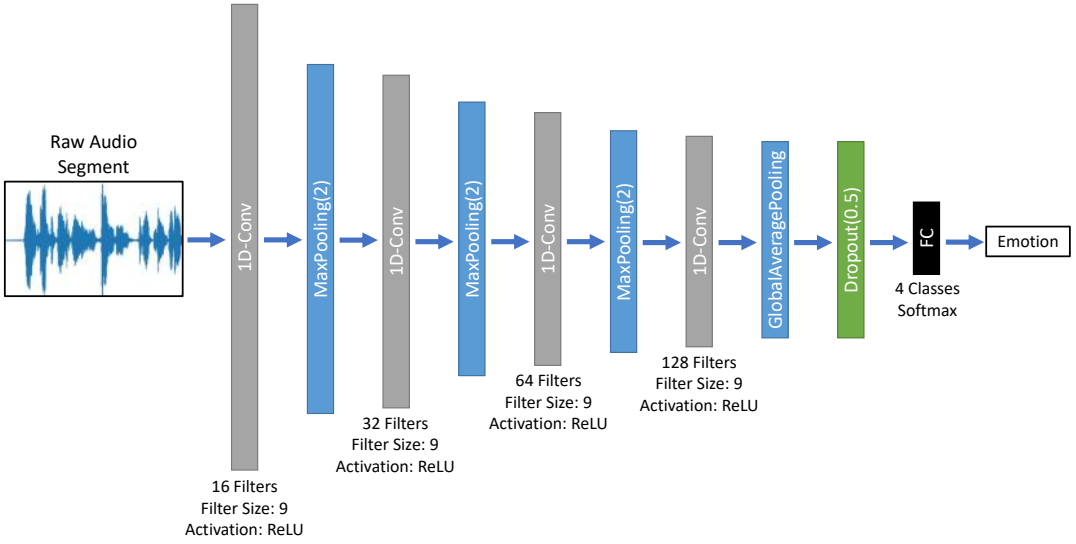

FIGURE 5.3: Raw audio architecture. Input: Audio segments of 66 or 100ms. Output: Label.

to reduce the dimensionality. Until this layer, the network can be seen as a feature extractor, which output is a feature vector of size 128.

In order to perform the emotion classification, a dropout layer with rate 0.5 is inserted and finally a fully-connected layer with 4 units (because there are 4 emotions) and **softmax** as the activation function.

Audio Feature Architecture

The architecture for classifying the audio feature vectors is simpler than the one explained for raw audio, but it requires an additional preprocess. First, each raw audio segment of 66 or 100ms is passed through the OpenSMILE toolkit to obtain the audio feature vectors. The configuration file *'IS13-ComParE'* is used which obtains a total of 6373 features (See Sec. 2.3.1). The next step is to normalize each of the features to the range [0,1]. For this task, the maximum and minimum for each feature is computed using the feature vectors of the training and validation set. This has been done like this because this two sets are the ones used during the training process. Once the 6373 maximums and minimums are computed, all the feature vectors of the training, validation and test sets are normalized using this formula

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{5.2}$$

. Another important part of this process is to delete the features that are not important to the training process. When computing the maximum and minimum, for some features their maximum and minimum are the same, which means that those features have the same value for all the vectors. These features are deleted, obtaining a total of **2099** features for raw audio segments of 66 ms and **5266** features for raw audio segments of 100ms.

After obtaining the feature vectors, they are passed through two consecutive fully-connected layers with 256 and 128 units respectively. This part of the model can be seen as a feature reduction. After that, the classification is performed using

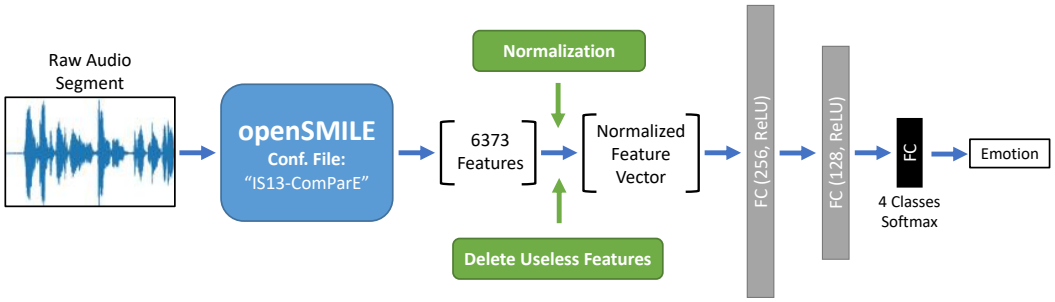

FIGURE 5.4: Audio feature architecture. OpenSMILE, feature normalization, removal of useless features, feature reduction and classification. Input: Audio segments of 66 or 100ms. Output: Label.

another fully-connected layer with 4 units and a **softmax** as the activation function (Fig. 5.4).

5.2.2 Video Architecture

For recognizing emotions from images, a CNN pretrained architecture is proposed. In this case, the **VGGFace** pretrained model from *keras* is used, with some modification for the right working of the process. In figure 5.5, the final architecture is shown and the next paragraph is going to describe the main changes that are performed to the original architecture.

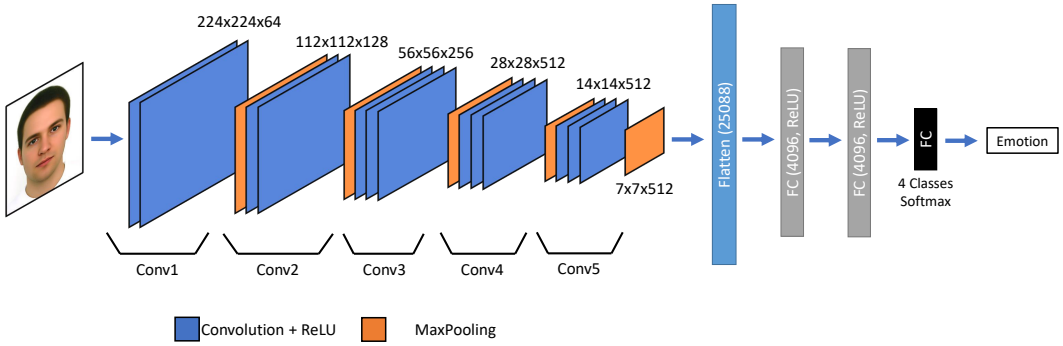

FIGURE 5.5: Video architecture. VGG Face for emotion recognition. Input: Face with size 224x224x3. Output: Label.

First of all, the pretrained model was trained with **VGG-Face database** [22]. The dataset contains up to 1000 images per identity and 2622 subjects. Images are downloaded from Google Image Search and have large variations in pose, age, illumination, ethnicity and profession (e.g. actors, athletes, politicians). For this reason, the original VGG Face architecture has a final softmax of 2622 neurons, to classify each subject. The objective of this project is to predict 4 emotions, so the final softmax is changed from 2622 to 4 neurons, keeping the rest of the architecture with the same weights. The idea of using this network is making **fine-tuning** over some layers of the VGG and freeze the rest, to get a better representation with our data. Therefore, the first three convolutional blocks are frozen and the next two convolutional layers with the fully-connected part are trained with new data.

The architecture is made of 5 convolutional blocks and the first convolution has an input of 224x224x3 (face in RGB). All convolution inside each block have a ReLU activation. The first convolutional block has 2 convolutions with 64 filters each one and *Zero padding* equal to 1, followed by a *Max Pooling layer* with stride equal to 2 to reduce dimensionality. The next convolutional block is the same, but increasing the numbers of filters to 128. The third convolutional block adds another convolution (a total of 3 convolution) with the same Zero padding and 256 filters instead of 128, computing a Max Pooling again at the end of the last convolution with stride equal to 2. The fourth and fifth convolutional block are the same, with 512 filters each convolution followed by the same Zero padding and a final Max Pooling. After the final Max Pooling, the final image is flattened and pass through 2 fully connected layers of 4096 neurons each with ReLU activation. The final 4096 neurons are connected to a fully connected layer with softmax activation for the emotion classification.

5.2.3 Fusion Architectures

Once all audio and video architectures are created and trained, the next step is to fuse them to obtain new representation of features. The final fusion architecture after some attempts is shown in figure 5.6. It is performed in three different ways: fusing raw audio architecture with video, fusing handcrafted features architecture with video and fusing all three architectures.

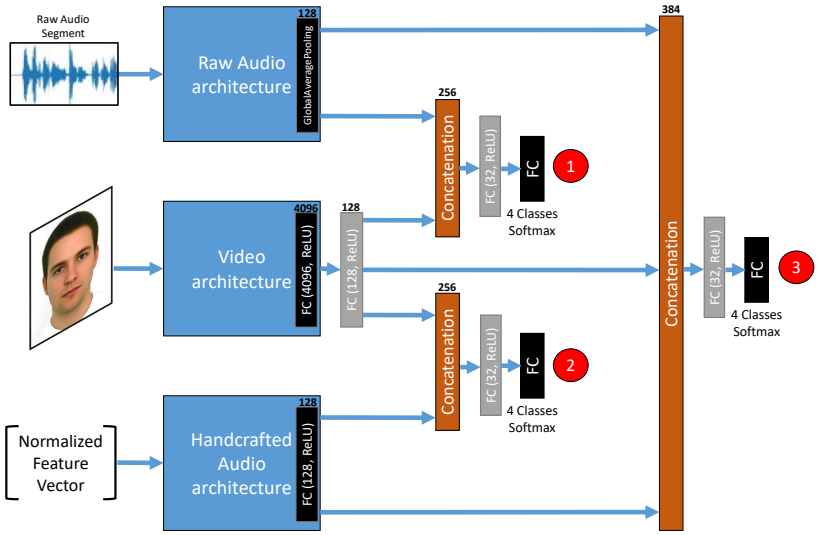

FIGURE 5.6: Fusion architecture. Raw audio and video fusion (1). Handcrafted feature and video (2). Raw audio, hancrafted features and video (3).

As it was explained in the previous chapters, there are the same number of video frames as number of windows, so the fusion is made at local level. The first step for the fusion is making a dimensionality reduction in video architecture. To do that, the output of last dense layer from the VGG Face architecture is taken and is connected with a fully connected of 128 neurons to reduce the vector of 4096 to 128, forgetting about the softmax layer. This new fully connected has learnable parameters and will be adapting them during training. For both audio architectures, the new system does not take into account the output of the softmax and uses the last feature vector of size 128 for the fusion. For the raw audio is the output of the Global Pooling and for the handcrafted is the output of the last fully connected layer.

The next step is to simply concatenate the feature vectors of size 128. For this, 3 different systems and combination are proposed. The first combination is between raw audio feature vector and video feature vector, with a final length of 256. The second combination is between handcrafted audio features and video obtaining the same length as the first combination and the last combination is between all feature vectors, obtaining a last feature vector of length 384. Once the concatenation is done, all architecture fusion have the same procedure for the classification. The new feature vector is connected to a new fully connected with 32 neurons and finally with a fully connected with softmax activation for the final emotion classification. Basically, what is done is just freezing all previous layers from all architectures and train the new fully connected layers added to the system.

Chapter 6

Method description - Speaker Segmentation

This section describes in detail a **Speaker segmentation** implemented in this project. This speaker segmentation system is based on feature extraction and clustering and is a baseline for future works.

6.1 Audio Preprocessing

For the speaker segmentation, the **IEMOCAP database** is used, but only one audio that does not have the problem of overlapping speech. The first step is to annotate each utterance of the audio. The utterances annotated by the authors are quite different than the ones this methods expected to have. The real ground truth (start and end of an utterance) should be very accurate to decrease the error (DER) of the segmentation. For this reason, a part of an audio speech is reannotated taking into account just the start and the end of the audio and not the facial expressions.

The speaker segmentation architecture is shown in figure 6.1. The next step in the preprocessing is to convert the audio from stereo to mono, because the first block of the audio segmentation (VAD) works better if the mono audio is obtained. This is done with a library in python called *pydub*. The audio was selected carefully because nowadays the diarization techniques does not work well with overlapped speech or with background noise. For this reason, an exhaustive search has been made to find a representative audio in IEMOCAP to perform this experiments correctly.

6.2 Speaker Segmentation Steps

After annotating again the audio speech, the speaker segmentation can be divided in 4 steps:

1. **VAD (Voice Activity Detection)**: first of all, is recommendable to eliminate those silence parts from the speech audio, because it can be a trouble for the classifier to classify the silence (the cluster would have to have 3 centroids). So it is applied (from a library in python called *auditok*) a VAD to the audio signal and it gives back the timestamp where a speech has occurred. This VAD works as follows: there is some parameters that can be controlled (Minimum length an accepted audio activity should have, Maximum length an accepted audio activity should reach, Analysis window length, etc). Table 6.1 shows the parameters that were selected from different experiments with the audio.

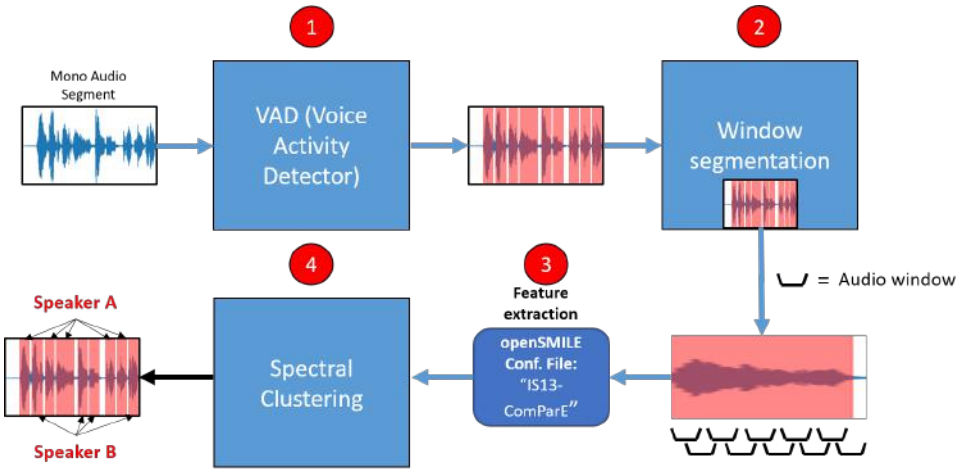

FIGURE 6.1: Speaker segmentation steps. All segmentation process after preprocessing the audio.

If the experiments were made with different audios, a normalization preprocessing would have been favorable to find the optimal parameters. This VAD takes small analysis windows and computes the energy inside those windows. Then, if the energy of the window rises a threshold, then the increase in energy is assumed to be to associated with voice.

	Unit	Value
Minimum length an accepted audio activity should have	second	0.4
Maximum length an accepted audio activity should reach	second	20
Maximum length of a continuous silence period within an accepted audio activity	second	0.3
Drop trailing silence from an accepted audio activity	boolean	False
Analysis window length	second	0.02
Energy threshold	-	50%

TABLE 6.1: Table with the paramters of the VAD.

2. **Window segmentation**: Once there are segments of audio with start and end, the next step is to divide these segments in windows. As it is described in the experiments section, a grid search has been made to find an optimal value of window and overlap between consecutive windows. The best window length for the cluster was 240 ms with an overlap between them of 30%. If it happens that a window goes off the segment, this window will be fulfill with zeros.

3. **Feature extraction**: For the feature extraction, the **OpenSMILE** is used as in emotion recognition systems, because it is a very reliable feature extractor and works very fast. The system uses the configuration file '*IS13-ComParE*', which obtains a total of 6373 features (See Sec. 2.3.1)

4. **Spectral clustering**: After obtaining the feature representation of each window, a *Spectral Clustering* is used to classify each window as Speaker A or Speaker B, after trying other type of clusters. First of all, a similarity matrix is constructed with the feature vectors. The final affinity matrix will be 6373x6373. These similarities are made with cosine distance, in this case. Then, a refinement based on the temporal locality of speech data is applied. This is

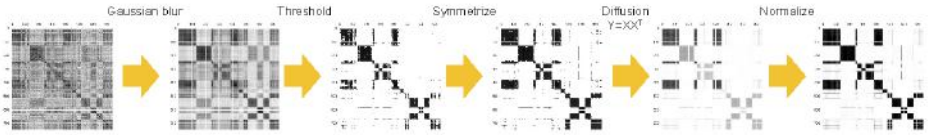

FIGURE 6.2: Refinement process before K-Means in Spectral Clustering.

important because contiguous speech segments should have similar embeddings, and hence similar values in the affinity matrix. The steps in the refinement are shown in figure 6.2. Row-wise thresholding serves to zero-out affinities between embeddings belonging to two different speakers. Symmetrization restores matrix symmetry which is crucial to the spectral clustering algorithm. The diffusion steps draws inspiration from the Diffusion Maps algorithm and serves to sharpen the image resulting in clear boundaries between sections of the affinity matrix belonging to distinct speakers. Finally, the row-wise max normalization serves to rescale the spectrum of the matrix to ensure undesirable scale effects do not occur during the subsequent spectral clustering step. After all refinement operations, it is performed an eigen-decomposition on the refined affinity matrix. The maximal eigen-gap is used to determine the number of clusters, but it is known a priori in the case of this project that is equal to 2. Finally, the k (k=2) eigen-vectors corresponding to the largest eigenvalues are selected for the classification. Then a K-Means (maxiter = 300 and random state = 0) is used to cluster these new embeddings, and produce speaker labels for each window. Once there is a prediction for each window, a majority vote inside each segment predicted by the VAD is applied. This last step is computed to have a single prediction for each utterance and compare the final DER with the ground truth.

Chapter 7

Experiments and Results

This section discusses the results obtained in chronological order using the architectures from Chapters 5 and 6, as well as the environment used for the experiments and the experimental protocol followed in each experiment. Remember that all experiments have been performed on the *IEMOCAP database*, except a proof of concept tested over a subset of our database.

7.1 Experimental protocol

In this section, the experimental protocol and the environment used for the experiment is going to be discussed.

7.1.1 Environment Adaptation

All the experiments and part of the database organization and preprocess have been done in a server from *HuPBA (Human Pose Recovery and Behavior Analysis)*. This server has 4-GPUs (GeForce GTX TITAN X) and the OS is Linux (Ubuntu 16.04.6). This allows the experiments to run in parallel and go faster parallelizing the code in more than one GPU. The GPU capacity is enough to use a batch size for the images of 16 and for audio a batch size of 128. To connect to the server a *ssh* protocol is used (it is a secure protocol to connect via internet). Tu upload data to the server, **Filezilla** was used with the **SFTP** protocol to transfer data from local computer to the server.

All code created is in **Python** and the code reused is written in *Java* and *C++*. The code is placed in this *GitHub repository*. For the python code, the main libraries used to develop the networks are **Keras** (over *Tensorflow*) and **Pytorch**. For the data processing it is used as well *Pandas* and *OpenCV*. For the speaker segmentation, the library **auditok** and *sklearn* for the spectral clustering was used. To edit and debug the code, the favorite editor was Spyder. To run the code in the server, **Docker** has been used to package all the code and its dependencies (libraries, data, etc) in a virtual container. For this process, two images have been created with the respective libraries and dependencies to run all the codes. These images are in charge of creating the final containers.

7.1.2 Emotion Recognition Protocol

As it has been explained in Sec. 5, the IEMOCAP database has been used for all the experiments. This database (see Sec. 2.4.2) contains 5 sessions with 151 videos in total (28, 30, 32, 30 and 31 respectively). As it is done in the state-of-the-art literature,

		Train	Validation	Test	Total
OpenFace	Frames	299.545	67.077	149.244	515.866
Proposed Method	Frames	451.696	104.389	143.648	699.733
	Utterances	3.275	814	1.093	5.182

FIGURE 7.1: Number of extracted faces with OpenFace. Number of extracted faces and utterances for the proposed method.

the 5^{th} session have been used as the test set and the other 4 sessions as training sets. Following the same strategy than **??**, the 20% of the training set have been used as the validation set, taking the first and last 3 videos of each session. Thus, the final division is: 96 videos for training, 24 for validation and 31 for testing.

The utterances of the IEMOCAP database are labeled with 10 different emotions: *frustration*, *surprise*, *fear*, *disgust*, *happy*, *neutral*, *sad*, *angry*, *excited* and *other*. Following the state-of-the-art, only four emotions have been used for the experiments: **happy**, **neutral**, **sad** and **angry**, considering all the *excited* utterances as *happy*. Each utterance is labeled by three annotators. When there is not a consensus between them, (for example, two annotators label a utterance as happy and the other one as sad), the emotion is labeled as *XXX* and this utterance is not taken into account for any of the sets. A Python script is coded to read all the annotation files and extract the beginning and final time of the utterances whose labels are the ones selected for this work.

For extracting the faces, the method explained at Sec. 5.1.2 is used. Table 7.1 shows the number of faces extracted with this method and with the *OpenFace* library. Using this method, only the **7.52%** of the faces is lost, while the percentage of lost faces for the OpenFace method is **31.82%**, so it was discarded due to the number of information lost. This table shows the number of images that are going to be used for training, validation and test sets, which are the **65%**, **15%** and **20%**, respectively. This division, in terms of number of utterances, means **3275** for the training set, **814** for the validation set and **1093** for the test set. All these values are the same for faces, raw audio segments and audio feature vectors, to facilitate pairing in fusion models.

With the initial and final time of each utterance and the extracted faces, audio segments and audio feature vectors, some scripts are coded to organize these files in three sets. Each folder (*train*, *val* and *test*) has 4 sub-folders (*ang*, *hap*, *neu*, *sad*), in which all the files belonging to that label and that set are placed. This organization is due to the **generators** that were coded for the training processes. A different generator was coded for each modality, changing the way the files are loaded in memory and the dimensions of the data.

The training of each modality and the fusions are performed using the *fit_generator* function of *keras* and the coded generators. After that, the best model is chosen and a **evaluation** process starts. During this processes, the model is evaluated as follows:

1. **At local level**. The performance of the model is evaluated at local level for the training, validation and test sets, which means that every input file has its own label.

2. **At utterance level**. In this case, two strategies are implemented: the prediction of an utterance is the **mean** or the **majority vote** computed from the value of the final softmax. In this evaluation, the performance of the model is also

computed at utterance level for each label for the training, validation and test sets. For this computations, some python dictionaries is generated in which each key is the number of the utterance and its corresponding value is a list with the files that integrate that utterance.

In order to make the accuracy tables more understandable, only the utterance results of the **mean** are going to be shown, because in general, they perform better than the majority voting. In order to compare with the state of the art methods, the macro **F-Score** is computed for the model with higher accuracy.

One important step in all this training process is the **data augmentation**. All the images were flipped in the vertical axis to have the double amount of data in the training set, obtaining a total of **903.392** training faces. This operation is computed because it makes the system more robust to different face positions. The validation and test sets are the same as before. In order to be coherent during the fusion methods training, the corresponding raw audio segment and feature audio vector for a face and its flipped face is the same, so they were loaded only two times when training.

7.1.3 Speaker Segmentation Protocol

As it has been explained in Sec. 6, for the Speaker segmentation the audio used to perform the operations was a single fragment of an audio selected from **IEMOCAP**. In particular, this audio was from the first session and is a conversation between a male and a female. It will be discuss later in Sec. 7.2 why this is an advantage, but makes the system less robust to new audios. This is because males and females has more difference between their features than males (Pitch, MFCC, etc). The length of the audio is **81 seconds**, enough to have a huge number of windows to be analyzed. This speech was the longer audio fragment found without overlapping between speakers. The total number of **utterances annotated is 29**, so the first block of the audio segmentation (VAD) expects to divide the audio in this number of utterances.

Because this is an unsupervised learning method, there is not need to divide the data in train, validation and test. For this reason, a representative audio was selected to test this segmentation method. The metrics used are *DER (Diarization Error Rate)* and *Accuracy at utterance level*. The DER is obtained for individual windows and the final ground truth and for the majority vote inside an utterance and the ground truth.

A **grid search** was made to find the best parameters for the window length and overlapping and for the spectral clustering. The results and parameters obtained will be discuss later in Sec. 6.

7.2 Experimental Development

In this section, the experiments are going to be introduced in chronological order and are going to be commented in detail through the explanation. First of all, for the metrics in emotion recognition, the accuracy has been computed in two different ways: at **frame level** (just the accuracy for each frame, window or feature vector) and at **utterance level**. To compare the results of this project, state of the art results from [8] are going to be shown with *F-Score* metrics.

7.2.1 Unimodal Faces Experiments

These experiments have been launched along 20 epochs. The final results are shown in Table 7.1 in terms of accuracy. The table shows the *accuracy* at frame level, at utterance level with mean score and the *F-Score* for the best model at utterance level in order to compare with the state of the art.

	Accuracy			F-Score
	Train	Validation	Test	Test
Frame level	78.50	39.20	36.63	-
Utterance level	91.87	45.33	40.62	37.72
Utterance level from [X]	-	-	41.79	41.79

TABLE 7.1: This table shows the **accuracy** for the VGG at frame level and utterance level and the **F-Score** at utterance level. It shows as well the F-Score of the state of the art.

At frame level, the model with **39.20%** of accuracy is kept (the one with higher accuracy at validation). This results is just at frame level, so it cannot be compared with other state of the art methods. When computing the test accuracy, a **36.33%** is obtained. This is not something striking because the validation set contains videos from actors that also appears in the training set, and the actors from the test set do not appear in the training or validation, meaning that the validation set is not representative with the test set. But this division in train, validation and test is the same that the state of the art literature uses. When choosing the final model only the information of validation is available, so this validation set is not going to let the system generalize to other videos. This can be seen in the low accuracy of test.

The IEMOCAP has a low resolution image, so the faces are not detected in a correct way. As it was commented in Subsec. 5.1.2, a new face detector was used to fix the problem of missing faces at some frames. This problem is highlighted looking at the accuracy and if it is compared with audio method or state of the art, for example. The network cannot achieve good results with faces that are not correctly aligned or just have resolution problems.

At utterance level, the accuracy increases as it was expected because, imagine that in a happy utterance with 200 frames, there are 120 with happy labels and at frame level is just a 60% of accuracy, while in utterance level is a 100%. The validation accuracy for utterance level is **45.33%** and for test is **40.62%**. This model is 4.07% under the state of the art model from [8] (**41.79%**) in terms of F-Score, where, to capture the temporal dependence, they transform each pair of consecutive images at$t$t and $t+1$ into a single image and then they provide this input to a multi-level CNN. This difference of modelling the temporal aspect may be the reason why they achieve this difference at utterance model. As well, comparing the training accuracy and the test accuracy in the proposed method, there is an intuition that the system is overfitting. Maybe, the network is very complex for recognizing emotions and in future work it will be proposed new architectures.

7.2.2 Unimodal Audio Experiments

In this subsection, the architectures of **raw audio** and **handcrafted features** are going to be analyzed together. These experiments have been launched along 1000 epochs.

	Architecture 1	Architecture 2	Architecture 3
66 ms	**42.10**	42.00	41.80
100 ms	**42.80**	42.40	42.50

TABLE 7.2: **Accuracy** for the different architecture of handcrafted features audio in validation set. The window analysis are 66 and 100 ms.

Table 7.2 shows the results obtained with handcrafted features for different architectures. The results are from the validation set in terms of *accuracy* for the best models. The first architecture shows the highest accuracy for a window analysis of 66 and 100 ms. This architecture is shown in Sec. 5.2.1 (Fig. 5.4). The other two architecture have the same structure except for the new fully connected layers that have been added. For architecture 2, a fully connected of 512 neurons is introduced before the fully connected of 256 neurons. In addition, for architecture 3 a fully connected of 1028 neurons is added before the fully connected of 512 neurons from architecture 2. Noticeably, the first architecture works better and this is because the number of parameters is smaller, so it generalizes better validation and test data than the other two architectures. If the number of data/samples increases, it could be reasonable to use architecture 2 or 3.

		Raw Audio				Handcrafted Features				Baseline Audio Method
		Frame level		Utterance level		Frame level		Utterance level		Utterance level
		66	100	66	100	66	100	66	100	100
Accuracy	Train	42.20	43.00	51.84	52.39	40.50	40.30	47.39	44.21	-
	Validation	41.40	41.80	51.84	51.10	42.10	42.80	50.98	50.12	-
	Test	41.43	41.84	52.24	**53.52**	39.28	39.72	49.31	49.22	-
F-Score	Test	-	-	48.87	**51.32**	-	-	47.32	45.40	**51.52**

TABLE 7.3: **Accuracy** for raw audio and handcrafted features audio at frame level and utterance level and the **F-Score** at utterance level. F-Score of the state-of-the-art baseline [8].

In table 7.3, the results of raw audio architecture and architecture 1 for hand crafted features are shown. The first analysis is going to be about the **window length**. For the Raw audio information, at frame level and using a window length of 100 ms, a **41.84%** of accuracy is obtained for test set. Compared with 66 ms for the same set, it is **0.41%** higher. It happens the same for validation set. At utterance level and for test set, a **53.52%** is obtained with 100 ms, **1.28%** higher compared with 66 ms. In general, when extracting the features directly from the raw audio, it is clear that for 100 ms of window length is going to work better.

For handcrafted features, at frame level it works better with 100 ms (39.72%) than 66 ms (39.28%) of window length. When computing the accuracy at utterance level, for 66 ms the accuracy is higher. For test set, it achieves a **49.31%** of accuracy. The number of features extracted from a window of 66 ms is 2099 and for 100 ms is

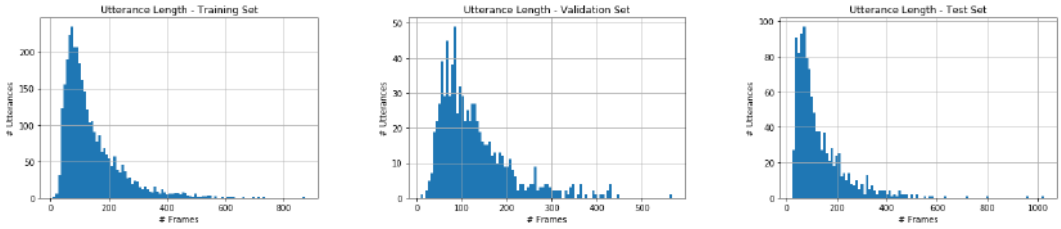

TABLE 7.4: Number of frames per utterance in each set.

5266. This feature vector of 100 ms is almost the double of 66 ms, so for our architecture it should be working better the 100 ms window length. This is something that happens at frame level as it was expected, but not for utterance level. One of the possible explanations is shown in table 7.4. Following the histogram, the number of frames per utterance (length of the utterance) is in general less than 100 frames for all sets. Although with a longer window the time resolution increases, there is a lost of frequency resolution [21] when computing the STFT to extract features, what could make the system fail in shortest utterances with 100 ms. This would explain why at utterance level with handcrafted features, the system performs better for 66 ms.

Comparing both methods at frame level, a higher accuracy in test set is achieved. For validation set, the accuracy is higher for handcrafted features, but this is not something significant. At utterance level, the accuracy of raw audio architecture is 3.02% higher than handcrafted feature for 100 ms. For this reason, extracting features directly from raw audio over handcrafted features has become very popular and is used more frequently.

The state of the art just extracting features with **OpenSmile** directly from a window length of 100 ms is **51.52%** in terms of F-Score. In this case, our result is very close to the state of the art (51.52%), just **0.30%** of difference . This confirms that extracting features directly from raw audio can work better than using an external feature extractor such as **OpenSmile**.

7.2.3 Fusion Experiments

In this section, the three fusion experiments explained on Sec. 5.2.3 are analyzed. As it has been seen in the previous section, at the unimodal experiments with raw audio segments and audio feature vectors, using 100 ms obtains better results than using 66ms. For this reason, only the fusion methods that use **segments of 100ms** are going to be analyzed.

To perform the fusion methods, the best model of each modality is loaded and all convolutional and fully connected layers are frozen. Only the new layers after the concatenation are trained. All the fusion methods are trained between 10 and 20 epochs because of the amount of the data, the computational time and the fast convergence of the training.

Table 7.5 shows the results obtained for the three fusion architectures. First, it is going to be analyzed the comparison between fusion methods at utterance level. Again, the utterance level performs better than the frame level. Looking at the first 4 columns of the table, it can be compared the performance of the fusion of faces with the raw audio and the fusion of faces with the audio feature vectors in terms of accuracy and F-Score (only for test set). As it could be expected from the unimodal

		Faces + Audio100		Faces + Feat100			Faces + Audio100 + Feat100	
		Frame Level	Utterance Level	Frame Level	Utterance Level	Baseline Utt. Level	Frame Level	Utterance Level
Accuracy	Train	86.40	96.06	86.10	96.09	-	**86.50**	**96.24**
	Validation	43.30	50.24	42.90	49.63	-	**43.70**	**50.73**
	Test	38.70	44.55	38.62	44.09	-	**39.17**	**44.73**
F-Score	Test	-	41.71	-	41.18	**52.15**	-	41.86

TABLE 7.5: Accuracy of fusion methods for training, validation and test. Utterance level. Raw Audio and Audio Features from 100ms. Baseline in terms of F-Score [8].

scenario, the fusion with raw audio obtains a better performance with respect to the fusion with feature vectors with a difference of **0.61%** of accuracy in the validation set and **0.46%** in the test set (at utterance level). The difference in terms of F-Score is **0.53%**. The last column shows the results of the trimodal fusion (Faces + Audio of 100ms + Feature Vectors of 100ms) at utterance level. This fusion method outperforms the results of the previous ones in accuracy and F-Score, which is indicating that both raw audio and feature vectors helps the faces model in different ways to obtain a better performance. With the trimodal fusion, the accuracy is **50.73%** for the validation set and **44.73%** for the test set. Adding the feature vectors improves **0.49%** in validation set and **0.18%** in the test set with respect to the bimodal fusion of faces with raw audio. But this methods are not capable of improving the state of the art [8], which obtains **52.15%** of F-Score for the fusion of faces and audio features.

Table 7.6 show the accuracy of the unimodal, bimodal and trimodal architectures proposed in this work in order to compare between modalities. Focusing only on the validation and test sets, it can be appreciated again that the accuracy for validation set is better than for the test set. This is due to the fact that the validation set is not representative with the test set. Besides, the fusion models are limited by the accuracy of the faces. Both bimodal fusions and the trimodal one outperforms the accuracy of the unimodal face model in **3.93%**, **3.47%** and **4.11%**, respectively. Despite this improvement, this models do not beat the performance of the unimodal raw audio or audio feature vectors. This could be because the fully connected layers are given more importance to the faces features than the audio ones after the concatenation. On the other hand, these results are showing that the three modalities provide useful and different information to the model, and each of them improves the classification of the emotion recognition problem at local and utterance level. This also shows that the information extracted from OpenSMILE is different and compatible with the information that the proposed method extracts from the raw audio segments.

		Train	Validation	Test
Unimodal	Faces	91.87	45.33	40.62
	Raw Audio	52.39	**51.10**	**53.52**
	Audio Features	44.21	50.12	49.22
Fusion	Faces + Raw Audio	96.06	50.24	44.55
	Faces + Features	96.06	49.63	44.04
	Faces + Audio + Feat.	96.24	**50.73**	**44.73**

TABLE 7.6: Accuracy for all the scenarios considered at this work at utterance level. Raw Audio and Audio Features from 100 ms.

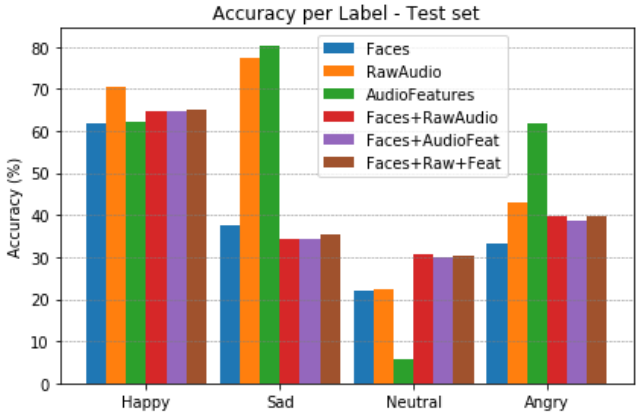

FIGURE 7.2: Bar charts of unimodal and fusion methods.

One way to extract some conclusions about the behaviours of the models is to analyze the performance of each of the emotions individually. Fig. 7.2 is a bar chart where the accuracy of each method for each emotion is represented.

The first conclusion is that the **neutral emotion** is the most difficult to classify for all the proposed methods and modalities, while the **happy emotion** is the one that, on average, obtains the best accuracy. This is probably because both face expression and speech during a happy utterance are very representative and different from the rest: the face is always characterize by the smile and the voice suffers a rise in tone. Raw audio and audio features perform better for sad emotions almost obtaining the double of accuracy than the rest of methods. This is reasonable because it is easier to know if a person is sad by the way he/she speaks rather than the expression of his/her face. Another important point is that the accuracy of the neutral emotion using audio features is really bad (less than 10%), whereas for angry emotion it manages an accuracy very similar to the happy emotion. This could be because OpenSMILE (or the complexity of the proposed method) obtains better features when there are changes in tone, amplitude, etc., and worse when the speech is monotonous. By last, the performance of the faces model and the three fusions is very similar. This confirms again than the face model and its overfitting have a lot of weight on the fusion models.

7.2.4 Speaker Segmentation Experiments

In this subsection, the speaker segmentation method proposed is analyzed and commented. First of all, the minimum DER that the system could have is computed, to have an idea of the results obtained through the processes. The **minimum DER is 0.282** and is not 0 because the VAD (Voice Activity Detector) introduces an error with respect the ground truth. The ground truth can have some errors when annotating, so is not significant if the minimum DER is 0.282. For this reason, the utterance accuracy is computed as well, not looking exactly at the start and end of the utterance, just at the utterance itself.

Table 7.7 shows the different parameters and results for the **grid search**. The grid search has been made for 6 different parameters: *Windows length* (66,100,160,200,240,300,500 ms), *Overlapping* between consecutive windows (30%,50%,70% and 90%), *P-percentiles* for the row wise thresholding (0.99,0.95,0.90,0.85,0.8), sigma value of the *Gaussian*

		66 ms	100 ms	160 ms	200 ms	240 ms	300 ms	500 ms
Parameters	Overlapping	30%	30%	30%	30%	70%	30%	70%
	Percentiles	0.99	0.99	0.99	0.95	0.99	0.99	0.99
	Gaussian blur	3	3	3	2	2	3	3
	Thresholding	0	0	0.5	0.5	0	0.5	0
	Stop eigenvalues	10^{-5}	10^{-5}	10^{-5}	10^{-5}	10^{-5}	10^{-5}	10^{-5}
DER		0.734	0.711	**0.634**	0.649	0.681	0.683	0.671
DER (Majority voting)		-	-	**0.535**	0.608	-	-	-
Utterance accuracy		-	-	**65.60**	55.20	-	-	-

TABLE 7.7: Grid search experiments for speaker segmentation. The metrics are for DER at windows level, DER at utterance level with majority voting and the accuracy at utterance level.

blur operation (1,2,3), the multiplier for *soft threhsold* (0,0.01,0.02,0.05,0.1,0.5), the *minimum eigenvalue* that is looked ($10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}$).

The results for each window length is shown. In general, the best overlap for consecutive windows is 30%. Except for 240 ms and 500 ms, is the one that obtains the best DER. The next parameter is the p-percentile for the row wise thresholding and in this case is 0.99 except for 200 ms. The best sigma for the Gaussian blur is 3, except for 200 ms and 240 ms. The multiplier for the soft threshold is between two values (0 and 0.5), but the best model was obtained with 0.5. The minimum eigenvalue is 10^{-5} for all windows and all overlap values. This grid search obtained the best results (**0.634**) for a **windows length of 160 ms and 30% of overlapping**. For small windows (66 and 100 ms) the DER value is over 0.700 and are the worst results. The best two windows length values are 160 and 200 ms and if the value is increased it gets lower DER. This is because there are a huge number of utterance that are smaller than 1 second, so using a windows analysis length of 300 or 500 ms is not going to work well. For the overlapping between windows, the smaller one (30%) has been selected and this is because there is not more than two speaker in one segment obtained from the VAD. For this reason, this two reasonable values are chosen from the grid search.

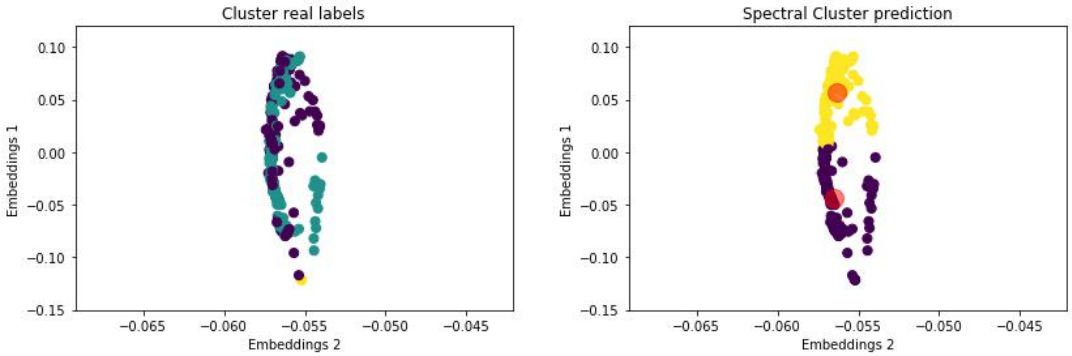

FIGURE 7.3: Embedding distribution for each audio window. In the left the real label and in the right the predicted values are shown.

For the best two models, a majority voting inside each utterance has been made (similar to the emotion recognition system) and than the DER has been computed

again. As well, the accuracy at utterance level (without taking into account the limits between start and end of each utterance) is computed. For the best model (160 ms and 30% overlapping) a DER of **0.535** at utterance level is achieved. Taking into account that the minimum DER is 0.282, the results are very optimal for this example. The accuracy is 65.60%, and it means that the cluster is learning something from the features extracted with the spectral clustering, but are not fully discriminant. In Fig. 7.3, a representation of the real values of the cluster and the segmentation is shown. The real values are not very discriminant, as it was expected, because most of them are sharing the space representation. The prediction has only two embeddings to cluster, so taking into account this information, the cluster works fine. The yellow point at the bottom of the first figure is an outlier (background noise) that was detected from the VAD, so it can be avoided increasing the number of clusters to 3 and cluster this kind of points as *no-talking*.

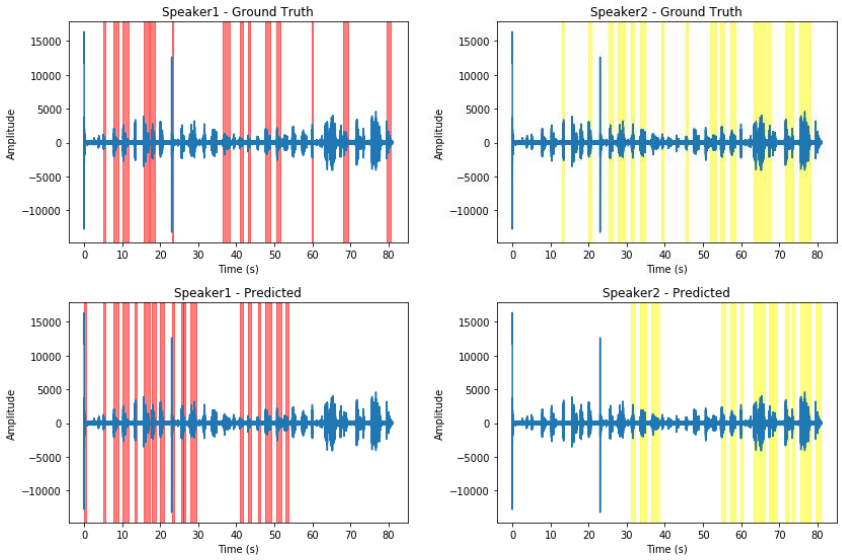

FIGURE 7.4: Final results of the audio segmentation compared with the ground truth.

A representation of the audio speech with the real utterances and the utterances predicted is shown in figure 7.4. This graphics represent the 65.60% of accuracy of the system. It is clear that the first part of the audio has more utterances from speaker 1 and it is recognized by the system. For the last part of the audio, the speaker 2 has more influence and the system can detect it just missing 3 utterances in the last 30 seconds of the audio. Remember that this audio fragment is from a conversation between a male and a female. This helps the system to perform better because some features differ a lot (Pitch, MFFC) when comparing male and female voice. If a conversation between two males or two females is tested, the results would not be so satisfactory. To improve this accuracy for the unsupervised learning, new methods for feature extraction are proposed in future work. In general, this kind of features is difficult to cluster with *K-Means*, so finding a new representation for the data in the future can be a great idea.

7.2.5 Proof of Concept

After the analysis of the experiments using the IEMOCAP database, one last thing to do is to make a Proof of Concept. This Proof of Concept consists of evaluating

the performance of the trained models with some new data from the Face-to-face Dyadic Interaction Dataset. For this purpose, one video of this dataset is used. This video is a conversation between two participants. For the analysis, only one view is taken for each person, named as *P1* and *P2*. For *P1*, 5 utterances are analyzed: one happy, one sad, one angry and two neutral. For *P2*, only 3 utterances: one happy and two neutral.

The first evaluation have been done with the unimodal faces model. For *P1*, three out of five utterances are well predicted. The model have failed in the sad and the angry utterances. This prediction was expected because the sad utterance is not a 100% sad utterance, it could be also a neutral face because the expression is neutral, even though the voice looks sad. In this case, the utterance would be well predicted. The angry utterance is not also a pure angry utterance, it could be detected as a sad or neutral emotion. Although only one video is not representative of the entire dataset, these results are indicating that the model is able to predict well the happy and neutral emotions, when for the IEMOCAP database the neutral emotion was the one with less accuracy. This could be because of the improvement of the image quality in this new database. For *P2*, only one neutral utterance is predicted as happy while the other two are well predicted. The face expression in a happy utterance is slightly exaggerated, so the system works well for this kind of utterances.

The next two analysis are done with the unimodal raw audio and the unimodal handcraft audio features. These experiments perform really bad in comparison with the faces one. Only one utterance is well predicted for each person, a neutral one (three if the sad is considered as neutral and the angry as sad). This results are also expected because there is a lot of difference between the audio signals from this new database and the IEMOCAP audios. The amplitude rank is very high in the IEMOCAP database, so the raw audio and the handcrafted features are going to be very different from the ones extracted from the new database. Another important difference is that this new database have a sampling rate of $44.100Hz$, while in IEMOCAP is $16.000Hz$. The needed down-sampling to evaluate the new database in the proposed models could mean a lost of important temporal and frequency information. In the future, normalizing the audio before entering the network could be worthy for testing the model with different audios.

Chapter 8

Conclusion and future work

8.1 Conclusion

This project has mainly been focused on the development of an real-time **Emotion Recognition System** working with **Deep Learning** architectures and using video and audio sources at frame level in a **dyadic** environment from **IEMOCAP** database, in addition to the **creation and recording of a database** that represents this dyadic environment. On the other hand, to facilitate the annotation for the database, a baseline system of **speaker segmentation** has been developed. Therefore, the conclusions obtained from this project are the following:

- Creating a **face-to-face dyadic database** is important to research and implement new paradigms and technologies of interpersonal behavior understanding. Currently, databases to recognize emotions do not usually contain dyadic conversations between individuals, in addition of having worse image quality and much less diversity and amount of data.
- The performance of the models indicate that the local level analysis is learning some patterns and features from the data, but it is yet far from the context-aware methods. Emotions such as happy and angry, in which there are more changes in the face, are better classified in the face model than sad and neutral. The utterance level analysis using the mean improves in all the cases the performance of frame by frame. This means that, for most of the cases, the frames that are not well classified are not far away from the true label.
- The **VGG architecture** for the face emotion analysis is overfitting with the amount of data that is used for the training. The results obtained are close to the state of the art results, so our network is learning useful features at frame level to predict some emotions correctly, in particular the happy emotion achieves more than 60% of accuracy. An important step in face analysis is the face alignment. For occlusion or non-frontal problems a new face detector is proposed.
- When analyzing the emotions using the audio source, extracting the features directly from a **deep learning model** (53.52%) achieves more accuracy than extracting these features with a **handcrafted model** (49.22%). In particular, the accuracy for sad labels is 80% in both models, but for neutral labels the accuracy goes under 10% of accuracy. As well, the best windows length for the analysis is 100 ms, performing better in most experiments and indicating than, for less window length, it is more difficult to have discriminatory information.

- The **fusion between face and audio** architectures results are far away from the state of the art results. This is because the network is giving more importance to the face features and this features are overfitted. This conclusion is clear looking at the emotion accuracy, which distribution is very similar to the only face analysis distribution. Another conclusion from this experiment is that fusing face with the raw audio and the handcrafted features improves the accuracy, because the features extracted from the deep learning model and the handcrafted are different and compatible.

- A baseline system for speaker segmentation has been created using only audio, using an unsupervised method to avoid annotating the utterance. The best window length for analyzing the audio after computing a grid search is 160 ms with an overlap of 30% between consecutive windows. Using an audio example from IEMOCAP database with no overlapping segments, the accuracy obtained at utterance level is 65.60% and a DER of 0.535. The voice activity detector works fine, but the features extracted from the windows do not have a clear distribution that helps the spectral clustering perform better.

8.2 Future work

This project shows that there is a lot of work and experiments to study. One improvement could be a different **fusion method**. Only the concatenation has been tested, but other methods like a weighted fusion or other architectures of the state-of-the-art literature could improve the performance of the fusion experiments. Other possible fusion method is to give more importance to the raw audio and audio features during the fusion process because they obtain the best performances in the unimodal scenarios. This statement also encourages the idea of fusing this modalities for obtaining even a better classification of the emotions. As well, new techniques could be included in the deep learning architecture, for example using an attention block to give more importance and refine the features.

Another possible experiments to run in a future are related with the layers and the architectures. For example, the VGG is overfitting with the data of the IEMOCAP database and the state-of-the-art has proved that models like **3D-CNN** perform better. Other way to improve the results is to model the **context** and using **memory blocks**, introducing a recurrent model to improve the results.

For the speaker segmentation, new techniques for feature extraction could be tried to facilitate the cluster classify each sample correctly. Using the video as the same time as the audio is something that works and improves the results, exploiting the facial characteristics of the person speaking, instead of only using the audio.

But it is clear that the literature is a little bit stagnant with respect to the databases. The existing ones do not have properly visual or auditory quality, or the participants are actors, or there is only one recorded view for each speaker, etc. Here is the opportunity of achieving a great improvement in the state-of-the-art with the **Face-to-face Dyadic Interaction Dataset**, which is going to provide more visual quality, nearly a 360^{o} view of the recordings, natural behaviour of the participants, a wide range of ages and ethnicities and 81 hours of recordings of dyadic conversations. Exploiting this database is one of the most important future works.

Bibliography

[1] Anurendra Kumar Akshay Kumar. "Unsupervised Speaker Diarization". In: 2012.

[2] Gholamreza Anbarjafari et al. "Review on Emotion Recognition Databases". In: Jan. 2017. DOI: 10.5772/intechopen.72748.

[3] Dario Bertero et al. "Real-Time Speech Emotion and Sentiment Recognition for Interactive Dialogue Systems". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1042–1047. DOI: 10.18653/v1/D16-1110. URL: https://www.aclweb.org/anthology/D16-1110.

[4] Vinay Bettadapura. "Face Expression Recognition and Analysis: The State of the Art". In: *CoRR* abs/1203.6722 (Mar. 2012).

[5] Adrian Bulat and Georgios Tzimiropoulos. "How Far are We from Solving the 2D 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks)". In: Mar. 2017. DOI: 10.1109/ICCV.2017.116.

[6] Carlos Busso et al. "IEMOCAP: Interactive emotional dyadic motion capture database". In: *Language Resources and Evaluation* 42 (Dec. 2008), pp. 335–359. DOI: 10.1007/s10579-008-9076-6.

[7] Teah-Marie Bynion and Matthew Feldner. "Self-Assessment Manikin". In: Jan. 2017, pp. 1–3. DOI: 10.1007/978-3-319-28099-8_77-1.

[8] Erik Cambria et al. "Benchmarking Multimodal Sentiment Analysis". In: (July 2017).

[9] Pawel Cyrta, Tomasz Trzcinski, and Wojciech Stokowiec. "Speaker Diarization using Deep Recurrent Convolutional Neural Networks for Speaker Embeddings". In: *CoRR* abs/1708.02840 (2017). arXiv: 1708.02840. URL: http://arxiv.org/abs/1708.02840.

[10] Abhinav Dhall et al. "Acted Facial Expressions In The Wild Database". In: (Oct. 2011).

[11] Florian Eyben, Martin Wöllmer, and Björn Schuller. "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor". In: Jan. 2010, pp. 1459–1462. DOI: 10.1145/1873951.1874246.

[12] Olivier Galibert. "Methodologies for the evaluation of Speaker Diarization and Automatic Speech Recognition in the presence of overlapping speech". In: Aug. 2013.

[13] Sanaul Haq, Philip JB Jackson, and J Edge. "Speaker-dependent audio-visual emotion recognition." In: *AVSP*. 2009, pp. 53–58.

[14] Devamanyu Hazarika et al. "Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 2122–2132. DOI: 10.18653/v1/N18-1193. URL: https://www.aclweb.org/anthology/N18-1193.

[15] Devamanyu Hazarika et al. "ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 2594–2604. DOI: 10.18653/v1/D18-1280. URL: https://www.aclweb.org/anthology/D18-1280.

[16] Agata Kołakowska et al. "Emotion Recognition and Its Applications". In: *Advances in Intelligent Systems and Computing* 300 (July 2014), pp. 51–62. DOI: 10.1007/978-3-319-08491-6_5.

[17] *Language Development Project. 4.Utterance boundaries-*. http://ldp-uchicago.github.io/docs/guides/transcription/sect_4.html#tg-4-8-1.

[18] Qingjian Lin et al. *LSTM based Similarity Measurement with Spectral Clustering for Speaker Diarization*. July 2019.

[19] P. Lucey et al. "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. 2010, pp. 94–101. DOI: 10.1109/CVPRW.2010.5543262.

[20] Sergej Lugovic, Ivan Dunđer, and Marko Horvat. "Techniques and Applications of Emotion Recognition in Speech". In: May 2016. DOI: 10.1109/MIPRO.2016.7522336.

[21] Alexey Lukin and Jeremy Todd. "Adaptive Time-Frequency Resolution for Analysis and Processing of Audio". In: 4 (Jan. 2012).

[22] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. "Deep Face Recognition". In: vol. 1. Jan. 2015, pp. 41.1–41.12. DOI: 10.5244/C.29.41.

[23] N Majumder et al. "Multimodal Sentiment Analysis using Hierarchical Fusion with Context Modeling". In: June 2018.

[24] Navonil Majumder et al. "DialogueRNN: An Attentive RNN for Emotion Detection in Conversations". In: *CoRR* abs/1811.00405 (2018). arXiv: 1811.00405. URL: http://arxiv.org/abs/1811.00405.

[25] G. McKeown et al. "The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent". In: *IEEE Transactions on Affective Computing* 3.1 (2012), pp. 5–17. ISSN: 1949-3045. DOI: 10.1109/T-AFFC.2011.20.

[26] Alejandro Newell, Kaiyu Yang, and Jia Deng. "Stacked Hourglass Networks for Human Pose Estimation". In: (Mar. 2016).

[27] Michalis Papakostas et al. "Deep Visual Attributes vs. Hand-Crafted Audio Features on Multidomain Speech Emotion Recognition". In: *Computation* 5.2 (2017). ISSN: 2079-3197. DOI: 10.3390/computation5020026. URL: https://www.mdpi.com/2079-3197/5/2/26.

[28] Magda Piórkowska and Monika Wrobel. "Basic Emotions". In: July 2017. DOI: 10.1007/978-3-319-28099-8_495-1.

[29] Soujanya Poria et al. “A review of affective computing: From unimodal analysis to multimodal fusion”. In: *Information Fusion* 37 (2017), pp. 98 –125. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2017.02.003. URL: http://www.sciencedirect.com/science/article/pii/S1566253517300738.

[30] Douglas A. Reynolds et al. “The NIST Speaker Recognition Evaluation - Overview Methodology, Systems, Results, Perspective”. In: *Speech Commun.* 31.2-3 (June 2000), pp. 225–254. ISSN: 0167-6393. DOI: 10.1016/S0167-6393(99)00080-1. URL: http://dx.doi.org/10.1016/S0167-6393(99)00080-1.

[31] Nicu Sebe et al. “Multimodal approaches for emotion recognition: A survey”. In: *Proceedings of SPIE - The International Society for Optical Engineering* 5670 (Dec. 2004). DOI: 10.1117/12.600746.

[32] Kaiyu Shi, Xuan Liu, and Yanmin Qian. “Speech Emotion Recognition Based on SVM and GMM-HMM Hybrid System”. In: 2017.

[33] Vera Shuman, Katja Schlegel, and Klaus Scherer. *Geneva Emotion Wheel Rating Study*. Aug. 2015.

[34] Ingo Siegert et al. “Appropriate emotional labelling of non-acted speech using basic emotions, geneva emotion wheel and self assessment manikins”. In: Aug. 2011, pp. 1 –6. DOI: 10.1109/ICME.2011.6011929.

[35] Rohit Sinha et al. “The Cambridge University March 2005 speaker diarisation system”. In: Jan. 2005, pp. 2437–2440.

[36] A. Solomonoff et al. “Clustering speakers by their voices”. In: *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP ’98 (Cat. No.98CH36181)*. Vol. 2. 1998, 757–760 vol.2. DOI: 10.1109/ICASSP.1998.675375.

[37] Huan Song et al. “Triplet Network with Attention for Speaker Diarization”. In: Sept. 2018, pp. 3608–3612. DOI: 10.21437/Interspeech.2018-2305.

[38] Wei-Ho Tsai, Shih-Sian Cheng, and Hsin-min Wang. “Speaker clustering of speech utterances using a voice characteristic reference space.” In: Jan. 2004.

[39] Yi Yang and Deva Ramanan. “Articulated Human Detection with Flexible Mixtures of Parts”. In: *IEEE transactions on pattern analysis and machine intelligence* 35 (Dec. 2013), pp. 2878–90. DOI: 10.1109/TPAMI.2012.261.

[40] Z. Zeng et al. “A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.1 (2009), pp. 39–58. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2008.52.

[41] Aonan Zhang et al. *Fully Supervised Speaker Diarization*. Oct. 2018.