Automatic evaluation of Alzheimer's disease, a multimodal analysis of spontaneous conversations

by

Arlen PÉREZ ARANA

THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLEMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
Ph. D.

MONTREAL, MARCH 14, 2022

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mrs. Sylvie Ratté, Thesis Supervisor
Department of Software and IT Engineering, École de technologie supérieure

Mr. Luc Duong, Thesis Co-supervisor
Department of Software and IT Engineering, École de technologie supérieure

Mr. Maarouf Saad, President of the Board of Examiners
Department of Electric Engineering, École de technologie supérieure

Mr. Carlos Vázquez, Member of the jury
Department of Software and IT Engineering, École de technologie supérieure

Mr. Pierrich Plusquellec, External member of the jury

THIS THESIS WAS PRENSENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND PUBLIC

MARCH, 2022

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGMENTS

There are several persons and institutions to thank in this thesis.

First, to my supervisors Sylvie Ratté and Luc Duong. Thanks, professor Luc for been always there for me. You have been very supportive, but above all, your kindness always makes me feel confident about sharing my ideas with you and discussing the direction of our project.

A special thanks to Sylvie Ratté. Not just professionally, but as a friend and confidant and as a person. Thank you for making me part of this project and especially, thank you for being part of my life. She supports me in all the ways a person could support others. For that, thank you, Sylvie, you will always have my respect, admiration, and love.

This project could not have been possible without the financial support of CONACyT (*Comisión Nacional de Ciencia y Tecnología*) from Mexico City, the “Ministère des Relations internationals et de la Francophonie Quebec-CONACYT”, as well as by NSERC (the Natural Sciences and Engineering Research Council of Canada). These institutions supported my studies and allowed me to participate in this project. Thank you for affording and supporting the Mexican talent, and bringing it to the rest of the world.

Thank you to all my friends, there are a lot of names to mention here. Thanks to Remi, Fariba Gerardo, Neda, Helmi, Edgar, Frederik, Athefe, and Eric for answering my questions and have helped me when I requested. Additionally, thank you to Karen, Geoffrey, and Erika for always believing in me and encouraging me through the process of this thesis. Thanks to Ramat for all your support, you have helped me on each step, without you reaching this point would have being much more difficult, thanks for your dedication and effort.

A special thanks to Laura, you were the first reason I started this project, you suggested and encourage me to do this project so a special thanks to you.

Finally, I would like to thank my family. Thanks, Evan and Alexa, my son, and daughter, for being my inspiration, my impulse every day, I will always be there for you. Hopefully, one day you will include me in your acknowledgments too. A special thanks to my parents, Oscar and Teresa, without them, it would not have been possible to reach success in my life. They worked arduously all my life to give me all the tools for life. There are not enough words to describe my admiration and love for them. They have been always there for me on every success and every problem throughout my life. Thank you, father, for leading me in all my decisions, thanks for suggested me to choose software engineering as my profession, you were right, I love it.

Évaluation automatique de la maladie d’Alzheimer, une analyse multimodale de conversations spontanées

Arlen PÉREZ ARANA

RÉSUMÉ

Les patients atteints de la maladie d'Alzheimer (MA) présentent des difficultés de communication verbale et non verbale, ce qui a suscité un intérêt croissant pour le rôle que joue la communication non verbale dans la vie des personnes atteintes de démence (Rousseaux et al., 2010). On estime que 55 à 97 % du message communiqué lors d'une interaction entre adultes consiste en un comportement non verbal ((Gross, 1990), (Hargie et al., 1981)), qui comprend les mouvements du corps, les expressions faciales (FE), le toucher, l'apparence physique, l'espace personnel et les caractéristiques de la communication vocale telles que la hauteur, l'intonation et le débit de parole.

En conséquence de ce qui précède, certaines études liées à la maladie d'Alzheimer (MA) ont étudié la communication verbale et non verbale, par exemple les mouvements oculaires, les expressions faciales, le débit de la parole, la communication vocale et l'analyse des sentiments pendant l'exécution de certaines tâches. Selon ces études, les expressions faciales et les caractéristiques acoustiques des patients atteints de la MA pourraient suggérer certaines caractéristiques dans les premiers stades de la MA qui peuvent être analysées automatiquement.

Dans cette thèse, nous introduisons une méthode d'analyse automatique et d'évaluation de la corrélation entre le verbal et le non-verbal et la MA lors de conversations naturelles enregistrées sur vidéo. Notre objectif est de classer automatiquement les sujets atteints de la maladie d'Alzheimer ou les témoins sains (HC) par le biais des expressions faciales.

Nous analysons 23 conversations, d'une durée moyenne de 16 minutes. Pour l'analyse faciale, nous avons suivi 3 groupes de caractéristiques : le regard et les repères oculaires, les repères faciaux et les unités d'action faciale (UAF). De plus, pour l'analyse verbale, nous avons obtenu deux groupes de caractéristiques: les silences et les caractéristiques phonétiques (13 caractéristiques de coefficients cepstraux de fréquence Mel (MFCC)).

En général, nous avons utilisé quatre classificateurs pour distinguer la MA de l'HC: Random Forest Classifier (RFC), K Nearest Neighbor (KNN), Support Vector Machines (SVM) et Naïve Bayes (NB).

En ce qui concerne l'analyse des silences et des caractéristiques phonétiques, la meilleure performance obtenue était une précision de 81 % et un taux de sensibilité-spécificité de 91 % sur la courbe ROC (Receiver Operating Characteristic). De même, l'analyse multimodale a montré une précision de 90% et une courbe ROC de 93%. Ces résultats ont été obtenus avec le classificateur KNN entraîné avec toutes les caractéristiques (verbales et non verbales).

Notamment, le RFC a montré la meilleure performance dans toutes les expériences réalisées dans cette étude, en entraînant l'algorithme exclusivement avec les expressions faciales et les caractéristiques du regard, nous avons obtenu une précision de 93% et une courbe ROC de 98%. La précision de classification pour le classificateur KNN était de 91 %, la sensibilité-spécificité était de 97,7 %, entraîné avec les repères faciaux comme caractéristiques. Ces résultats présentent une meilleure performance lorsque le classificateur est entraîné avec les points de repère faciaux que lorsqu'il est entraîné avec toutes les caractéristiques.

En conclusion, les expressions faciales et les caractéristiques phonétiques pendant la conversation pourraient fournir des signes de la MA à un stade précoce.

Dans ce travail, nous avons présenté une méthodologie permettant de distinguer la MA de l'HC. L'objectif principal de cette méthode est d'utiliser un moyen non invasif d'analyse et de classification par l'enregistrement de conversations naturelles. Par conséquent, nous pouvons fournir aux cliniciens un outil non invasif et automatique pour la détection précoce des signes de la MA.

Mots-clés: Classification automatique, conversations naturelles, expressions faciales, phonétique, silences, maladie d'Alzheimer (MA), analyse multimodale.

Automatic evaluation of Alzheimer's disease, a multimodal analysis of spontaneous conversations

Arlen PÉREZ ARANA

ABSTRACT

Alzheimer's disease (AD) patients present verbal and nonverbal communication difficulties, which has led to growing interest in the role nonverbal communication plays in the lives of people with dementia (Rousseaux et al., 2010). It is estimated that 55-97% of the message communicated in adult interaction consists of nonverbal behavior ((Gross, 1990), (Hargie et al., 1981)), which includes body movement, facial expressions (FE), touch, physical appearance, personal space, and vocal communication features such as pitch, intonation, and speech rate.

As a result of the above, there are some studies related to Alzheimer's Disease (AD), where verbal and nonverbal communication has been studied, some examples are eye movements, facial expressions, speech rate, vocal communication, and sentiment analysis during performing some tasks. According to these studies, facial expressions and acoustic features of AD patients could suggest certain characteristics in early stages of AD that can be automatically analyzed.

In this thesis, we introduce a method to automatically analyze and evaluate the correlation between verbal and nonverbal and AD during video recorded natural conversations. Our objective is to automatically classify AD subjects or Healthy Controls (HC) through facial expressions features.

We analyze 23 conversations, with an average duration of 16 minutes. For the purpose of the facial analysis, we tracked 3 groups of features: eye gaze and landmarks, face landmarks, and Facial Action Units (FAU). Additionally, for the purpose of the verbal, analysis we obtained 2 groups of features: silences and phonetic features (13 Mel-Frequency Cepstral Coefficients (MFCC) features).

In general, we used four classifiers to discern between AD and HC: Random Forest Classifier (RFC), K Nearest Neighbor (KNN), Support Vector Machines (SVM), and Naïve Bayes (NB).

Regarding the analysis with the silences and phonetic features, the best performance obtained was 81% accuracy and 91% of sensitivity-specificity rate 'Receiver Operating Characteristic' (ROC) curve. Likewise, the multimodal analysis showed 90% accuracy and 93% ROC curve. These results were obtained with the KNN classifier trained with all the features (verbal and nonverbal).

Notably, the RFC showed the best performance in all the experiments performed in this study, training the algorithm exclusively using facial expressions and gaze features, we obtained a 93% accuracy and 98% ROC. The classification accuracy for the KNN classifier was 91%, the sensitivity-specificity was 97.7%, trained with facial landmarks as features. These results present a better performance trained with facial landmarks in comparison with training the classifier with all the features.

In conclusion, facial expressions and phonetic features while speaking could provide signs of AD in early stages.

In this work, we presented a methodology for discriminating between AD and HC. The principal objective while using this method is to use a non-invasive way to analyze and do classification through recording natural conversations. Consequently, we can provide clinicians a non-invasive and automatic tool for the early detection of signs of AD.

Keywords: Automatic classification, natural conversations, facial expressions, phonetic, silences, Alzheimer’s disease (AD), multimodal analysis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABREVIATIONS

AAIC	Alzheimer Association International Conference
Ab. Av.	Absolute average
AD	Alzheimer’s disease
ADNI	Alzheimer’s disease Neuroimaging Initiative
AS	Antisaccade
AUC	Area Under the Curve
Av.	Average
CCC	Carolinas Conversation Collection
CNN-LSTM	Long-short term memory neural network
Corr.	Correlation
eGeMAPS	Geneva Minimalistic Acoustic Parameter
EMFACS	EMotional Facial Action Coding System
EMG	Electromyographic
Exp.	Experiment
FACS	Facial Action Coding System
FAU	Facial Action Unit
FE	Facial Expressions
FER	Facial Emotion Recognition
FN	False Negative
FP	False Positive
HC	Healthy Controls

IRB	Institutional Review Board
KNN	Nearest Neighbor
LDA	Linear Discriminant Analysis
LFPW	Labeled Face Parts in the Wild
LiNCS	Laboratoire d'ingénierie Cognitive et Sémantique
LSTM	Long-short term memory
MCI	Mild Cognitive Impairment
MFCC	Mel-Frequency Cepstral Coefficients
MLP	Multi Layer Perceptron
MRCG	Multi-Resolution Cochleagram
MRI	Magnetic Resonance Imaging
NLP	Natural Language Processing
pBLSTM	pyramidal Bidirectional LSTM
PDM	Point Distribution Model
PET	Positron Emission Tomography
PreS	Presaccade
PS	Prosaccade
RFC	Random Forest Classifier
RMS	Root Mean Square amplitude
ROC	Receiver Operating Characteristic
SEM	Saccade Eye Movements
SVM	Support Vector Machines
tDCS	Transcranial Direct-Current Stimulation

TN	True Negative
TP	True Positive
VRM	Visual Recognition Memory

LIST OF SYMBOLS

TIME UNIT

ms	milisecond

FREQUENCY

Hz	Hertz

INTRODUCTION

Alzheimer’s disease (AD) patients present verbal and nonverbal communication difficulties, which has led to growing interest in the role nonverbal communication plays in the lives of people with dementia (Rousseaux et al., 2010). It is estimated that 55-97% of the message communicated in adult interaction consists of nonverbal behavior ((Gross, 1990), (Hargie et al., 1981)), which includes body movement, Facial Expressions (FE), touch, physical appearance, personal space, and vocal communication features such as pitch, intonation, and speech rate (see Buck & VanLear (2002), Berger (2014), for an overview).

Many tests have been proposed for the analysis of AD in early stages. Some are based on medical data relying on biospecimens or imaging techniques. These types of studies evaluating the physical impacts of AD can be invasive and are not appropriate for monitoring the evolution of the disease. Some other studies evaluate the cognitive conditions of patients and are precious tools for clinicians to gauge the advances of the disease. Recently, researchers have focused on the analysis of verbal and nonverbal communication in patients with some type of dementia. Several researches have studied the language and communication process during dementia, most of them have shown that the language disturbances may appear in the beginning of the disease and these disturbances become more substantial as the disease progress. Likewise, this decrement seems to be associated to cognitive impairment which can be study through verbal fluency, lexical, and discourse analysis (Caramelli et al., 1998).

In the context of communication, current research on cognitive impairment and dementia tends to be limited to the investigation of two channels:

First, some authors have focused on manual evaluations while subjects perform some cognitive tasks. These types of tests that relied on nonverbal communication have become promising in the latest years. One example is the study of Gill Hubbard (2002), where researchers describe the subjects' behavior in the context of social interactions. Other researchers have applied both automatic evaluation techniques and manual evaluation in the context of cognitive tests. In these studies, subjects performed some cognitive tasks, one remarkable example is the picture description task, some examples are: Hernández-Domínguez et al. (2018), Zargarbashi & Babaali (2019), Tomoeda & Bayles (1993), Vuorinen et al. (2000), Fraser et al. (2015), Mueller et al. (2016), Sajjadi et al. (2012), and Bschor et al. (2001).

An interesting approach is the study of Parra et al. (2019). They implemented the "Visual Short-Term Memory Binding Test," this test consists of pictures with two or three black shapes (shape only condition) or shapes with color (shape-color binding condition). These pictures are shown to participants. They are required to remember the pictures. The authors concluded that people with Mild Cognitive Impairment (MCI) presented selective binding impairments. Further analysis of MCI patients suggests that this binding impairment is a factor of prodromal AD.

Specifically, several studies exist in the context of speech analysis and the relationship with AD. Likewise, several approaches have been implemented as attempts to detecting the AD in early stages.

One of the attempts was done in the work of López-de-Ipiña et al. (2013), authors in this work studied the spontaneous speech fluency based on the duration of the dialog, the time domain (short time energy) and the frequency domain (spectral centroid). Also, they studied emotion in the speech through acoustic features like pitch, intensity, and Root Mean Square amplitude (RMS) between others. Additionally, they extracted the emotional temperature by means of prosodic and paralinguistic features related to the pitch and energy. Their results showed a 94.6% accuracy with all the features extracted used in the training of the Multi Layer Perceptron (MLP) neural network.

One of the most recent studies and relevant studies is the one held by Yeung et al. (2021). They selected 5 clinicians to select clinically relevant speech and language features to analyze. They chose word-finding difficulty, incoherence, perseveration, and errors in speech. They recorded, transcribed, and annotated the dialog obtained from the "Cookie Theft" picture description.

Later, they extracted Natural Language Processing (NLP) variables including lexical, semantic, and syntactic features. They extracted also acoustic features like properties of the sound wave, speech rate, and the number of pauses. These features were extracted through the Mel-Frequency Cepstral Coefficients (MFCC).

The MFCC is a short representation of the spectrum of an audio signal, that is to say, it is the digital representation of the waveform of a sound (Zheng et al., 2001). These coefficients are the results of the cosine transform of the real logarithm of the short-term energy spectrum denotating on a mel-frequency scale, i. e., expressed in the discrete Fourier transform. The extraction of the MFCC consist on the following steps:

1. To obtain the discrete Fourier transform from the audio signal. Thus, the short-term power spectrum is obtained in the frequency domain *P(f)*,
2. *P(f)* is warped along its frequency axis in Hz into the mel-frequency axis as *P(M)*, where M is the mel-frequency,
3. *P(M)* is then convolved with the triangular band-bass filter, P(M) to $\boldsymbol{\theta(M)}$**,**
4. *Finally, the MFCC are computed using (0.1):*

$$MFCC(d) = \sum_{k=1}^{K} X_k \cos\left[d(k-0.5)\frac{\pi}{K}\right], \qquad d = 1, \dots, D \tag{0.1}$$

$$X_k = \ln(\theta(M_k)) \tag{0.2}$$

Where:

K, number of filters in the implementation of the discrete convolution,

X, defined by equation 0.2,

$\theta(M_k)$, the power spectrum convolved in step 3.

Researchers observed that the best correlations obtained were the word-finding difficulty (corr. = 0.92), the incoherence (corr. = 0.91), and perseveration (corr. = 0.88) in MCI and AD. According to their results, researchers in this study sustain that the early identification of these markers could be a useful help for clinicians in the task of distinguishing AD from healthy people.

Although verbal and nonverbal communication analysis seems to be a good technique to address the problem of assessing AD in the early stages, it remains relatively unexplored (Khan et al., 2021). In this context, this thesis presents a feasibility study that aims to perform an automatic analysis of videos involving older adults, to characterize facial expressions, eye movements, silences, and phonetic features and; consequently, attempts to automatically differentiate between AD patients and Healthy Controls (HC) by simply analyzing verbal and nonverbal communication of elderly and patients with Alzheimer's. In other words, this work aims to implement an economical and non-invasive technique for facial expression, eye movement and phonetic characterization to classify AD patients and HC automatically.

CHAPTER 1

LITERATURE REVIEW

1.1 Studying the relationship between cognitive impairment and facial expressions

There is little research about the characterization of facial expressions. The problem with studying facial expressions was to find an excellent technique to measure each movement occurring on the face. Over the years, several techniques have been implemented to measure facial behavior. According to Ekman (2002), some authors who attempted first to describe and measure facial expressions were Landis in 1924, Frois-Wittmann in 1930, Fulcher in 1942, and Thompson in 1941. Later, in 1971 the first attempt to formally describe facial expressions was made in the study of P. Ekman and W. V Friesen (1971).

In this study, the analysis of facial expressions focuses on the study of emotions and the expression related to them. This study leads authors to the necessity of having a system that could measure each movement as a unit and then have a tool to measure facial expressions not just for analyzing emotions but rather to use it in any type of approach related to facial movements. Consequently, Ekman et al. created the "Facial Action Coding System" (FACS) (Ekman & Friesen, 1976). In the following section, the FACS and Facial Action Units (FAUs) are described.

1.1.1 Facial Action Units (FAUs)

FAUs were first described in the FACS, a manual composed of 44 basic FAUs (Ekman & Friesen, 1978). This manual was subsequently updated in 2002 (Ekman, 2002). FAUs are basic expression construction units. They represent minimal facial actions, and each one can involve several movements of facial muscles. At the same time, several muscles can be involved in more than one expression. Despite no consensus in the research community about facial expressions characterization, the FACS is widely used in most studies on facial expressions analysis and measuring facial behavior.

FAUs are divided into two groups: “Upper face” and "Lower face," with the upper face group affecting the eyes, the eyebrows, the forehead, and the eyelids, see Table 1.1. The FAUs in the lower face are composed of five groups: up/down, horizontal, oblique, orbital, and miscellaneous motions, see Table 1.2. Additionally, in Figure 1.1, we show some examples of FAUs detected in this study.

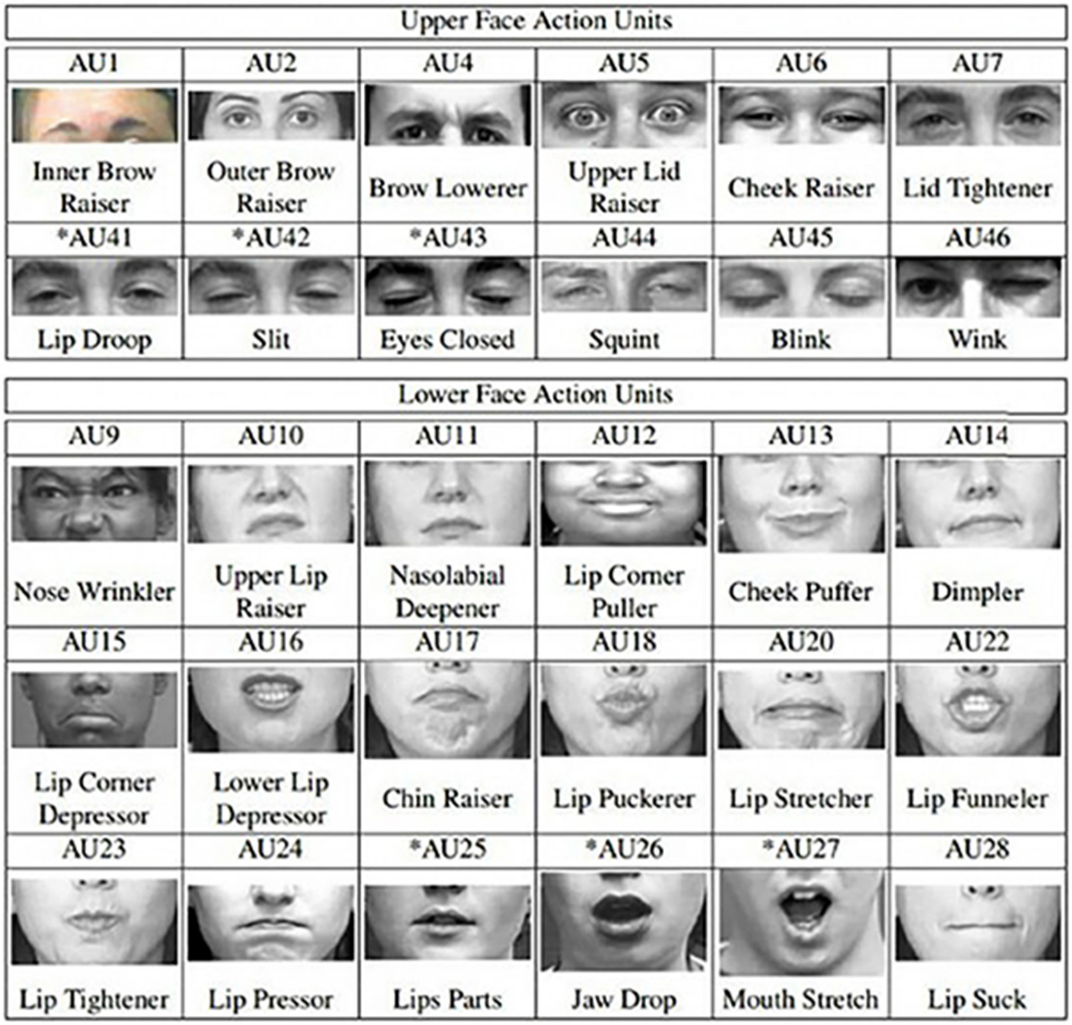

Figure 1.1 Some examples of FAUs. Upper and lower face action units. Copyright permission is obtained from De la Torre (2015, page 4)

Table 1.1 Description of FAUs in the upper face

FAU number	Description
1	Inner brow raiser
2	Outer brow raiser
4	Brow lowered
5	Upper lid raiser
6	Cheek raiser and lid compressor
7	Lid tightener
43	Eye closure
45	Blink
46	Wink

Table 1.2 Description of FAUs in the lower face

FAU number	Direction of the action	Description
9	Up/down	Nose wrinkle
10		Upper lip raiser
15		Lip corner depressor
16		Lower lip depressor
17		Chin raiser
14	Horizontal	Dimpler
20		Lip stretcher
11	Oblique	Nasolabial furrow deepener
12		Lip corner puller
13		Sharp lip puller
18	Orbital	Lip pucker
22		Lip funneler
23		Lip tightener
24		Lip presser
28		Lip suck

We used the FACS in this study because it is a complete and comprehensive system that helps researchers distinguish several visible facial movements. Consequently, it is a valuable tool for studies related to facial expressions. Despite there being no standard or consensus about the tools used for facial expressions characterization, FACS is used in a considerable number of recent studies. It is an adequate technique for this study.

1.1.2 Facial expressions and cognitive impairment

Several studies have been carried out about the analysis of facial expressions and their relation with cognitive impairment. In general, two lines can be addressed in this matter. First, the analysis of the performance of patients with cognitive impairment recognizing emotions in facial expressions, generally from pictures; second, the automatic or manual analysis of facial expressions from patients with cognitive impairment while performing some cognitive task.

In the case of the studies about the analysis of subjects' performance in Facial Emotion Recognition (FER), we can mention the work of Fernández-Ríos (2021). In this study, the authors reviewed relevant literature related to emotion recognition in AD. Authors conclude AD patients present a deficit in FER compared with healthy older adults. Except for the recognition of joy, AD patients tend to preserve this ability despite the pathology.

They evaluated the type of emotion (joy, sadness, anger, disgust, fear, surprise, and neutral). In most articles, the evaluation of FER was based on the identification of these basic emotions. Authors found many articles comparing AD with other diseases, mainly frontotemporal dementia (see Table 1.3). Likewise, there were three studies including groups with MCI to compare.

Regarding the authors' instruments in their experiments, the most used instrument was the "Pictures of facial affect" created by Ekman and Friesen in 1976. These pictures represent the six basic emotions and neutral expressions. In some other studies, the authors used the "Facial Expressions of Emotions- Stimuli and Test," this test is composed of two sets, "The Ekman 60 Faces Test" and "The Emotion Hexagon Test," for details refer to https://www.paulekman.com/.

In addition, some authors implemented assessments with neuroimaging, biospecimens samples, and/or electrodermal conductance response. These assessments were performed as support of the results obtained with FER testing.

Table 1.3 Relevant studies where neurodegenerative diseases were compared with AD

Disease compared with AD	Studies
Frontotemporal dementia	Bertoux et al. (2015); Chiu et al. (2016); Fernandez-Duque & Black (2005); Freedman et al. (2013); Hsieh et al. (2013); Kipps et al. (2009); Kumfor et al. (2014); Narme et al. (2013); Narme et al. (2017); Park et al. (2017).
Vascular dementia	Shimokawa et al. (2000).
Semantic dementia	Hsieh et al. (2012).
Progressive non-fluent aphasia	Kumfor et al. (2014); Kumfor et al. (2016).
Parkinson's Disease	Narme et al. (2017); Martinez et al. (2018).
Lewy Body Dementia	Heitz et al. (2016).

Contrasting findings are obtained in these studies. Some authors suggest the presence of deficit in FER in AD patients Bertoux et al. (2015); Dourado et al. (2019); Henry et al. (2008); Kohler et al. (2005); Kumfor et al. (2014); Martinez et al. (2018); Park et al. (2017); Sapey-Triomphe et al. (2015); Sheardova et al. (2014); Shimokawa et al. (2000); Weiss et al. (2008). Other authors found deficits in the performance of AD patients compared with the younger elderly control group; on the contrary, no differences were found compared with older adults. Regarding the age variable, as expected, their review revealed that the older subjects, the worst the recognition in facial expressions.

Regarding the gender variable, the authors pointed out the limited research about gender as a variable that can modulate the results of FER analysis. Most of the studies analyze gender as a sociodemographic variable. Moreover, some studies suggest gender and other variables as modulators of emotional processes (Demenescu et al., 2014) and (Lichtenstein-Vidne et al., 2017). According to those mentioned above, it is suggested to study the gender variable in further studies.

Moreover, other relevant research on emotion recognition is the work of González-Alcaide (2021). Similar to Fernández-Ríos, authors in this study perform a survey about FER in dementia.

They analyzed 345 documents. Unlike Fernández-Ríos, they used clusters to group documents. The first analysis made was the co-citation analysis which yielded 4 clusters: frontotemporal dementia with the highest degree of citation (35 documents), autism spectrum disorders with the oldest cited references (32 documents), AD (31 documents), and Parkinson's and Huntington's disease (26 documents). The second approach was the analysis of bibliographic coupling. Authors obtained three clusters: first, studies focus on Theory of Mind (ToM) and social cognition. The principal diseases studied are Huntington's, Parkinson's, and Alzheimer's diseases. The principal subjects include empathy and perception of emotions. Second, a cluster of frontotemporal dementia, social cognition, and ToM. Similar to the first cluster, most of the studies focus on empathy and emotion recognition. The third cluster address MCI as a precursor to dementia.

Likewise, there is a significant amount of studies about the relationship between ToM and AD. The ToM is the capacity of a subject to construct a state of mind of others and of itself. This process might be manifested in facial expressions. The interpretation of emotions in other subjects might elicit changes in AD patients' facial expressions. Some ToM tests have been performed to analyze the ability of AD patients to infer experience in others. Literature has suggested that ToM is not significantly impaired in AD and that any manifestation of impairment seems to be due to the impairment in other cognitive abilities (Freedman et al., 2013), (Gregory et al., 2002), (Zaitchik et al., 2004), (Zaitchik et al., 2006).

Some other relevant studies about the study of patients and the perception of emotions are Davidson & Irwin (1999), Allender & Kaszniak (1989), (Keane et al., 2002), (Shimokawa et al., 2003), (Schreiner et al., 2005), (Maki et al., 2013), (Torres et al., 2015), (Dourado et al., 2019), (Torres Mendonça De Melo Fádel et al., 2019), (Jiskoot et al., 2021), (P. Yang et al., 2016), and (Hayashi et al., 2021), among others.

One important point to highlight, is that the perception of emotion in other individuals should be carefully distinguished from the production of emotions. As is described above, several studies use facial expressions as stimuli to elicit emotions. In contrast, other studies focus on analyzing facial expression and the relationship with different types of dementia (Davidson & Irwin, 1999). Specifically, there are studies where the facial expressions of AD patients are assessed. In the following paragraphs, we will describe some of the most representative studies.

In the work of U. Seidl, U. Lueken, P. A. Thomann, A. Kruse, J. Schröder, and J. Schro (2012), an investigation is carried out about cognitive deficits' impact neuropsychiatric symptoms on facial expressions in AD, with a specific focus on apathy. The authors explore the determinants of emotional facial expressions in AD and the impact of cognitive and neuropsychiatric symptoms considered potential moderating variables. A neutral picture is shown to subjects, followed by emotional ones, and the facial expressions of each subject are recorded.

A trained psychologist analyzed the videos on the EMotional Facial Action Coding System (EMFACS). The EMFACS measures the most common emotions: anger, fear, distress and or sadness, disgust or contempt, surprise, and happiness. Each emotion consists of a group of FACs that describes each of the emotions already mentioned.

The authors then rated the frequency of each FAU in each video. To assess the expressivity in each participant, the authors calculated the mean frequencies of all 27 FAUs in each picture, as well as the average over all the pictures.

The results showed that cognitive deficits are associated with the loss of specific facial expressions, indicating a progressive loss of control of facial expressions and corroborating other studies. Apathy is a frequent symptom in AD and increases as the AD increases as well. Furthermore, in a study held by K. Burton and A. Kaszniak (2006), the authors demonstrate that specific pathways in the limbic system mediate neural substrates of voluntary and spontaneous emotional expressions. Likewise, these pathways become increasingly disrupted as dementia progresses. This is especially true in the cases of AD and Vascular Dementia, where there is a resulting relative loss of voluntary facial control.

G. Hubbard, A. Cook, S. Tester, and M. Downs (2002) followed a different track. They first observed the nonverbal communicative behavior of older adults with dementia. This behavior was observed in normal daily activities. After the observations, researchers compared their observations with the self-interpreted descriptions from each subject and the center staff interpretation.

Researchers in this study adopted a qualitative approach to nonverbal communication of older adults with dementia. They described their observations, observations from the center's staff, and interpretations from the subjects. No automatic nor quantitative evaluation was done in this study.

Additionally, in the study of V. Bevilacqua (2011), the authors propose a tool to help detect some dementias or neurological disorders by monitoring facial expressions. The authors state that the process of affect is supported by neural systems, specifically the right hemisphere. The authors conclude that there is a direct relationship between neurological distress and the expressivity of some emotions. For example, depression shows an increase in the expressivity of sadness and anger. The proposed tool aims to recognize facial expressions and then state if they represent positive or negative attitudes.

Other researchers who have studied facial expressions and cognitive impairment include the studies of A. S. Schreiner, E. Yamamoto, and H. Shiotani (2005a), and Y. Liu, Z. Wang, and G. Yu (2021).

In conclusion, according to several studies, it is suggested that there is a link between the damage to the orbitofrontal cortex and motivational behavior. The lateral prefrontal cortex is involved in integrating emotion, motivation, and cognition (Watanabe, 2016). Other Magnetic

Resonance Imaging (MRI) studies showed that damage in the amygdala and hypothalamus are involved in emotional body expression (Bachmann et al., 2018).

Following this line, our hypothesis of the feasibility study is that facial expressions in AD patients can differ from facial expressions of healthy older adults. In this manner, the analysis of facial expressions in AD patients became the main objective of this study. It is essential to highlight that this study aims to analyze facial expression as a whole face behavior. Without labeling the expression, we do not determine emotions. Instead, we analyze the movement of muscles on the face through time.

1.2 Studies related to eye movements and cognitive impairment

In addition to the facial movements' analysis, several authors have found interesting to study the eye movements related to cognitive impairment, some other focused specifically on AD. Different approaches have been proposed in this subject. In the following paragraphs we will describe some of the most representative studies.

In the work of D. Lagun, C. Manzanares, S. M. Zola, E. A. Buffalo, and E. Agichtein (2011), the authors used an infrared eye-tracker to record subjects' eye movements, a chinrest was used to fix the head of subjects in front of the screen. Authors used the VPC technique, where subjects are allowed to look at two identical pictures; later, the old visual stimulus and a novel stimulus are shown to the subjects. Features in this study were: the fixation duration of the gaze, the number of re-fixations, and the saccade orientation. The authors used Naïve Bayes (NB), Logistic Regression, and Support Vector Machines (SVM) classifiers with these features.

SVM showed an accuracy of 0.869, a sensitivity of 0.967, a specificity of 0.772, and an Area Under the Curve (AUC) of 0.869. Results indicate that machine learning techniques can help to automatic classification using eye movements features.

In the study of G. Fernández, P. Mandolesi, N. P. Rotstein, O. Colombo, O. Agamennoni, and L. E. Politi (2013), authors state that deficits in memory, attention processes, and inhibitory control may be found by analyzing eye movement patterns.

Later, in 2017, Noiret et al. (2018) held experiments analyzing the Saccade Eye Movements (SEM). Participants' eyes movements were recorded with an eye-tracking system. They compared 20 AD patients with 35 older adults in prosaccade (PS), antisaccade (AS), and presaccade (PreS) tasks.

Using the results of the PS, AS, and PreS tasks, the authors provided evidence of abnormal SEM in AD patients compared with healthy older adults. AD patients demonstrated longer latencies in all tasks. They also showed more significant latency variability in comparison with control subjects.

Several studies have examined the relationship between eye movements and the analysis of dementia, such as AD, in its early stages. Some other studies related to cognitive impairment and eye movements include (Lagun et al., 2011), (De Santi et al., 2011) , (Armstrong, 2008), (Dongheng Li et al., 2005), (Rizzo et al., 2000) and (Armstrong, 2009).

According to the studies described above, eye movements were related to cognitive impairment, specifically to AD. Studies have demonstrated that patients with AD present different eye movement patterns from HC during cognitive tasks.

1.3 Pauses and silences

Pauses and silences in the context of the speech have been studied since ancient times. In the following paragraphs, we will give a brief introduction to pauses in the context of speech, and the relationship with cognitive impairment.

1.3.1 Pauses and silences in the context of speech

It seems that the main problem with silence is to know the meaning depending on the context where it occurs, e. g., the meaning of silence while answering a question could not be the same as the silence evoked while a sentence is finished.

Following this line, Kurzon (1998) attempted to analyze and interpret silences. Authors in this study, state that silences have a semiotic relation with speech. They describe the silences in two lines: a lack of communication where no communicative activity is sighted and, second, in non-verbal communication, where the body language, gestures, and eye movements are meaningful and represent a way of communicating.

How long should a silence last to be meaningful during speech?

Goldman and Eisler (1958) in their study attempt to answer this question. In their study, they recommended the threshold for a silence as 250 ms, silences with less than 250 ms duration are considered pauses for breathing or articulation. Consequently, the authors treated these pauses out of the interest of the language process.

On the other hand, pauses over 250 ms duration are considered as the hesitation of preparation for cognitive processes like sentence planning or retrieving words. Additionally, other authors have adopted the same threshold in their experiments.

On the other hand, recent studies questioned the implementation of a fixed and pre-established threshold to discern between a significant silence and not significant silence in the context of the speech, i. e., Campione & Véronis (2002) performed experiments avoiding setting a predefined threshold, they argued that despising the longest duration values and/or the smallest duration values can lead to disregard meaningful information related to speech and communication skills.

Additionally, authors have mentioned in their studies that the threshold differentiating long and short pauses vary depending on some demographic aspects like age, gender, level of education, among others. Consequently, any pauses despite the duration should be included in the study if a complete dataset is desired.

1.3.2 Studies related to pauses and cognitive impairment

Pauses in the context of speech were studied by several researchers. In the following paragraphs, we will describe some of the most remarkable studies.

The authors of Angelopoulou et al. (2018) analyze the correlation between aphasia and silences during speech. Aphasia is a medical condition where impairment of language exists, commonly this impairment occurs due to an injury to the brain evoked from a stroke. Additionally, aphasia could appear due to a brain tumor or some infections in the brain (National Aphasia Association, 2021).
Authors argue in their work that silences and pauses during speech exhibit an association to cognitive processes, such as word access, word production, selection and retrieving of words, planning, and memory.

Subsequently, they indicate pauses in speech can provide information about the language processes necessary for communication. Pauses can be seen in two ways; first, short pauses for breading and articulate and, second, long pauses express internal cognitive processes while speaking.

Specifically, long pauses manifest more frequently when a selection of words occurs: lemma access process. Also, pauses tend to occur before less frequent words, in consequence, it seems long pauses occur as a consequence of difficulties for producing words (word retrieval and encoding) (Beattie & Butterworth, 1979).

According to the above, the authors of (Angelopoulou et al., 2018) asses speech recording from eighteen patients with aphasia and healthy controls. All the injuries of patients were on the left hemisphere of the brain. Patients were told to talk about their experience of the stroke, while healthy participants described the stroke's story of the patients.

Subsequently, they measure the number of pauses, number of short and long pauses, number of utterances per 100 words, among others. The objective of their work is to determine if a correlation between lesion score and number of pauses per hundred words exists. Thus, they obtained a Spearman rho correlation. As result, they found a higher frequency for short pauses in the patients group compared to controls, while the opposite occurs for the long pauses frequency.

In addition, they found that the log-normal bimodal pattern allowed them to classify between impaired patients and controls. Also, impaired patients showed a higher median for long pauses compared to healthy controls.

As conclusion, the authors in this study state that there are quantitative differences between aphasia patients and healthy controls, such as pause rate and duration, standing out the long pauses where they are considered as a consequence of the process of sentence planning and retrieving words: common cognitive processes.

Authors of Pistono et al., (2019) analyze the pause frequency while picture-based narrative in AD patients and HC. They selected the AD participants using brain MRI, they excluded patients with hyperintensities in the white matter. Also, they analyzed the amyloid from the cerebrospinal fluid from patients.

Participants, AD patients, and HC were told to describe a story given 5 sequential pictures. The description of the story was recorded. Later, all the oral production obtained was manually transcribed and orthographically checked.

Three assessments were performed. First, the discourse organization, total number of words in the complete description, total time duration. Second, lexical content, the proportion of closed and open words (nouns, verbs, adjectives, etc.). And third, pauses rate per one hundred words, and duration of the pauses.

As result, the authors found that AD patients elicit the same amount of words as HC, but AD patients produce more and longer pauses. In addition, they surprisingly found that the number of pauses has a positive correlation with the abilities in the narrative task. Authors concluded that, according to their results, pauses during narrative tasks could exhibit the time necessary for compensation and memory processes.

Finally, we can conclude that all these studies support the idea that pauses can be considered an important feature that describes certain language processes and, can unveil remarkable information about cognitive impairment since it is related to language processes.
For a deeper study about silences in speech and cognitive impairment, refer to (Gayraud et al., 2011), (Beattie & Butterworth, 1979), (Butterworth & Beattie, 1978), (Qiao et al., 2020), Toth et al. (2017), Tóth et al. (2015), Weiner & Schultz (2016).

1.4 Speech analysis and cognitive impairment

Several researchers have been analyzed speech as a method for early detection of AD. This is most probably due to the fact that dementia affects speech and language abilities. For a deeper study about speech and language in AD refer to Rochon et al. (2018).

Studies in this field seem to be in two lines: the analysis of the acoustic of the speech combined with linguistic characterization and, acoustic characteristics solely. Therefore, researchers have been used different methodologies to extract acoustic features. In contrast, most of the studies are based on cognitive task, the most common is the picture description task.

One of these studies is the work of (Hernández-Domínguez et al., 2018). Authors in this study performed experiments base on the Cookie Theft picture description task for eliciting spontaneous speech from subjects. They used the Pitt Corpus (Fleisher & Corey-Bloom, 2010) of the DementiaBank database.

They analyzed the information coverage given by subjects during the task; linguistic characteristics, i.e., vocabulary richness, frequency of verbs, nouns, adjectives, prepositions, conjunctions, prepositions; and, as some other previous studies, they used phonetic characteristics in their study, extracting the 13 Mel-Frequency Cepstral Coefficients (MFCC) features from speech.

For automatic classification, the SVM, and Random Forest Classifier (RFC) machine learning algorithms were implemented. Authors found that the best classifier was the RF with 94% accuracy and 93% AUC with the linguistic features set. In the case of the phonetic features performance, 72% accuracy and 71% AUC were obtained with SVM as classifier.

Later in 2020, the authors Haider et al. (2020), used the Pitt Corpus from the Alzheimer Research Program at the University of Pittsburgh (Fleisher & Corey-Bloom, 2010), this corpus contains speech recordings from AD and non-AD subjects. The tasks recorded in this corpus are the picture description task, word fluency task, story recall task, and sentence construction task.

With the audio recording, authors performed an acoustic feature extraction, these features are: MFCC, voice quality, fundamental frequency, LSP, energy measures, spectral, voicing related low-level descriptors. Additionally, they extracted the extended Geneva Minimalistic Acoustic Parameter eGeMAPS (Eyben et al., 2016), acoustic features that are based on their potential to detect physiological characteristics in the human voice. Specifically, this set contains fundamental frequency semitones, loudness, spectral flux, MFCC, jitter, shimmer, alpha ratio, Hammarberg index, and slope. Finally, they used the Multi-Resolution Cochleagram (MRCG) set of features. These features represent the multi-resolution power distribution of the audio signal in the time-frequency domain.

Aiming to perform classification, authors implemented five machine learning techniques: decision trees, Nearest Neighbor (KNN), Linear Discriminant Analysis (LDA), RF, and SVM.

As result, authors reported the best performance with all the acoustic features combined in the RF classifiers, they obtained 78.7% accuracy. Additionally, using simply the eGeMAPS set authors obtained 77.4% accuracy, with the LDA classifier.

Recently, another attempt in the line of phonetic analysis from speech, is the work of (Meghanani et al., 2021). Researchers used log-Mel spectrograms and MFCC for speech analysis. Distinctly, from previous studies, 3 models of Neural Networks were implemented: long-short term memory network (CNN-LSTM), pre-trained ResNet18 network followed by LSTM (ResNet-LSTM), and pyramidal bidirectional LSTM followed by a CNN (pBLSTM-CNN). They obtained 64.58% accuracy as the best performance of the classification task.

According to the research described above, AD patients seem to present alterations on speech that can be used to satisfactory distinguish between AD and HC.

Other relevant studies analyzing the relation between the speech production and cognitive impairment, can be found in these studies: Yeung et al. (2021), Haulcy & Glass (2021), Qiao et al. (2020), Ben Ammar & Ben Ayed (2019), Zargarbashi & Babaali (2019), Tóth et al. (2015a), Tóth et al. (2015b), Cera et al. (2018), Rochon et al. (2018), Toth et al. (2017), Fraser et al. (2016), Tóth et al. (2015), König et al. (2015), Satt et al. (2014), Bschor et al. (2001), and others.

CHAPTER 2

RESEARCH OBJECTIVES

2.1 Problem statement

Most studies that have covered the detection of early stage AD are based on manual evaluations of biomarkers and cognitive ability tests. Methods to analyze signs of AD are often invasive, leading to discomfort for the subjects involved in data collection experiments.

It seems that the study of AD in early stages has followed three approaches:

First, the most invasive approach, several studies have been held based on biomarkers like cerebrospinal fluid extracted from participants, others authors have taken MRIs and Positron Emission Tomography (PET) images from participants, these types of techniques have demonstrated to be invasive, causing discomfort and significant expenses.

The second approach seems to be the studies based on the acoustic and linguistic features during speech. Several studies support the fact that AD patients, even in early stages, suffer a detrimental in verbal communication. There are several studies associating deficits in speech production and cognitive impairment. Some of them used some picture description technique, others have memory test techniques, and some others, spontaneous speech.

The third line are the studies related to nonverbal communication and, the relationship with cognitive impairment. Interest in cognitive ability measurements has shot up due to the fact that they are non-invasive, in addition to being inexpensive.

To our knowledge, there are sparse findings on the relationship between nonverbal communication and AD, likewise, there is little research about studies relating verbal and nonverbal communication in automatic approaches. Most of the studies in this line, study the eye movements and facial expression with a sentimental analysis approach, some others have centered the analysis of silences as a marker of deficit in the abilities necessary for communication.

Additionally, most of these last studies involve manual evaluations. Subjects are told to perform specific tasks, i. e., to follow a dot on the screen, tell a story or describing a picture.

2.2 Main objectives

In this study, we follow this track and implement a non-invasive method, in which the facial expressions, eye movements, silences, and acoustic features of natural conversations among elderly subjects are analyzed from video recordings.

The three main objectives are:

- Automatic classification of AD and HC through facial expressions and gaze. We characterize the facial expressions through specific landmarks on the face and FAUs. In this way, we could track key movements on the face per frame. Regarding the gaze, we track it through a vector for each eye that describes the direction of the gaze, and through landmarks over the eyes tracking specific parts of the eyes for each frame.
- Automatic classification of AD and HC with silences and MFCC features. We extracted the silences in the audio of the conversations. Additionally, we describe the phonetic features of the subject's voice through the 13 MFCC.
- Automatic classification of AD and HC through a multimodal approach. Analysis of correlation between facial expressions, eye movements, and silences in AD. Finally, in this objective, with all the features obtained in the previous two objectives, we performed the training of the classifiers with several combinations of them.

CHAPTER 3

GENERAL METHODOLOGY

In order to reach the objectives stablished, we follow several steps to extract the features from the video recordings and perform the classification tasks. These steps are described as follow:

3.1 Corpus

For our study, we used the cohort "Conversing with the elderly in Latin America: a new cohort for multimodal, multilingual longitudinal studies on aging" (Pope & Davis, 2016) recollected in Ecuador with Lain-American Spanish speakers. This cohort is part of a project that started in 2008, The Carolinas Conversations Collection (Pope & Davis, 2011). The recollection was made with the collaboration with "*Universidad Técnica Particular de Loja*" and the "*Perpetuo Socorro*" Foundation, a home for elderly people. For more details about the physical setup used to create the database, refer to (Pope & Davis, 2016).

This collection is composed of audio and video recordings of natural conversations in Spanish among older adults. The authors of this project applied the Institutional Review Board (IRB) from their institution "École de Technologie supérieure" in Montreal, Quebec, Canada (H20150301), and the IRB from the Medical University of South Carolina in the United States (HR# 17575). They assigned a certified person, trained on American or Canadian ethics for research with human subjects, to be present in every recording. All participants provided their agreement to participate in the recordings. In the case where participants could not sign the agreement, the legal guardian consented to participate. Aliases replaced all names to protect the privacy of participants.

This collection is composed of 25 conversations, with 16 participants, 12 women, and 4 men. The average length of the conversations was 16 minutes, and they were video recorded at 30 frames per second. The range of the age participants is 70 to 91 years old. The demographic information of participants and their medical condition is described in Table 3.1. Additionally, we obtained the medical condition of participants from the psychiatric authorities in the institution. Specifically, we request if patients suffer from AD, or they are HC, see Table 3.1. We excluded other diagnoses from this study as they were not relevant for our purposes.

The Linguistic Engineering Group manually transcribed this collection at the "*Universidad Nacional Autónoma de México*" from Mexico City, a group with expertise in creating transcription from audio recordings.

We used the Ecuador section because it is composed of subjects with similar ethnicity and with AD patients and HC. The CCC corpus from Carolina contains just healthy elderly. For classification, we need samples from both populations.

Table 3.1 Demographic information of participants in the CCC Ecuador

	Women	**Men**	**Total**
Participants	12	4	16
Conversations	18	7	25
Av. age	83.9	83	83.6
Av. education (years)	6.2	8.2	6.7
AD patients	7	1	8
HC	5	3	8

3.2 Data pre-processing

The videos must be excellent (sound, resolution, clarity) to extract the required characteristics. Consequently, videos with not good enough quality were excluded from the experiments.
To obtain purely natural conversations, we cleaned the videos as follows:

1. We excluded poor-quality videos since feature detection performed poorly in them.
2. We excluded parts of the videos, including no conversations; for example, the camera, microphone, interviewer, and interviewee were set up or distracted by external stimuli.
3. We excluded parts in which the subject spent a long time in profile towards the camera.

4. We use only the parts where the participant, not the interviewer, is talking.
5. We excluded parts where noise cover the dialog or dialog was not clear enough.
6. Finally, we excluded parts of the videos in which feature detection was unsuccessful, frames where the subjects' faces were not detected.

After applying the criteria described above, we excluded two subjects from the experiments, using conversations from 14 elderly subjects. 4 males and ten females. Details about the sociodemographic information are shown in Table 3.2.

Table 3.2 Demographic information of subjects after applying the pre-processing criteria

	Women	**Men**	**Total**
Participants	10	4	14
Conversations	16	7	23
Av. age	83.9	83	83.6
Av. education (years)	6.2	8.2	6.7
AD patients	6	1	7
HC	4	3	7

CHAPTER 4

AUTOMATIC EVALUATION OF ALZHEIMER’S DISEASE, ANALYSIS OF FACIAL EXPRESSIONS AND EYE MOVEMENTS IN NATURAL CONVERSATIONS

4.1 Introduction

In this chapter, we will implement experiments aiming to analyze all the features related to facial expressions, and their relationship to AD. We will extract facial and eye landmarks. Additionally, we will obtain the gaze vector for each eye as features.

4.2 Specific methodology

4.2.1 Extraction of facial expression characteristics

We used the OpenFace 2.2.0, a facial behavior analysis toolkit (Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, 2018), an application to extract the required features in each video and on each frame. In Figure 4.1, it is shown an example of the features extracted. This application is an open-source project designed to analyze facial expressions and gaze through different approaches. In the following paragraphs, we will describe the features extracted.

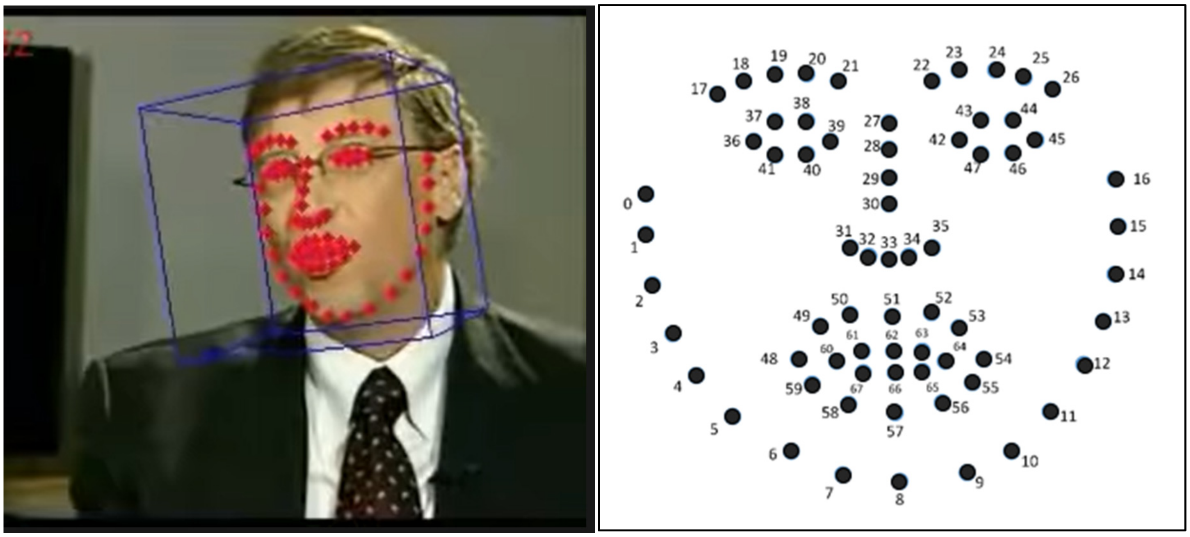

Figure 4.1 The left picture shows an example of how facial landmarks are tracked in the videos (red points tracking the movement on the face) and head tracking (blue cube over the head) in OpenFace 2.2.0. Example extracted from the video "CLNF tracking on YouTube Celebrities dataset" (Tadyla, 2014). On the right, it is presented the index of the face landmarks extracted

We extracted the following facial characteristics:

1. Facial landmark detection. We tracked the coordinates of specific points (landmarks) on the face, the index is shown in Figure 4.1;
2. Head pose tracking. Pose_Tx, pose_Ty, pose Tz. These features describe the position of the head with respect to the camera in millimeters in the x, y, and z axes, see Figure 4.1;
3. Rotation of the head. Pose_Rx, pose_Ry, pose_Rz. These features describe the rotation of the head in radians. The convention used is the positive left-handed sign (R = Rx * Ry * Rz);
4. FAU_c. Features to describe the FAUs in binary. If FAU appears, the value equals 1; if the FAU is not detected, the value will be 0. The system can detect 18 FAUs: 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 28, and 45;
5. FAU_r. Describes the intensity of the FAUs number from 0 to 5. Only FAU 28 is excluded in this set. FAUs 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, and 45;
6. Parameters of the Point Distribution Model (PDM). This model describes the rigid face shape: scale (p_scale), rotation (p_rx, p_ry, and p_rz), and translation (p_tx and p_ty);

Description of gaze features. OpenFace performs gaze analysis using five sets of features:

- Gaze_0, this set describes the direction vector in world coordinates for the leftmost eye in the image.
- Gaze_1, this set describes the direction vector in world coordinates for the rightmost eye in the image.
- Gaze_angle is the average of the gaze directions in both eyes in radians, in world coordinates.
- Eye_landmarks 2D, this set describes the coordinates of the landmarks of the eyes in 2D. See Figure 4.2**.**
- Eye_landmarks 3D, this set describes the coordinates of the landmarks of the eyes in 3D in millimeters. See Figure 4.2**.**

It is important to note that tracking facial and eye landmarks are independent from the head movements. The performance of Openface is described in detail in (Baltrusaitis et al., 2018).

First, OpenFace detects the faces for each frame in the video recordings, for that purpose OpenFace uses the Multi-task Convolutional Neural Network (MTCNN) face detector (Zhang et al., 2016). This face detector was trained with Wider Face (S. Yang et al., 2016), and CelebA (Liu et al., 2015) datasets.

For extracting the facial and eyes landmarks, OpenFace implements the Convolutional Experts Constrained Local Model (CE-CLM) (Zadeh et al., 2017), a model for tracking the landmarks, this model has two main components:

First, the PDM that detects the variation of each landmark and, second, a patch expert that model the local appearance variation for each landmark detected. In order to train the PDM, the OpenFace authors used the Labeled Face Parts in the Wild (LFPW) (Belhumeur et al., 2013) and Helen (Le et al., 2012) datasets.

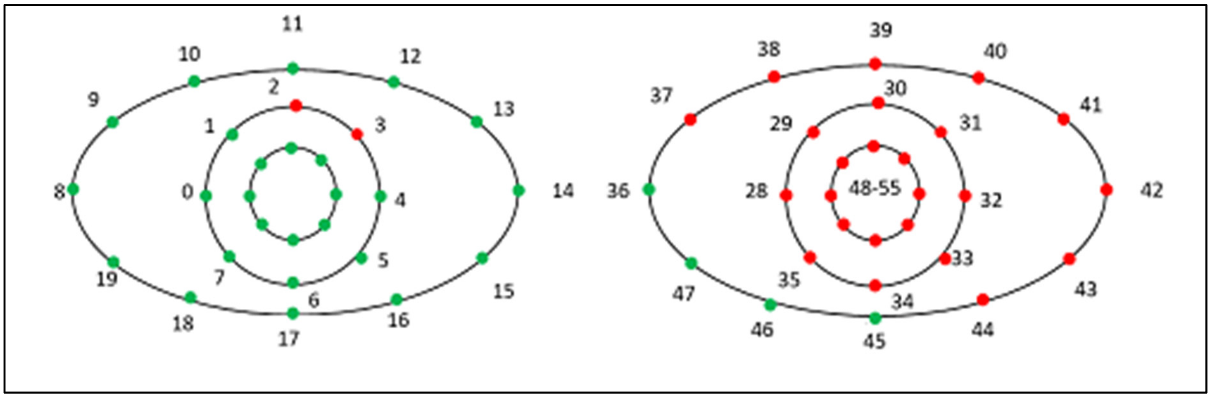

Figure 4.2 Index of the eyes landmarks extracted. The red points composed the best features in the automatic classification task

4.2.2 Automatic classification

The purpose of automatic classification is to discriminate HC from AD patients using machine learning techniques. In this study, we used three machine learning algorithms, namely, RFC, KNN, and NB. We used these algorithms because they are well known for their classification ability. Moreover, these algorithms showed a better performance in terms of execution time (for both training and prediction tasks). We used the data obtained in the feature extraction process to train the algorithms.

We performed ten-folds cross-validation to evaluate the algorithms. We took 90% of the samples as the training set and the remaining 10 % as the test set. Each sample corresponds to one frame in a video recording. We report the results as the average of these ten results. In this experiment, we did not consider the AD level, but rather, we performed a binary classification between HC and AD.

4.3 Results

In the present section, we will explain the results of the classification task obtained. The results of the correlation analysis as well as the FAU occurrences in HD and AD groups are presented.

4.3.1 Features performance

Among the 713 features extracted, as described in section “3.2.1 *Extraction of facial expression characteristics*", we retained 162 features following the ANOVA F-value. This statistic is well known for measuring the correlation between features and was thus used to obtain the features in terms of best classification performance.

After measuring the F-value for the classification task with a different combination of features, we found that the best performances were obtained with the features related to the eyes landmarks in the y- axis, the facial landmarks in the y-axis, and pose_Rx, which is the rotation in radians of the head in the x-axis.

These features are described below:

1. Eye_landmarks 2D, points 2, 3, 28-35, 37-44, 48-55. These features describe the location of the eye landmarks in 2D, in the y-axis. See Figure 4.2;
2. Eye_landmarks 3D, points 28-45, 48-55. These features describe the location of the eye landmarks in 3D in the y-axis. See Figure 4.2;
3. Face landmarks, y-axis. Points: 0, 1, 10-26, 42-47. These features describe the location of the face landmarks in 2D. See Figure 4.3.

Face landmarks, y-axis. Points: 0-2, 10-16, 19, 20, 22-26, 42-47. These features describe the location of face landmarks.

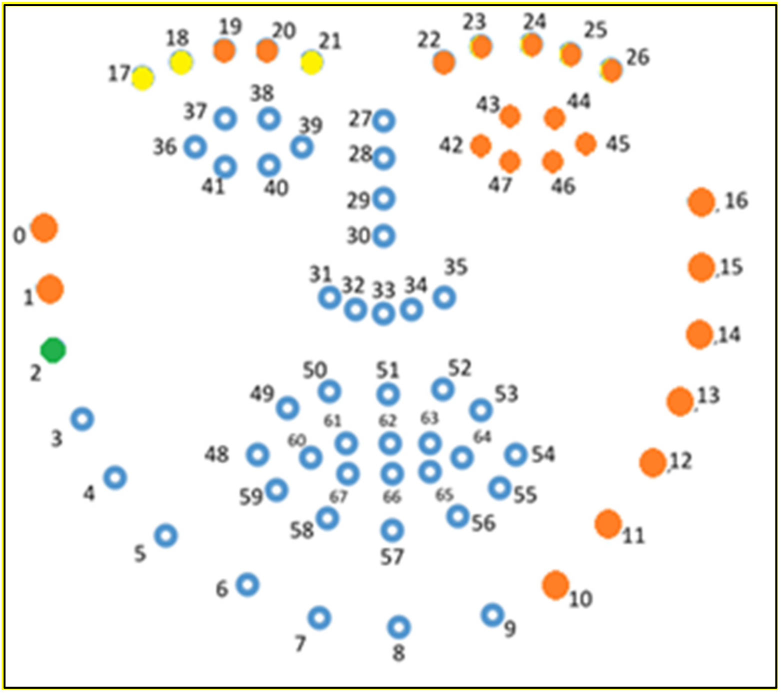

Figure 4.3 Face landmarks. The points shown are tracked on each frame of the videos. The points in yellow, green, and orange compose the best features in the automatic classification task. The yellow points are the points in 2D exclusively. The green point is the point in 3D exclusively. The orange points composed the best features in both 2D and 3D groups

4.3.2 Automatic classification

We tested the RFC, KNN, and NB to obtain the best classification performance, using the Gaussian and Bernoulli approaches in the NB classifier. Additionally, we trained the algorithms using different combinations of sets of features. Our models were trained with gaze features, facial landmarks, and FAUs, independently and combined. Table 4.1 and Table 4.2 show the results obtained. Our best performance is the RFC trained with all the features; we obtained a 93.3 % accuracy and a 98.38% ROC.

Table 4.1 Description of the performance of the models in classifying HC and AD with facial features, RFC, and KNN classifiers

Model	***Exp.***	***Features***	***Accuracy***	***Precision***	***Recall***	***F1***	***ROC***
RFC	F1	Gaze	0.7633	0.8444	0.7703	0.8003	0.8272
	F2	Facial landmarks	0.8629	0.9395	0.8347	0.8747	0.9320
	F3	**FAU**	**0.9118**	**0.9453**	**0.9164**	**0.9279**	**0.9716**
	F4	Gaze and Facial landmarks	0.8787	0.9435	0.8612	0.8932	0.9460
	F5	Gaze and FAU	0.8139	0.9040	0.7940	0.8319	0.8888
	F6	Facial landmarks and FAU	0.8849	0.9570	0.8557	0.8960	0.9505
	F7	Gaze, facial landmarks and FAU	0.8933	0.9590	0.8698	0.9057	0.9563
	F8	Best features (F-value)	0.8793	0.9462	0.8592	0.8933	0.9503
	F9	**All features**	**0.9330**	**0.9847**	**0.9099**	**0.9381**	**0.9838**
KNN	F1	Gaze	0.8786	0.9411	0.8756	0.8971	0.9637
	F2	**Facial landmarks**	**0.9144**	**0.9721**	**0.8940**	**0.9252**	**0.9769**
	F3	FAU	0.8928	0.9130	0.9225	0.9161	0.9586
	F4	Gaze and Facial landmarks	0.8999	0.9683	0.8758	0.9108	0.9759
	F5	Gaze and FAU	0.8786	0.9411	0.8756	0.8971	0.9637
	F6	**Facial landmarks and FAU**	**0.9144**	**0.9721**	**0.8939**	**0.9251**	**0.9769**
	F7	Gaze, facial landmarks and FAU	0.8999	0.9683	0.8758	0.9108	0.9759
	F8	Best features (F-value)	0.8660	0.9227	0.8635	0.8853	0.9333
	F9	**All features**	**0.9006**	**0.9685**	**0.8765**	**0.9114**	**0.9765**

Table 4.2 Description of the performance of the models in classifying HC and AD, Naïve Bayes classifier

Model	*Exp.*	*Features*	*Accuracy*	*Precision*	*Recall*	*F1*	*ROC*
Gaussian	F1	Gaze	0.7484	0.8606	0.7710	0.7896	0.7854
Bernoulli	F2	Facial landmarks	0.6860	0.7829	0.6749	0.6364	0.8064
Gaussian	F3	Facial landmarks	0.7462	0.8643	0.7633	0.7805	0.7871
Gaussian	**F4**	**FAU**	**0.7402**	**0.8388**	**0.7381**	**0.7718**	**0.8168**
Gaussian	F5	Gaze and Facial landmarks	0.7466	0.8627	0.7655	0.7842	0.7872
Gaussian	F6	Gaze and FAU	0.7548	0.8612	0.7820	0.7968	0.7903
Bernoulli	**F7**	**Facial landmarks and FAU**	**0.7265**	**0.7837**	**0.7384**	**0.7145**	**0.8126**
Bernoulli	F8	Gaze, facial landmarks and FAU	0.6623	0.7788	0.6366	0.6100	0.7973
Gaussian	F9	Best features (F-value)	0.7280	0.8261	0.7810	0.7769	0.7530
Gaussian	F10	All features	0.7484	0.8607	0.7710	0.7896	0.7854

4.3.3 Correlation analysis

We additionally analyzed the Pearson correlation for each feature to the AD condition. We found that the correlation obtained is consistent with the ANOVA F-value. The results showed the best features are part of the y-axis, except for the 'Pose_Rx' (which describes the position of the head in the x-axis), which is the feature that is the second most correlated to AD.

Furthermore, we analyzed the average of each group of features obtained: Pose_Rx, Face landmark 2D, Face landmarks 3D and, Eye landmarks 2D. Our findings are reported in Table 4.3.

It can be seen that the pose of the head (Pose_Rx), with one feature, has a stronger correlation than the rest of the features. Thus, the landmarks on the face (2D and 3D) have a lower correlation. Finally, the group of features in "Eye landmarks 2D" presents the lowest correlations.

Table 4.3 Comparison between Pearson correlation and t-test (t-statistic), grouped by type of feature and ordered with respect to the strength of their correlation

Features	Axis	Pearson Corr.	T-statistic
Pose_Rx	x	0.5170	462.1364
Face landmark 2D	y	-0.5079 (av.)	-451.1914 (av.)
Face landmarks 3D	y	-0.5029 (av.)	-445.5095 (av.)
Eye landmarks 2D	y	-0.5001 (av.)	-441.7983 (av.)

Table 4.4 Comparison of features Pearson correlation with AD by age and gender

Features	Axis	Corr. by age	Features	Axis	Corr. by gender
Pose_Rx	x	0.5073	Pose_Rx	x	0.5260
Face landmarks 2D	y	-0.4983 (av.)	Face landmarks 2D	y	-0.5041 (av.)
Face landmarks 3D	y	-0.4940 (av.)	Face landmarks 3D	y	-0.4965 (av.)
			Eye landmarks 2D	y	-0.4933 (av.)
			Translation of the face	y	-0.4923

We also performed a t-test (t-statistic) for each feature. First, we analyzed the normal distribution with the response variable AD. Also, we analyzed the variances and checked for the similarity between them (relation 4:1). The results showed a very small P-value (very close to zero) for all of them. Our results are shown in Table 4.3. As can be seen, the t-test results were very similar to the Pearson correlation results.

We also analyzed the Pearson correlation by gender and age, and our findings are described in Table 4.4. According to the results, we can notice no strong correlation between AD and gender or age. The best result is with the feature describing the pose of the head in the axis “x” (Pose_Rx) with 0.52. This is most probably due to our small population (ten females and four males).

4.3.4 FAU occurrences in HC and AD groups

As described in the FAUs section, the Facial Action Units describe specific movements of the face muscles. According to the results in the classification task, it is notable that AD and HC groups display specific movements that can be distinguished with machine learning techniques. For this reason, we considered it interesting to analyze the number of incidences in the HC group and AD group of each FAU.

In Figure 4.4, it is shown the number of each FAU detected in the binary features in the AD and HC groups. Regarding the FAU features that measure the intensity, we also counted for each FAU. We consider as an incidence a value between 2.8 and 5. The results related to the AD group and HC are shown in Figure 4.5.

Additionally, we analyze the amount of FAUs detected in the AD and HC groups, respectively. Likewise, we analyze the quantity of FAUs detected for the group of "Upper face" and "Lower face." The results are shown in Figure 4.6 and Figure 4.7. Remarkably, the HC group presents more FAUs in both the 'Lower face' and 'Upper face' groups; comparing with the AD group.

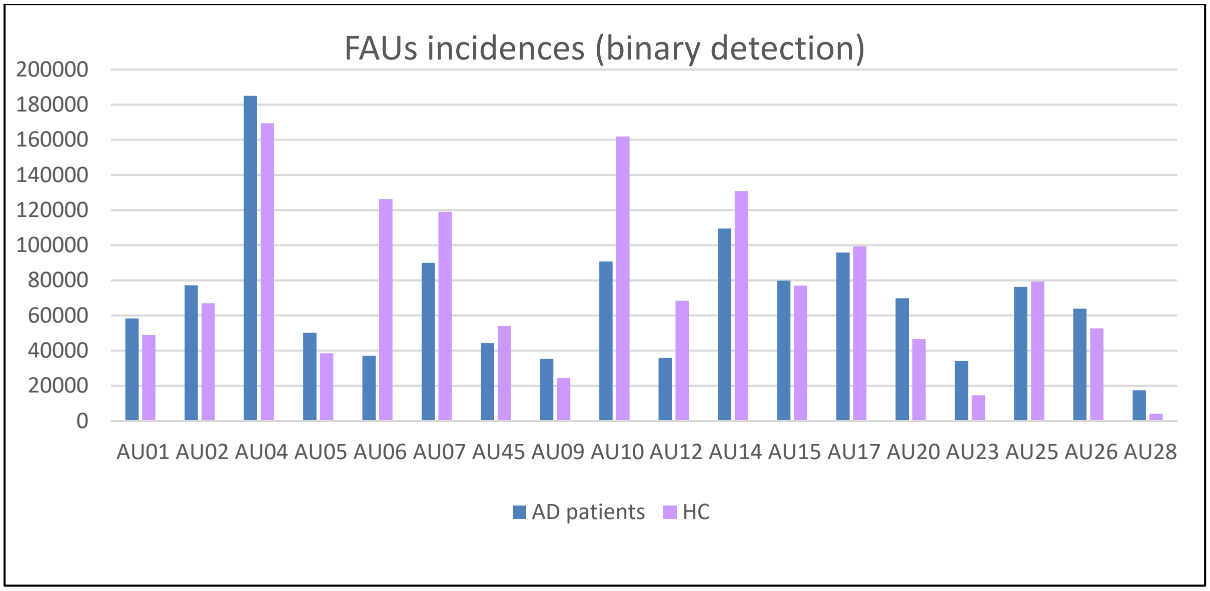

Figure 4.4 Number of incidences detected for each FAU binary detected. Blue bars describe the number of incidences for AD patients. Purple bars describe the number of incidences for HC

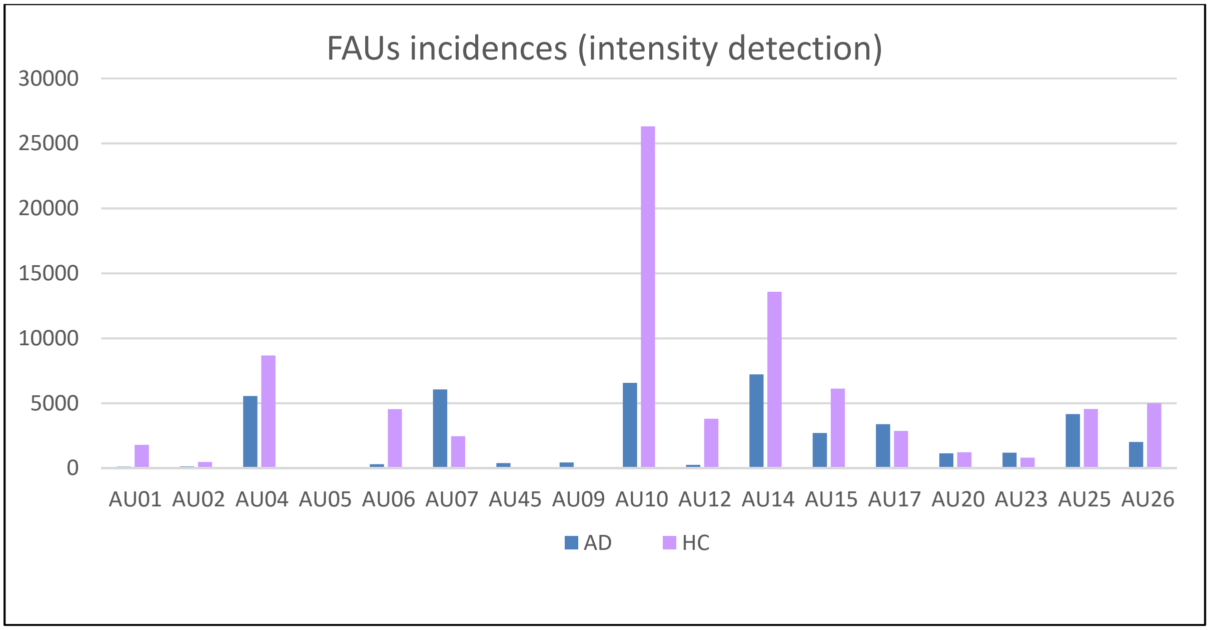

Figure 4.5 Number of incidences detected for each FAU in the intensity set of features. Blue bars describe the number of incidences for AD patients. Purple bars describe the number of incidences for HC. We consider as an incidence the values from 2.8 to 5

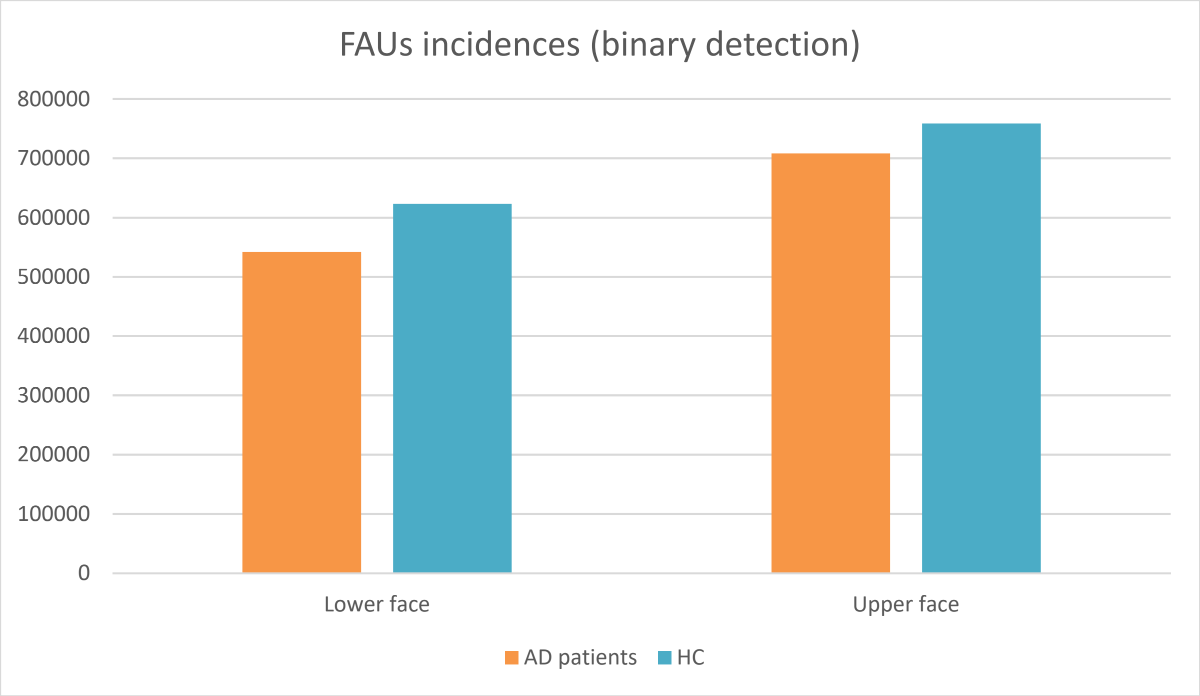

Figure 4.6 FAUs incidences in binary detection. FAUs are grouped by Lower and Upper face

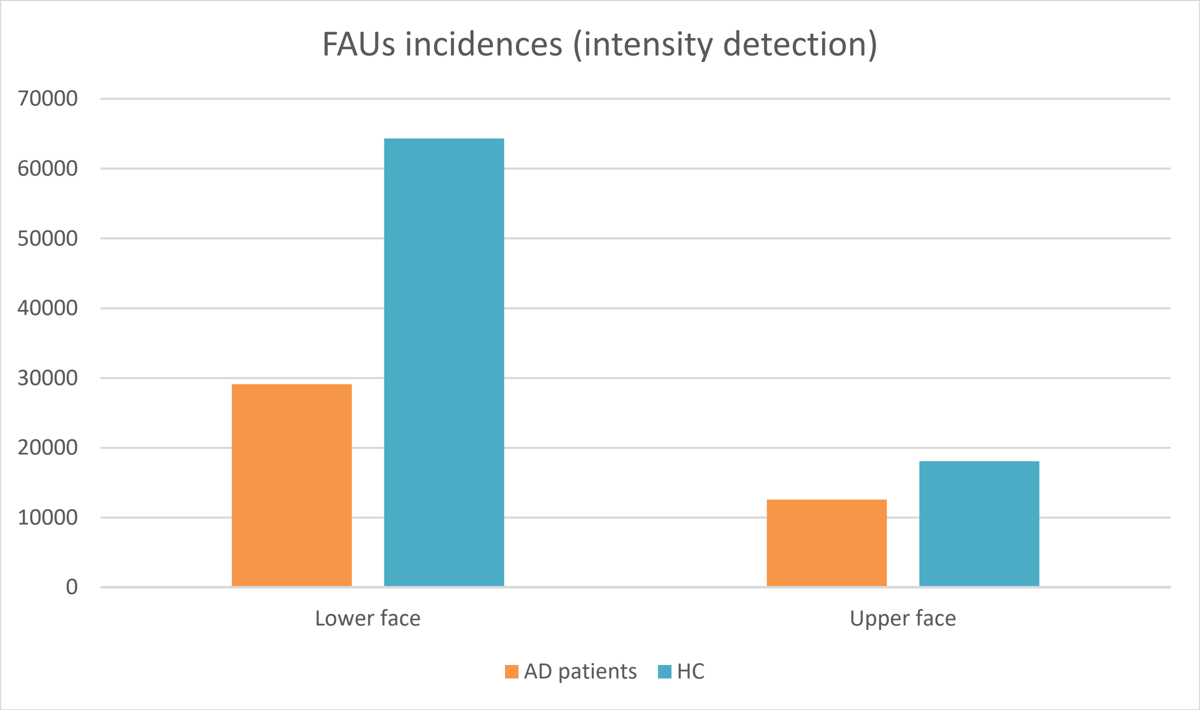

Figure 4.7 FAUs incidences in the intensity detection. FAUs are grouped by Lower and Upper face

4.4 Discussion

Regarding the performance of each set of features, it is remarkable that the features in the “y” axes are the features that have a better performance among all features in the classification task, as was expected. When subjects were recorded in natural conversation, the most common movements on the head occurred more in the axis "y." This is because we tend to move the head left to right, more than up and down.

We performed experiments with the different features from facial expressions and presented the results in Table 4.1 and Table 4.2.

As is shown in Table 4.1, the RFC showed the best performance in training the algorithm with all the features, and we obtained 93 % accuracy and 98 % sensitivity. With the Facial Action Units, we obtained 91 % accuracy and a 97.1 % ROC measure. This indicates that we can obtain significant results using just FAUs to train the RFC. Comparing the performance of the KNN classifier, we obtained an 89% accuracy and a 95% ROC with FAUs features.

On the other hand, KNN performs better with "Facial landmarks" and "Facial landmarks combined with FAUs," obtaining, in both cases, the same performance (91.44 % accuracy and 97.69 % ROC). In addition, when we train KNN with all features, it performs worse than with only the FAUs group, with a 90% accuracy and a 97.65% ROC.

Besides, in the case of the NB classifier, we have lower performance measures. The best NB performance was the Gaussian approach, training the algorithm with FAUs features. With it, we have a 72 % accuracy and an 82 % ROC. However, this result is not as good as obtained with the other two classifiers: KNN and RFC.

In the case of the RFC, it can be seen that it performs well when trained with the group of features independently and even better if it is trained with all the features together. By contrast, the rest of the classifiers tested performed better just with FAU and landmarks features.

However, these results may be limited because our database is small and composed of 16 subjects, and facial expressions strongly correlate with cultural and ethnic aspects. This limitation has been compensated by the nature of facial expressions in video recordings. We have 30 images to analyze per second, and also, per frame, we obtained 713 features. This means that, even if the collection used for the experiments is small at first sight, we have a considerable quantity of data.

We also analyzed the occurrences of each FAU on each group, AD and HC; results are reported in Figure 4.4 and Figure 4.5. Results indicate that the three FAUs with the highest number of occurrences in the binary detection are FAU04 (brow lowerer) with 185,020 occurrences in the AD group and 169,410 occurrences in the HC group; FAU10 (Upper lip raiser) with 90,730 occurrences in the AD group, and 161,860 occurrences in HC group; and FAU14 (Dimpler) with 109,460 occurrences in the AD group, and 130,780 occurrences in HC group.

In contrast, in the intensity detection, FAU10 (Upper lip raiser) was the FAU with the highest number of occurrences with 6,570 occurrences in the AD group and 26,270 occurrences in the HC group; FAU14 (Dimpler) with 7,220 occurrences in the AD group, 13,580 occurrences in the HC group; and FAU04 (brow lowerer) with 5,560 occurrences in the AD 8,670 occurrences in the HC group. The results obtained in binary detection are consistent with the intensity detection. We have the same three FAUs in both cases but with different quantities of occurrences.

The large amount of FAU04 detected is probably due to all subjects being older adults. Therefore, the wrinkles around the eyebrows could be detected as FAU04.

Remarkably, the quantity of FAU10 detected in both intensity and binary detection. The number of incidences in the HC group doubles the number of incidents in the AD group in binary detection. In the case of the intensity, the number of incidences in HC is remarkably higher than AD group.

Regarding the analysis grouping FAUs by upper and lower face, we can observe in the intensity detection that the ‘Lower face’ group has remarkable more incidences than the 'Upper face' group. On the contrary, the binary detection 'Upper face' group presents a slightly higher quantity of incidences than the 'Lower face.' The most remarkable result is that we obtained more FAU incidences of FAUs in all HC groups, intensity, and binary detection compared with the amount of FAUs detected in the AD group. Likewise, the AD group demonstrates less expressiveness than HC, according to the FAUs detected in our experiments. As we mentioned in the introduction, the brain parts involved in the emotional process tend to be impaired in AD patients. Our findings confirm that AD patients present less expressiveness.

CHAPTER 5

AUTOMATIC EVALUATION OF ALZHEIMER’S DISEASE, ANALYSIS OF SILENCES AND ACOUSTIC FEATURES IN NATURAL CONVERSATIONS

5.1 Introduction

In this study, we implement a multimodal approach for the analysis of AD in natural conversations. Specifically, in this section, we will analyze the phonetic features in AD patients and HC. We will use machine learning techniques to automatically distinguish between AD and HC groups with the 13 MFCC features. Additionally, we evaluated the mean, skewness, kurtosis, and variance of these 13 MFCC features.

5.2 Specific methodology

In the general methodology we pre-processed the data to have the video recordings without noise and/or external sounds. In this section, we will focus exclusively in the sound of the videos, we will explain on detail the steps we followed.

5.2.1 Pre-processing data

As we proceeded with the facial and gaze analysis, it is necessary to exclude some parts of the videos with certain criteria. In this case, we made the exclusion following the steps described as follows:

1. We extracted the audio file (.wav) for each video,
2. We kept the parts of the audio where the interviewee is talking. In Figure 5.1, it is shown an example of the cutting process. The silences of the subjects of interest where included in the audio. Likewise, the parts where the interviewer is talking were excluded even if the subjects were talking in this time span,
3. We dismissed also, the parts where the quality of the sound was not good enough, i. e., parts whit noise, external sounds or parts where the interviewee cannot be heard clearly,

4. Finally, we segmented the audios obtained every 3 seconds, we extracted the phonetic features on each of these segments.

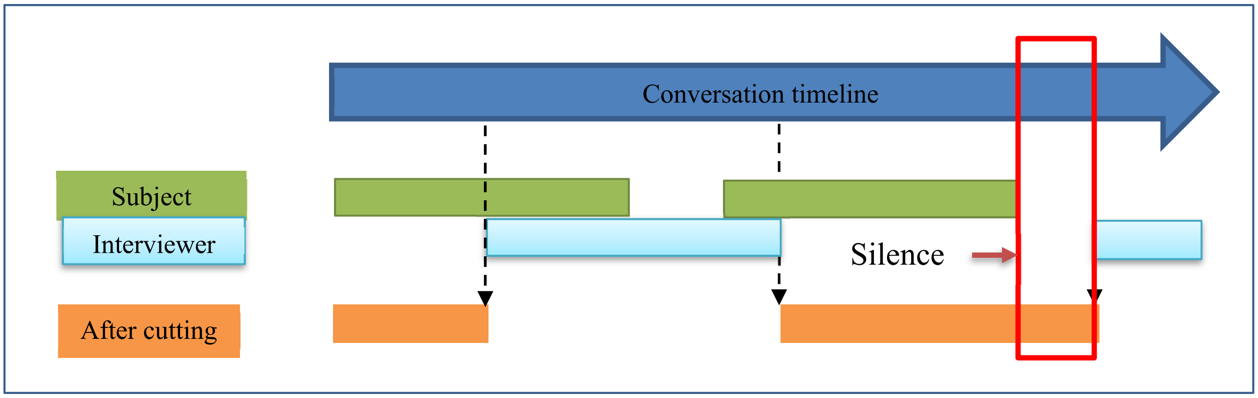

Figure 5.1 Timeline of the audio cutting process

5.2.2 Extraction of silences and phonetic features

Once we obtained all the audio files cleaned, the silences were extracted with SilenceDetection app (https://github.com/LiNCS-lab/SilenceDetection, Baissus 2020), this application was made during an internship at LiNCS at the École de technologie supérieure. An application that detect time spans where human voice is not detected. In Figure 5.1 we can see an example of how a silence was detected. The red rectangle shows the segment where a silence is detected.

Time spans where obtain as ‘Start_time’, ‘End_time’, in seconds for each audio file obtained in the cleaning process described above, we consider a silence as a minimum of 250 ms duration, we chose this threshold according to the literature reviewed related to this topic. With this information, we estimate the number of silences, total duration (the duration accumulative of all the silences in the span time) and the duration average per span time.

Additionally, we extracted the 13 MFCC features from audio files. For this analysis we used the phonetic measurement which is a part of the application implemented in the article (Abiven & Ratté, 2021). This phonetic extraction estimates the 13 MFCC. Then, this application obtains the mean, kurtosis, skewness, and variance of these coefficients. Likewise, we obtained 52 phonetic features. For more details about this application refer to the article (Abiven & Ratté, 2021). Finally, we obtained a total of 55 features, 52 phonetic and 3 silence related.

5.2.3 Classification

Similar to the methodology for the facial and eye movements features, we performed an automatic classification with machine learning techniques. We implemented four machine learning techniques: RF, KNN, NB, and SVM. The phonetic data obtained is considerably smaller comparing with facial and eye movement features. The execution time in the task of SVM training with facial and eye movement features was considerably large. Due to the small data obtained in the phonetic analysis, the execution time was considerably smaller than the execution time with the facial and eye movement features. For this reason, we implemented the analysis with SVM technique.

We performed experiments with different combination of features in order to analyze the performance on each machine learning technique. A ten-folds cross-validation was performed for every model training, a sample is considered as a frame of video recordings. We obtain 90 % of the dataset as the training set and the 10 % as the test set. Results are reported as the average of these ten results. Likewise, results are shown in following paragraphs.

5.3 Results

5.3.1 Automatic classification

We performed different combination of features to train our models, phonetic, silence, and phonetic and silence. Results are reported in Table 5.1and Table 5.2. The Table 5.1 presents the results obtained with RFC, KNN, and SVM models. We observed that the best result is with the model RFC trained with “silence and phonetic” features sets with 81 % accuracy and 91 % ROC.

Table 5.1**.** Performance of the models in classifying HC and AD, RFC, KNN, and SVM classifiers with phonetic and silence features

Model	*Exp.*	*Features*	*Accuracy*	*Precision*	*Recall*	*F1*	*ROC*
RFC	P1	Silence	0.5659	0.6113	0.6651	0.6366	0.5652
	P2	**Phonetic**	**0.80**	**0.8570**	**0.8592**	**0.8455**	**0.8751**
	P3	**Silence and phonetic**	**0.8136**	**0.8778**	**0.8194**	**0.8155**	**0.9132**
KNN	P1	Silence	0.6093	0.6014	0.9436	0.7345	0.6204
	P2	Phonetic	0.7090	0.7318	0.8953	0.7955	0.7934
	P3	Silence and phonetic	0.6434	0.6573	0.8223	0.7054	0.8184
SVM	P1	Silence	0.6110	0.6101	0.8862	0.7228	0.6169
	P2	Phonetic	0.8008	0.8597	0.8652	0.8433	0.8341
	P3	Silence and phonetic	0.8076	0.8706	0.7917	0.7892	0.8642

Table 5.2 Performance of the models in classifying HC and AD, Naïve Bayes classifier with phonetic and silence features

Model	*Exp.*	*Features*	*Accuracy*	*Precision*	*Recall*	*F1*	*ROC*
Gaussian	P1	Silence	0.5706	0.5745	0.9602	0.7187	0.5680
	P2	Phonetic	0.7665	0.8503	0.8110	0.8140	0.7763
	P3	**Silence and phonetic**	**0.7909**	**0.8486**	**0.7899**	**0.7862**	**0.8269**
Bernoulli	P4	Silence	0.5723	0.5723	1.0	0.7279	0.5664
	P5	Phonetic	0.7665	0.8503	0.8110	0.8140	0.7763
	P6	Silence and phonetic	0.7717	0.8026	0.7686	0.7640	0.7862
Multinomial	P7	Silence	0.5728	0.5739	0.8532	0.7250	0.6151

In addition to the classification analysis, we also analyzed the correlation between each feature and AD. We performed correlation analysis with different metrics. First, we obtained the Pearson correlation for the 55 features. Then, we obtained the average for each set of phonetic features (mean, kurtosis, skewness, and variance). We did the same for the P-value for each set of phonetic features. Also, we obtained the absolute average of the t-statistic metric and the average of each set of phonetic features was obtained.

It is worth noting that the best performance is the 'mean' set of phonetic features with a very small p-value, a considerably large t-statistic and with the largest negative Pearson correlation. Results are reported in Table 5.3.

Table 5.3 Pearson correlation and t-test statistics with phonetic and silence features

Features	**Set type**	**Pearson Corr.**	**P-value**	**T-statistic**
Phonetic	Mean	-0.347 (av.)	3.93×10^{-30} (av.)	28.94 (ab. av.)
	Kurtosis	-0.183 (av.)	4.40×10^{-2} (av.)	8.82 (ab. av.)
	Skewness	-0.066 (av.)	6.29×10^{-3} (av.)	15.22 (ab. av.)
	Variance	0.184 (av.)	2.26×10^{-7} (av.)	17.59 (ab. av.)
Silences	Number	0.115	5.31×10^{-7}	5.03
	Cumulative duration	-0.046	4.53×10^{-2}	-2.00
	Duration av.	-0.180	3.70×10^{-15}	-7.93

Additionally to the classification task, we analyzed the incidences of silences on each AD and HC groups. For this purpose, we first measure the amount on data on each group to be equal. We consider a frame as the unit to make this measure. We choose to define a sample as a frame instead of, for example, a 3 seconds segment because choosing the latter would have reduced the number of samples drastically. After detecting the largest group, we randomly chose data so we could have the same amount of data in both AD and HC group.

We found that AD subjects present a slightly higher number of silences than HC subjects, AD subjects present around 15 % more occurrences than the number of occurrences in HC.

5.4 Discussion

We performed experiments with phonetic and silence features and we present the results in Table 5.1 and Table 5.2. Results obtained showed that the best model is RFC with silence and phonetic features, model performance presented 81 % accuracy and 91 % ROC. Interestingly, the performance of the same RFC model trained exclusively with the phonetic features, has an accuracy of 80 % and 87 % of ROC. Training the model with phonetic and silence features presents an improvement of just 1 % accuracy and 4 % of ROC, comparing with the model trained solely with phonetic features.

On the other hand, silence features present a deficient performance comparing with the rest of the combination of features and models. It is notable that training with silence features any of the models implemented in this study, they failed on distinguish between AD and HC groups. This result could be attributed to the number of features, silences features are three and phonetic features thirteen. Additionally, previous studies have demonstrated that phonetic features have a significant correlation with some types of dementia, specifically, some other studies have shown a correlation with AD. On the other hand, as it is explained in the literature review section, the relationship between silences and AD have been studied. In general, studies analyzing phonetic features seem to have a better performance comparing to studies analyzing just silences.

Regarding the statistics measuring the performance, in general, the best is the ‘recall’ statistic. It is remarkable that even in the cases where the accuracy, precision, F1, and ROC do not present a satisfactory performance, the recall presents satisfactory results. One remarkable example is the Multinomial NB model, see Table 5.2, this model presents 57 % accuracy and 98% of recall. The recall statistic is the rate of the true positives with respect to the true positives plus false negatives. Therefore, having satisfactory recalls suggest all models tested tend to have a significative small number of false negatives; i.e., it is highly unlikely models predicts a HC subject incorrectly.

In the matter of correlation analysis, we performed correlation analysis with different metrics. First, we analyzed the Pearson correlation which determines the linear correlation between two samples. From the Pearson correlation results, we can conclude that the linear correlation is weak between the features and AD. Also, we obtained the p-value for each feature and then as average per set of phonetic features. In general, all phonetic features present satisfactory p-values, less to 0.007, except for the kurtosis with 0.044, even this last result could be considered a good performance. Regarding the t-statistic metrics, in general, for the phonetic features, we obtained large t-statistic, which is consistent with the respective p-values obtained.

In the case of the silence features, they present deficient performance in the Pearson correlation analysis, similar to the analysis with phonetic features, we can conclude that there is no linear correlation with AD. In contrast, the p-value metrics present satisfactory performance (less than 0.046) if we consider a p-value threshold of 0.05 or less as satisfactory.

According to the above, we can conclude that there is no linear correlation to any of the analyzed features. However, there is a correlation according to the p-values and t-statistics presented, specially, in the ‘mean phonetic features’ with a p-value very close to zero.

Finally, we did not obtain a significant difference of the number of silences between AD and HC group. In natural conversations it is more complicated to extract accurately each silence. First, if the interviewer’s voice overlaps a silence from the interviewee, this duration is interrupted. Then it is impossible to extract the actual duration of the silence.

CHAPTER 6

MULTIMODAL ANALYSIS: THE RELATIONSHIP BETWEEN VERBAL AND NONVERBAL COMMUNICATION IN ALZHEIMER'S DISEASE

6.1 Introduction

In chapters 3 and 4 we performed the analysis of the verbal and nonverbal communication in AD, respectively. In the present chapter we will analyze the performance of the models already analyzed (RF, KNN, SVM, and NB) trained with a combination of verbal and nonverbal communication features. Additionally, we will analyze the relationship between verbal and nonverbal communication features with different approaches described in the following sections.

6.2 Specific methodology

As is described in previous chapters, we obtained several types of features from videos with natural conversations. Specifically, we obtained facial and gaze features related to nonverbal communication. On the other hand, we obtained phonetic features such as silences and MFCC related to verbal communication features.

With all the features obtained, we performed different combination of features in order to evaluate the performance, we keep the models evaluated in previous chapters, and we trained them with distinct combinations of verbal and nonverbal communication features.

6.3 Results

6.3.1 Automatic classification

The results of the classification task are described in Table 6.1 and Table 6.2. The results obtained indicate that the best performance was the RFC classifier with the facial and silence features with 87% accuracy and 93% ROC.

Table 6.1 Performance of the models in classifying HC and AD, RFC, KNN, and SVM classifiers with all features in different compositions

Model	*Exp.*	*Features*	*Accuracy*	*Precision*	*Recall*	*F1*	*ROC*
RFC	**M1**	**Facial and silence**	**0.8690**	**0.9179**	**0.8548**	**0.8797**	**0.9335**
	M2	Facial and phonetic	0.8697	0.9167	0.8577	0.8813	0.9301
	M3	Gaze and silence	0.8121	0.8756	0.8148	0.9809	0.9049
	M4	Gaze and phonetic	0.8044	0.8752	0.7986	0.8194	0.9028
	M5	Facial, gaze, phonetic	0.8572	0.8997	0.8539	0.8697	0.9293
	M6	**All the features**	**0.8684**	**0.9151**	**0.8563**	**0.8790**	**0.9321**
KNN	M1	Facial and silence	0.8650	0.9191	0.8741	0.8733	0.9082
	M2	Facial and phonetic	0.8741	0.9261	0.8782	0.8809	0.9098
	M3	Gaze and silence	0.8921	0.9182	0.9188	0.9076	0.9173
	M4	Gaze and phonetic	0.8985	0.9247	0.9220	0.9135	0.9230
	M5	Facial, gaze, phonetic	0.9081	0.9385	0.9238	0.9220	0.9310
	M6	**All the features**	**0.9081**	**0.9386**	**0.9238**	**0.9220**	**0.9310**
SVM	M1	Facial and silence	0.8924	0.9031	0.9449	0.9188	0.8807
	M2	Facial and phonetic	0.8905	0.8995	0.9470	0.9176	0.8841
	M3	Gaze and silence	0.7662	0.8148	0.8410	0.8120	0.8381
	M4	Gaze and phonetic	0.7909	0.8273	0.8589	0.8325	0.8450
	M5	Facial, gaze, phonetic	0.8698	0.8881	0.9231	0.8988	0.8992
	M6	All the features	0.8700	0.8881	0.9234	0.8990	0.8992

Table 6.2 Performance of the models in classifying HC and AD, Naïve Bayes classifier with all features in different compositions

Model	*ID*	*Features*	*Accuracy*	*Precision*	*Recall*	*F1*	*ROC*
Gaussian	M1	Facial and silence	0.7163	0.7013	0.9782	0.8086	0.7045
	M2	Facial and phonetic	0.7202	0.7057	0.9673	0.8084	0.7069
	M3	Gaze and silence	0.7633	0.7865	0.8457	0.8092	0.7523
	M4	Gaze and phonetic	0.7794	0.7963	0.8586	0.8191	0.7632
	M5	Facial, gaze, phonetic	0.7350	0.7712	0.8472	0.7839	0.7128
	M6	**All the features**	**0.7631**	**0.7594**	**0.9210**	**0.8243**	**0.7475**
Bernoulli	M7	Facial and silence	0.7295	0.7529	0.8227	0.7792	0.7089
	M8	Facial and phonetic	0.7412	0.7609	0.8313	0.7866	0.7138
	M9	Gaze and silence	0.4964	0.4651	0.5985	0.4705	0.6698
	M10	Gaze and phonetic	0.4993	0.4647	0.5976	0.4703	0.7130
	M11	**Facial, gaze, phonetic**	**0.7631**	**0.7594**	**0.9210**	**0.8243**	**0.7473**
	M12	All the features	0.7362	0.7766	0.8480	0.7853	0.7133

6.3.2 Multimodal correlation analysis

In this section, the correlation analysis of all the features extracted in this study is performed. We obtained facial, gaze, phonetic and silence features, related to verbal and nonverbal communication. With all these features extracted we performed Pearson correlation and T-statistic test for each feature, the Table 6.3, describe the results obtained.

Table 6.3 Best feature performances in T-test (t-statistic) and p-value, grouped by type of feature and ordered with respect to the strength of their correlation

Features	Axis	P-value	T-statistic
PDM translation	y	Very close to zero	-45.39
Eye landmarks 2D	y	Very close to zero	-44.98 (av.)
Pose_Ty	y	Very close to zero	-44.61
Facial landmarks 2D	y	Very close to zero	-44.51 (av.)
Facial landmarks 3D	y	Very close to zero	-44.43 (av.)
Eye landmarks 3D	y	Very close to zero	-44.42 (av.)
FAU 06	-	Very close to zero	-38.79
Phonetic (mean)	-	Very close to zero	|30.69|
FAU 10	-	Very close to zero	-28.60
Phonetic (skewness)	-	Very close to zero	|14.58|
Facial landmarks 3D	z	Very close to zero	10.45

Table 6.4 Best Pearson correlation feature performances, ordered with respect to the strength of their correlation

Features	Axis	Pearson corr.
Facial landmarks 2D	y	-0.50878978 (av.)
PDM translation	y	-0.508386492
Facial landmarks 3D	y	-0.505845919 (av.)
Eye landmarks 2D	y	-0.504992323 (av.)
Eye landmarks 3D	y	-0.503182834 (av.)
Pose_Ty	y	-0.501908124

Table 6.3 and Table 6.4 present the PDM translation in the axis "y" as the feature most correlated to the AD. Followed by the 2D facial and eye landmarks in the axis "y".

6.3.3 FAU occurrences in silences

Finally, we analyzed the occurrences of each FAU on each period of time where there is a silence. The Figure 6.1 shows the results obtained, the FAU 04 "brow lowerer" is the one with more occurrences during the silence periods followed by FAU 10 "upper lip raiser".

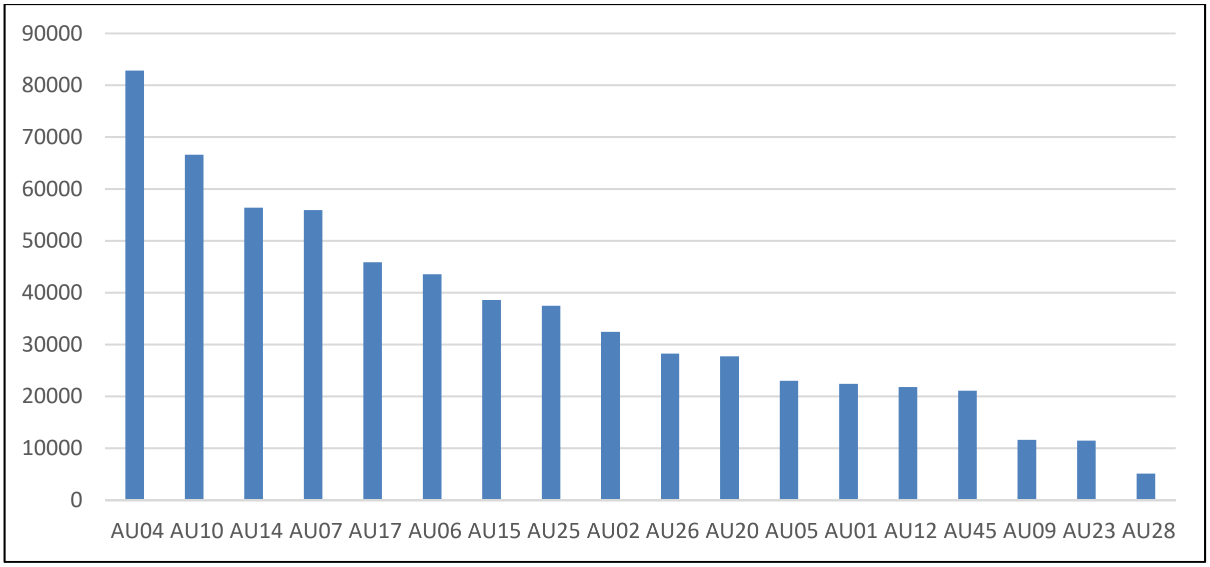

Figure 6.1 FAU incidences during silences

In order to analyze the relationship between silences and facial expressions, we consider also the analysis of the incidences of each FAU during silences and while subjects were speaking. The comparison of the results obtained is shown in Figure 6.2.

Finally, we performed experiments with the aim of evaluating facial expressions and silences in AD, we wanted to explore and characterize facial expression during a silence, and while subjects were speaking. We performed this in both groups of interest AD and HC in order to compare, and thus, will be able to analyze the relationship of silences, facial expressions and AD. The results compared are shown in Figure 6.3. Remarkably, FAU 04 and FAU 10 present the highest number of incidences in both, silence and dialog. Likewise, similar results were obtained for the AD and HC group, FAU 04 and FAU 10 present also the highest number of incidences.

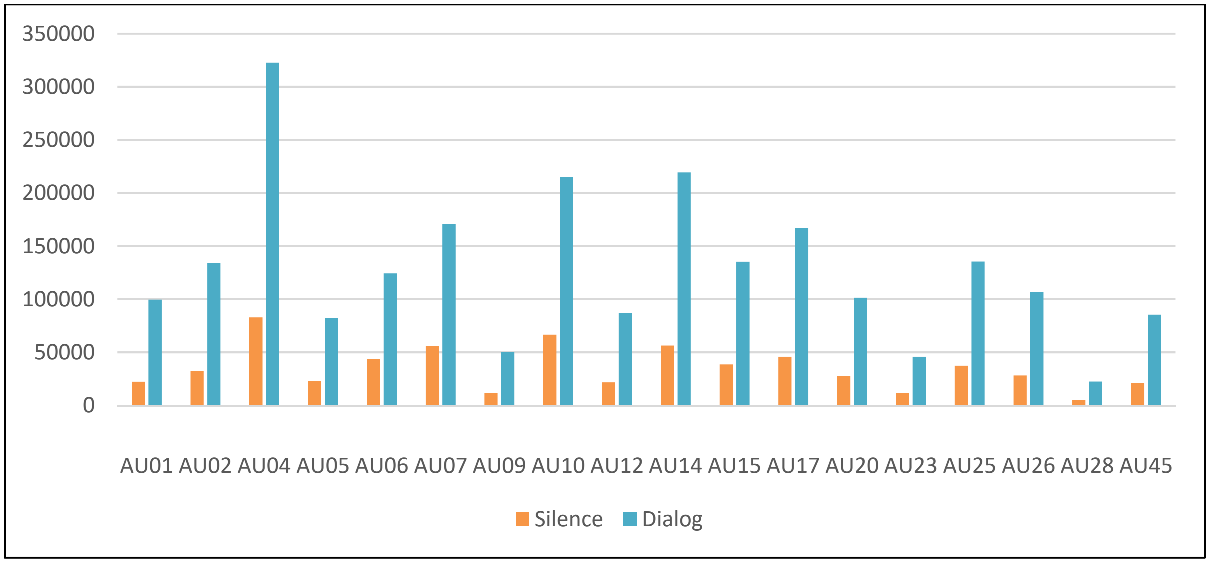

Figure 6.2 FAU incidences, comparison between incidences during silences and dialogs

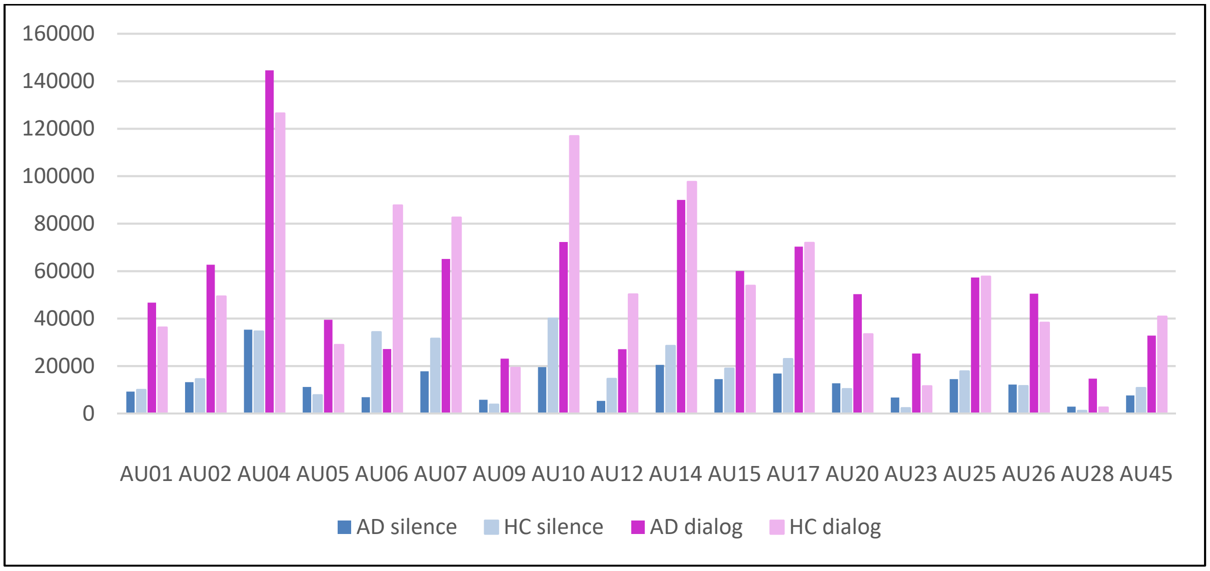

Figure 6.3 FAU incidences, comparison between incidences during silences and dialogs, group by AD and HC

6.4 Discussion

In this chapter, we analyzed all the types of features obtained in this thesis. Verbal and nonverbal communication features related were studied throughout this work. First, the features related to facial expressions: facial landmarks and gaze; second, phonetic and silence features.

Our best performance obtained is the KNN classifier trained with all the features with 91 % accuracy and 93 % ROC. Likewise, the RFC presented 87 % accuracy and 93 % ROC trained with all the features.

Additionally, we decided to include the SVM classifier in this chapter, due to the reduction of the data we did in the last chapter, SVM was suitable for being tested since the time of the training was reduced too. Therefore, we obtained an 89 % accuracy and 88 % ROC, trained with facial, silence, and phonetic MFCC features. Likewise, SVM trained with all the features obtained 87 % accuracy and 90 % ROC. In conclusion, SVM presents a satisfactory performance, however, KNN and RFC perform better than SVM.

On the other hand, the worst performance was the one obtained by the NB classifier. The NB classifier trained with gaze and silence performance practically randomly. The best performance for this classifier is the Gaussian implementation trained with all the features and trained with facial, gaze, and phonetic we obtained 76 % accuracy and 75 % ROC. These results show that the silence features in this classifier did not contribute to any relevant information for the process of training and performing classification.

With respect to the correlation analysis, several highlights were found. First, facial and gaze landmarks are more correlated in the "y" axis. The feature most correlated to AD is the PDM translation in axis "y". Second, the group of features describing the facial and eyes landmarks in 2D and 3D in axis "y" presented also a high correlation. The pose of the head in axis "y" was also strongly correlated. In general, the findings were consistent in Pearson, p-value, and T-statistic correlation measures.

As was expected, the number of FAUs incidences was higher during dialog compared to silence periods, all FAUs presented a higher number of incidences during dialog. Our results showed that FAU 4, 10, 14, and 7 presented the highest number of incidences with more than 6 x10^4, FAU 4 obtained more than 8 x 10^4 incidences.

Finally, unlike the training with the facial and phonetic features separately, in this case, where the classifiers are trained with all the features, the KNN performed better than the RFC.

CHAPTER 7

GENERAL DISCUSSION

In this thesis, we implemented three main experiments: first, we analyzed facial features in the context of natural conversations; second, we analyzed phonetic and silences, and third, we implemented multimodal experiments so we could compare the performance of each classifier in different configurations for training.

Of all the experiments carried out in this study, our best performance was obtained with the RFC trained with facial and eye landmarks, and gaze features (see Table 3.1, exp. (RFC, F9)). An important point to highlight is that we obtained a better performance exclusively with features related to nonverbal communication compared with any other combination of features performed in our experiments. Facial features performed even better than all the features together. This performance was obtained most probably as a result of the reduction of the data that had to be implemented in the phonetic analysis. We obtained 93% accuracy and 98% ROC training the RFC with facial features (see Table 3.1, exp. (RFC, F9)); on the other hand, we obtained 90% accuracy and 93% ROC training the KNN classifier with all, verbal and nonverbal features (see Table 5.1, exp. (KNN, M6)).

For the phonetic analysis, it was necessary to cut the sound every 3 seconds. Therefore, the videos were cut every 3 seconds too. Less than 3 seconds would tend to lose information related to pauses. Pauses could last more than one second approximately, thus, selecting a window shorter than 3 seconds could yield erroneous results like a pause counted as 2 small pauses. We recall that, videos were recorded as 30 frames per second, likewise, 90 frames had to be reduced at one span of time. Instead of having 90 data we reduced to 1 data point. For the above reasons, we believe that the best result was obtained with training exclusively with facial features.

Regarding the facial features (see Table 3.1), we note that the best on classifying was the RFC. At the same time, the best classifier over all the experiments carried out was also the RFC, its performance was satisfactory in most of the experiments. However, in general, the KNN classifier obtained satisfactory results too.

Regarding the FAU occurrences analyzed, we can mention several highlights.
As expected, the FAU occurrences are higher during dialog compared to silence periods for both groups AD and HC. All FAUs together have a higher number of occurrences in dialog than silences.

In general, the highest number of FAU occurrences obtained were FAU number 4, 10, and 14 in all, verbal and nonverbal approaches (consistent in all the experiments, facial, phonetic, and the combination of facial and phonetic features). FAU 4 describes the “brow lowerer” movement, this movement is common in elderly people, so our findings are consistent with what we expected. Regarding the FAU 10 and 14, which describes the “upper lip raiser” and “dimpler” movement could be facial expressions commonly occurring during natural conversations.

An interesting finding was the FAU 6 and 10, results showed that the number of occurrences of these FAU in the HC group was significantly higher than the number of occurrences in AD. In general, a proportion of 1:2 was obtained in the FAU 10 number of occurrences. For the FAU 6 (cheek raiser), we obtained a proportion of 1:3 (AD: HC), we can notice that the number of occurrences in HC is significantly higher than the occurrences in AD.

In general, the AD group seems to show less facial expressiveness according to the FAU analysis, refer to Figure 3.6 and Figure 3.7. In Figure 3.7, remarkably, the intensity of the expressions is significative higher in the HC group compared with the AD group in both, lower and upper face groups of FAUs. Likewise, FAUs, in the intensity detection in the lower face, showed more occurrences than the upper face in both groups: HC and AD.

For each FAU results presented a different number of incidences in the AD and HC groups. In the binary detection, FAUs numbers 6, 7, 45, 10, 12, 14, and 17 presented a higher number of incidences in the HC group compared to the AD group. On the other hand, in the intensity occurrences, the results presented a higher number of occurrences in the FAUs number 1, 2, 4, 6, 10, 12, 14, 15, 25, and 26. Likewise, just three occurrences were found higher in AD than HC group, these FAUs are the number 7, 17, and 23.

Additionally, we trained and tested the NB classifier with several combinations of features. In our experiments the worst performance obtained was the NB classifier with phonetic features with a 50 % accuracy with all the models tested Gaussian, Bernoulli, and Multimodal (see Table 4.2, exp. P1, P4, and P7), this result shows that the NB classifier performed by chance which in the context of the objectives of this thesis is not a satisfactory performance.

In the context of the confusion matrix analysis, the results obtained with accuracy, precision, recall, F1, and ROC shows the behavior of the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Likewise, we obtained the ROC curve measure for every classifier tested for several combinations of features for training. In the case of the RFC, SVM, and KNN we obtained a high value of ROC. As we know, ROC is the representation of the True Positive (TP) rate over the False Positive (FP) rate. Our results show that the TP rate was significative large and FP obtained was small. In the context of our problem, when an FP occurs means that the classifier has failed to distinguish between AD and HC, and a subject with HC has been classified as AD, likewise, a clinician could treat a healthy person for AD, wasting resources on a healthy person, and, in the worst case, clinicians could provide medication for AD treatment to a healthy individual. On the other hand, we obtained also a small number of False Negatives (FN), we could note this in every table where the performance of the classifiers was tested. In general, the recall statistic obtained shows the small number of FN in all the performances obtained. The recall is the TP rate over FN plus TP. If our results present values very close to 100 % it means that the number of FN is very small compared to the TP. Likewise, in the context of our problem, obtaining a FN means that a subject with AD was classified as HC. Consequently, the misclassified subject could not be receiving the proper treatment, and, as a consequence, AD could progress normally, losing the opportunity for the AD patient to retard the effects of the AD.

Likewise, the number of True Negatives (TN) in RFC, SVM, and KNN were significantly small. This small number of TN shows that our classifiers barely label an HC subject as AD. Phonetic features are useful, however, phonetic features did not improve the results obtained with training classifiers with exclusively facial features.

In the context of our objectives, we can say that our results are satisfactory since 98% of the time the classifier predicts correctly if a subject presents signs of AD or not. Thus, we successfully implemented an automatic tool that can predict with satisfactory results early signs of AD in a noninvasive way. We are aware and confident that it exists facial markers that indicates AD (already explained above). As we studied in the literature related, phonetic features are important, however, facial features could contribute quite good with biomarkers of AD.

CONCLUSION AND RECOMMENDATIONS

At this time, the Alzheimer is a neurodegenerative disease that has no cure. A specific diagnostic of AD exists only post mortem. Therefore, the join effort of the community research in this field focus on slowing down the disease progress, and at the same time, on allowing the patients to have a better quality of life. For this purpose, community research has been centered their efforts on methods to detect the disease on early stages so it will be possible to bring them an appropriate treatment.

The principal objective of this work is to implement a noninvasive method as a tool to help clinicians on detecting the AD in early stages. Specifically, this study implemented a non-invasive technique to automatically classify AD and HC with a multimodal analysis in natural conversations. During this thesis we published the article “Automatic analysis of Alzheimer’s disease, evaluation of eye movements in natural conversations” in the Alzheimer Association International Conference (AAIC) 2020 (Pérez-Arana & Ratté, 2020). We also published the article “Automatic evaluation of Alzheimer’s disease, analysis of facial expressions in natural conversations” in the “Journal of the Alzheimer disease reports”, article under revision (Pérez -Arana, Arlen; Ratté, Sylvie; Duong, Luc, 2021). Finally, we have the article “Multimodal analysis: the relationship between verbal and nonverbal communication in Alzheimer’s disease” ready for submission.

Likewise, the principal objective is divided in three:

First, the automatic classification through silences and phonetic features analysis. We extracted phonetic features from the audio of the videos. We tested four classifiers RFC, SVM, NB and KNN. In this case our best classifier was our KNN, this classifier could discern between AD and HC with 81% accuracy and 91% ROC, solely with silence and phonetic features.

Second, RFC could discern between AD and HC with a 93% accuracy and a 98% ROC trained with exclusively facial expressions and gaze.

Finally, in the multimodal analysis, we obtained the best performance training our KNN classifier with all the features (facial expressions, gaze, silences, and phonetic). We obtained 90% accuracy and 93% ROC, a better performance than the nonverbal features.

The best performance obtained of all the experiments performed in this thesis was the RFC trained with exclusively facial and gaze features in comparison with all the combinations of features tested in this thesis, even better than training any classifier tested in this thesis with all the features.

According to the above, our results obtained show that certain facial expressions and speech captured while the subject is speaking could provide us with important signs of AD in its early stages with considerable good accuracy.

The classifiers implemented were trained and tested with the Carolinas Conversation Collection (CCC) from Ecuador corpus with video recordings of natural conversations, including subjects from Latin America. In future work, we propose to extend the scope of our research to a corpus richer in diversity, i.e., a corpus that includes more ethnicities, different languages, or levels of education among the subjects. It is expected that with more diversity, we can obtain a classifier that performs better without having a strong correlation with the ethnicity or language of the subjects. Likewise, it is important to add more approaches like the linguistic feature analysis, there are several studies related to the analysis of linguistic characteristics and the relationship with AD. In this way, we will have a more complete analysis of the relationship between communication and AD.

In conclusion, we implemented a non-invasive technique with simply analyzing through machine learning techniques, facial expressions and acoustic features; we believe that this technique could be useful for clinicians in the early detection of AD. And thus, clinicians could proportionate opportune treatment to AD patients.

LIST OF BIBLIOGRAPHICAL REFERENCES

Abiven, F., & Ratté, S. (2021). Multilingual automation of transcript preprocessing in Alzheimer's disease detection. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, *7*(1), 2–7. https://doi.org/10.1002/trc2.12147

Allender, J., & Kaszniak, A. W. (1989). Processing of Emotional Cues in Patients with Dementia of the Alzheimer's Type. *International Journal of Neuroscience*, *46*(3–4), 147–155. https://doi.org/10.3109/00207458908986252

Angelopoulou, G., Kasselimis, D., Makrydakis, G., Varkanitsa, M., Roussos, P., Goutsos, D., Evdokimidis, I., & Potagas, C. (2018). Silent pauses in aphasia. *Neuropsychologia*, *114*(January 2017), 41–49. https://doi.org/10.1016/j.neuropsychologia.2018.04.006

Armstrong, R. A. (2008). Visual signs and symptoms of Parkinson's disease. *Clinical and Experimental Optometry*, *91*(2), 129–138. https://doi.org/10.1111/j.1444-0938.2007.00211.x

Armstrong, R. A. (2009). Alzheimer's disease and the eye. *Journal of Optometry*, *2*(3), 103–111. https://doi.org/10.3921/joptom.2009.103

Bachmann, J., Munzert, J., & Krüger, B. (2018). Neural underpinnings of the perception of emotional states derived from biological human motion: A review of neuroimaging research. *Frontiers in Psychology*, *9*(SEP), 1–12. https://doi.org/10.3389/fpsyg.2018.01763

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L. P. (2018). OpenFace 2.0: Facial behavior analysis toolkit. *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, 59–66. https://doi.org/10.1109/FG.2018.00019

Beattie, G. W., & Butterworth, B. L. (1979). Determinants of pauses and errors in spontaneous speech. *Language and Speech*, *22*(3), 201–211.

Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., & Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(12), 2930–2940. https://doi.org/10.1109/TPAMI.2013.23

Ben Ammar, R., & Ben Ayed, Y. (2019). Speech Processing for Early Alzheimer Disease Diagnosis: Machine Learning Based Approach. *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA*, *2018-Novem*. https://doi.org/10.1109/AICCSA.2018.8612831

Berger, C. R. (2014). Interpersonal communication. In *Interpersonal Communication*. https://doi.org/10.1515/9783110276794

Bschor, T., Kühl, K.-P., & Reischies, F. M. (2001). Spontaneous Speech of Patients With Dementia of the Alzheimer Type and Mild Cognitive Impairment. *International Psychogeriatrics*, *13*(3), 289–298. https://doi.org/10.1017/S1041610201007682

Buck, R., & VanLear, C. A. (2002). Verbal and Nonverbal Communication: Distinguishing Symbolic, Spontaneous, and Pseudo-Spontaneous Nonverbal Behavior. *Journal of Communication*, *52*(3), 522–541. https://doi.org/10.1111/j.1460-2466.2002.tb02560.x

Butterworth, B., & Beattie, G. (1978). Gesture and Silence as Indicators of Planning in Speech. In R. N. Campbell & P. T. Smith (Eds.), *Recent Advances in the Psychology of Language: Formal and Experimental Approaches* (pp. 347–360). Springer US. https://doi.org/10.1007/978-1-4684-2532-1_19

Campione, E., & Véronis, J. (2002). A large-scale multilingual study of pause duration. *Speech Prosody 2002. Proceedings of The1st International Conference on Speech Prosody*, 199–202. http://www.isca-speech.org/archive/sp2002/sp02_199.html

Caramelli, P., Mansur, L. L., & Nitrini, R. (1998). Chapter 32 - Language and Communication Disorders in Dementia of the Alzheimer Type. In B. Stemmer & H. A. Whitaker (Eds.), *Handbook of Neurolinguistics* (pp. 463–473). Academic Press. https://doi.org/https://doi.org/10.1016/B978-012666055-5/50036-8

Cera, M. L., Ortiz, K. Z., Bertolucci, P. H. F., & Minett, T. (2018). Phonetic and phonological aspects of speech in Alzheimer's disease. *Aphasiology*, *32*(1), 88–102. https://doi.org/10.1080/02687038.2017.1362687

Davidson, R. J., & Irwin, W. (1999). The functional neuroanatomy of emotion and affective style. *Trends in Cognitive Sciences*, *3*(1), 11–21. https://doi.org/10.1016/S1364-6613(98)01265-0

De Santi, L., Lanzafame, P., Spanò, B., D'Aleo, G., Bramanti, A., Bramanti, P., & Marino, S. (2011). Pursuit ocular movements in multiple sclerosis: A video-based eye-tracking study. *Neurological Sciences*, *32*(1), 67–71. https://doi.org/10.1007/s10072-010-0395-1

Demenescu, L. R., Mathiak, K. A., & Mathiak, K. (2014). Age- and Gender-Related Variations of Emotion Recognition in Pseudowords and Faces. *Experimental Aging Research*, *40*(2), 187–207. https://doi.org/10.1080/0361073X.2014.882210

Dongheng Li, Winfield, D., & Parkhurst, D. J. (2005). Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, *3*, 79–79. https://doi.org/10.1109/CVPR.2005.531

Dourado, M. C. N., Torres Mendonça de Melo Fádel, B., Simões Neto, J. P., Alves, G., & Alves, C. (2019). Facial Expression Recognition Patterns in Mild and Moderate

Alzheimer’s Disease. *Journal of Alzheimer’s Disease*, *69*, 539–549. https://doi.org/10.3233/JAD-181101

Ekman, P. (2002). Facial action coding system (FACS). *A Human Face*.

Ekman, P., & Friesen, W. V. (1976). Measuring facial movement. *Environmental Psychology and Nonverbal Behavior*, *1*(1), 56–75. https://doi.org/10.1007/BF01115465

Ekman, P., & Friesen, W. V. (1978). *Manual for the facial action coding system*. Consulting Psychologists Press.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., Andre, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, *7*(2), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

Fleisher, A., & Corey-Bloom, J. (2010). The natural history of Alzheimer's disease. *Dementia*, 405–416.

Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2015). Linguistic features identify Alzheimer’s disease in narrative speech. *Journal of Alzheimer’s Disease*, *49*(2), 407–422. https://doi.org/10.3233/JAD-150520

Fraser, K. C., Rudzicz, F., & Hirst, G. (2016). *Detecting late-life depression in Alzheimer’s disease through analysis of speech and language*. 1–11. https://doi.org/10.18653/v1/w16-0301

Freedman, M., Binns, M. A., Black, S. E., Murphy, C., & Stuss, D. T. (2013). Theory of mind and recognition of facial emotion in dementia: Challenge to current concepts. *Alzheimer Disease and Associated Disorders*, *27*(1), 56–61. https://doi.org/10.1097/WAD.0b013e31824ea5db

Gayraud, F., Lee, H.-R., & Barkat-Defradas, M. (2011). Syntactic and lexical context of pauses and hesitations in the discourse of Alzheimer patients and healthy elderly subjects. *Clinical Linguistics \& Phonetics*, *25*(3), 198–209. https://doi.org/10.3109/02699206.2010.521612

Gregory, C., Lough, S., Stone, V., Erzinclioglu, S., Martin, L., Baron-Cohen, S., & Hodges, J. R. (2002). Theory of mind in patients with frontal variant frontotemporal dementia and Alzheimer’s disease: Theoretical and practical implications. *Brain*, *125*(4), 752–764. https://doi.org/10.1093/brain/awf079

Gross, D. (1990). Communication and the Elderly. *Physical & Occupational Therapy In Geriatrics*, *9*(1), 49–64. https://doi.org/10.1080/J148V09N01_05

Haider, F., De La Fuente, S., & Luz, S. (2020). An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer's Dementia in Spontaneous Speech. *IEEE Journal on Selected Topics in Signal Processing*, *14*(2), 272–281. https://doi.org/10.1109/JSTSP.2019.2955022

Hargie, O., Saunders, C., & Dickson, D. (1981). Social skills in interpersonal communicationCroom Helm. *Beckenham, Kent*.

Haulcy, R., & Glass, J. (2021). Classifying Alzheimer's Disease Using Audio and Text-Based Representations of Speech. *Frontiers in Psychology*, *11*(January), 1–13. https://doi.org/10.3389/fpsyg.2020.624137

Hayashi, S., Terada, S., Takenoshita, S., Kawano, Y., Yabe, M., Imai, N., Horiuchi, M., Miki, T., Yokota, O., & Yamada, N. (2021). Facial expression recognition in mild cognitive impairment and dementia: is the preservation of happiness recognition hypothesis true? *Psychogeriatrics : The Official Journal of the Japanese Psychogeriatric Society*, *21*(1), 54–61. https://doi.org/10.1111/psyg.12622

Hernández-Domínguez, L., Ratté, S., Sierra-Martínez, G., & Roche-Bergua, A. (2018). Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, *10*, 260–268. https://doi.org/10.1016/j.dadm.2018.02.004

Jiskoot, L. C., Poos, J. M., Vollebergh, M. E., Franzen, S., van Hemmen, J., Papma, J. M., van Swieten, J. C., Kessels, R. P. C., & van den Berg, E. (2021). Emotion recognition of morphed facial expressions in presymptomatic and symptomatic frontotemporal dementia, and Alzheimer's dementia. *Journal of Neurology*, *268*(1), 102–113. https://doi.org/10.1007/s00415-020-10096-y

Keane, J., Calder, A. J., Hodges, J. R., & Young, A. W. (2002). Face and emotion processing in frontal variant frontotemporal dementia. *Neuropsychologia*, *40*(6), 655–665. https://doi.org/10.1016/S0028-3932(01)00156-7

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., Manera, V., Verhey, F., Aalten, P., Robert, P. H., & David, R. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, *1*(1), 112–124. https://doi.org/10.1016/j.dadm.2014.11.012

Kurzon, D. (1998). *Discourse of Silence*. John Benjamins Publishing Company.

Lagun, D., Manzanares, C., Zola, S. M., Buffalo, E. A., & Agichtein, E. (2011). Detecting cognitive impairment by eye movement analysis using automatic classification algorithms. *Journal of Neuroscience Methods*, *201*(1), 196–203.

https://doi.org/10.1016/j.jneumeth.2011.06.027

Le, V., Brandt, J., Lin, Z., Bourdev, L., & Huang, T. S. (2012). Interactive facial feature localization. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7574 LNCS*(PART 3), 679–692. https://doi.org/10.1007/978-3-642-33712-3_49

Lichtenstein-Vidne, L., Gabay, S., Cohen, N., & Henik, A. (2017). Lateralisation of emotions: evidence from pupil size measurement. *Cognition and Emotion*, *31*(4), 699–711. https://doi.org/10.1080/02699931.2016.1164668

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. *Proceedings of the IEEE International Conference on Computer Vision*, *2015 Inter*, 3730–3738. https://doi.org/10.1109/ICCV.2015.425

López-de-Ipiña, K., Alonso, J. B., Travieso, C. M., Solé-Casals, J., Egiraun, H., Faundez-Zanuy, M., Ezeiza, A., Barroso, N., Ecay-Torres, M., Martinez-Lage, P., & De Lizardui, U. M. (2013). On the selection of non-invasive methods based on speech analysis oriented to automatic Alzheimer disease diagnosis. *Sensors (Switzerland)*, *13*(5), 6730–6745. https://doi.org/10.3390/s130506730

Maki, Y., Yoshida, H., Yamaguchi, T., & Yamaguchi, H. (2013). Relative preservation of the recognition of positive facial expression “happiness” in Alzheimer disease. *International Psychogeriatrics*, *25*(1), 105–110. https://doi.org/DOI: 10.1017/S1041610212001482

Meghanani, A., C. S., A., & Ramakrishnan, A. G. (2021). An Exploration of Log-Mel Spectrogram and MFCC Features for Alzheimer’s Dementia Recognition from Spontaneous Speech. *2021 IEEE Spoken Language Technology Workshop, SLT 2021 - Proceedings*, 670–677. https://doi.org/10.1109/SLT48900.2021.9383491

Mueller, K. D., Koscik, R. L., Turkstra, L. S., Riedeman, S. K., LaRue, A., Clark, L. R., Hermann, B., Sager, M. A., & Johnson, S. C. (2016). Connected language in late middle-aged adults at risk for Alzheimer’s disease. In *Journal of Alzheimer’s Disease* (Vol. 54, Issue 4, pp. 1539–1550). IOS Press. https://doi.org/10.3233/JAD-160252

Natinal Aphasia Association. (2021). *Aphasia Definitions*. 15 July. https://www.aphasia.org/aphasia-definitions/

Perez, A., & Ratté, S. (2020). Automatic analysis of Alzheimer�s disease, evaluation of eye movements in natural conversations. *2020 Alzheimer’s Association International Conference*.

Pérez Arana, Arlen; Ratté, Sylvie; Duong, L. (2021). Automatic evaluation of Alzheimer’s disease, analysis of facial expressions in natural conversations. *Journal of the Alzheimer Disease Reports*.

Pistono, A., Pariente, J., Bézy, C., Lemesle, B., Le Men, J., & Jucla, M. (2019). What happens when nothing happens? An investigation of pauses as a compensatory mechanism in early Alzheimer's disease. *Neuropsychologia*, *124*(July 2018), 133–143. https://doi.org/10.1016/j.neuropsychologia.2018.12.018

Pope, C., & Davis, B. (2016). *Conversing with the elderly in Latin America : a new cohort for multimodal , multilingual longitudinal studies on aging*. *April*, 16–21. https://doi.org/10.18653/v1/W16-1903

Pope, C., & Davis, B. H. (2011). *Finding a balance : The Carolinas Conversation Collection*. *1*, 143–161.

Qiao, Y., Xie, X.-Y., Lin, G.-Z., Zou, Y., Chen, S.-D., Ren, R.-J., & Wang, G. (2020). Computer-Assisted Speech Analysis in Mild Cognitive Impairment and Alzheimer's Disease: A Pilot Study from Shanghai, China. *Journal of Alzheimer's Disease*, *75*, 211–221. https://doi.org/10.3233/JAD-191056

Rizzo, M., Anderson, S. W., Dawson, J., & Nawrot, M. (2000). Vision and cognition in Alzheimer's disease. *Neuropsychologia*, *38*(8), 1157–1169. https://doi.org/10.1016/S0028-3932(00)00023-3

Rochon, E., Leonard, C., & Goral, M. (2018). Speech and language production in Alzheimer's disease. *Aphasiology*, *32*(1), 1–3. https://doi.org/10.1080/02687038.2017.1390206

Rousseaux, M., Sève, A., Vallet, M., Pasquier, F., & Mackowiak-Cordoliani, M. A. (2010). An analysis of communication in conversation in patients with dementia. *Neuropsychologia*, *48*(13), 3884–3890. https://doi.org/10.1016/j.neuropsychologia.2010.09.026

Sajjadi, S. A., Patterson, K., Tomek, M., & Nestor, P. J. (2012). Abnormalities of connected speech in semantic dementia vs Alzheimer's disease. *Aphasiology*, *26*(6), 847–866. https://doi.org/10.1080/02687038.2012.654933

Satt, A., Hoory, R., König, A., Aalten, P., & Robert, P. H. (2014). Speech-based automatic and robust detection of very early dementia. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *September*, 2538–2542.

Schreiner, A. S., Yamamoto, E., & Shiotani, H. (2005). Positive affect among nursing home residents with Alzheimer's dementia: The effect of recreational activity. *Aging & Mental Health*, *9*(2), 129–134. https://doi.org/10.1080/13607860412331336841

Shimokawa, A., Yatomi, N., Anamizu, S., Torii, S., Isono, H., & Sugai, Y. (2003). Recognition of Facial Expressions and Emotional Situations in Patients with Dementia of the Alzheimer and Vascular Types. *Dementia and Geriatric Cognitive Disorders*, *15*(3), 163–

168. https://doi.org/10.1159/000068479

Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and L.-P. M. (2018). OpenFace 2.0: Facial Behavior Analysis Toolkit. *IEEE International Conference on Automatic Face and Gesture Recognition*.

Tadyla. (2014). *CLNF tracking on YouTube Celebrities dataset*. https://www.youtube.com/watch?v=vYOa8Pif5lY

Tomoeda, C. K., & Bayles, K. A. (1993). Longitudinal effects of Alzheimer disease on discourse production. In *Alzheimer Disease and Associated Disorders* (Vol. 7, Issue 4, pp. 223–236). Lippincott Williams & Wilkins.

Torre, F. De, Chu, W., Xiong, X., Vicente, F., Ding, X., & Cohn, J. (2015). *IntraFace*.

Torres, B., Santos, R. L., De Sousa, M. F. B., Neto, J. P. S., Nogueira, M. M. L., Belfort, T. T., Dias, R., & Dourado, M. C. N. (2015). Facial expression recognition in Alzheimer's disease: A longitudinal study. *Arquivos de Neuro-Psiquiatria*, *73*(5), 383–389. https://doi.org/10.1590/0004-282X20150009

Torres Mendonça De Melo Fádel, B., Santos De Carvalho, R. L., Belfort Almeida Dos Santos, T. T., & Dourado, M. C. N. (2019). Facial expression recognition in Alzheimer's disease: A systematic review. *Journal of Clinical and Experimental Neuropsychology*, *41*(2), 192–203. https://doi.org/10.1080/13803395.2018.1501001

Tóth, L., Gosztolya, G., Vincze, V., Hoffmann, I., Szatlóczki, G., Biró, E., Zsura, F., Pákáski, M., & Kálmán, J. (2015). Automatic detection of Mild cognitive impairment from spontaneous speech using ASR. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2015-Janua*, 2694–2698.

Toth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatloczki, G., Banreti, Z., Pakaski, M., & Kalman, J. (2017). A Speech Recognition-based Solution for the Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech. *Current Alzheimer Research*, *15*(2), 130–138. https://doi.org/10.2174/1567205014666171121114930

Vuorinen, E., Laine, M., & Rinne, J. (2000). Common Pattern of Language Impairment in Vascular Dementia and in Alzheimer Disease. *Alzheimer Disease & Associated Disorders*, *14*(2). https://journals.lww.com/alzheimerjournal/Fulltext/2000/04000/Common_Pattern_of_Language_Impairment_in_Vascular.5.aspx

Watanabe, M. (2016). [Emotional and Motivational Functions of the Prefrontal Cortex]. *Brain and nerve = Shinkei kenkyu no shinpo*, *68*(11), 1291–1299. https://doi.org/10.11477/mf.1416200593

Weiner, J., & Schultz, T. (2016). Detection of intra-personal development of cognitive impairment from conversational speech. *Speech Communication; 12. ITG Symposium*, 1–5.

Yang, P., Liu, Q., Metaxas, D. N., Ekman, P., Friesen, W. V., Davidson, R. J., Irwin, W., Davis, M., Whalen, P. J., Martins, A., Muresan, A., Justo, M., Simão, C., Kurzon, D., De Santi, L., Lanzafame, P., Spanò, B., D'Aleo, G., Bramanti, A., … Schröder, J. (2016). Emotional experience and facial expression in Alzheimer's disease. *Journal of Alzheimer's Disease*, *1*(1), 1–12. https://doi.org/10.1016/j.dadm.2014.11.012

Yang, S., Luo, P., Loy, C. C., & Tang, X. (2016). WIDER FACE: A face detection benchmark. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 5525–5533. https://doi.org/10.1109/CVPR.2016.596

Yeung, A., Iaboni, A., Rochon, E., Lavoie, M., Santiago, C., Yancheva, M., Novikova, J., Xu, M., Robin, J., Kaufman, L. D., & Mostafa, F. (2021). Correlating natural language processing and automated speech analysis with clinician assessment to quantify speech-language changes in mild cognitive impairment and Alzheimer's dementia. *Alzheimer's Research and Therapy*, *13*(1), 1–10. https://doi.org/10.1186/s13195-021-00848-x

Zadeh, A., Lim, Y. C., Baltrušaitis, T., & Morency, L. P. (2017). Convolutional experts constrained local model for 3D facial landmark detection. *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, *2018-Janua*, 2519–2528. https://doi.org/10.1109/ICCVW.2017.296

Zaitchik, D., Brownell, H., Winner, E., Koff, E., & Albert, M. (2006). Inference of beliefs and emotions in patients with Alzheimer's disease. *Neuropsychology*, *20*(1), 11–20. https://doi.org/10.1037/0894-4105.20.1.11

Zaitchik, D., Koff, E., Brownell, H., Winner, E., & Albert, M. (2004). Inference of mental states in patients with Alzheimer's disease. *Cognitive Neuropsychiatry*, *9*(4), 301–313. https://doi.org/10.1080/13546800344000246

Zargarbashi, S. S. H., & Babaali, B. (2019). *A Multi-Modal Feature Embedding Approach to Diagnose Alzheimer Disease from Spoken Language*. http://arxiv.org/abs/1910.00330

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, *23*(10), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342

Zheng, F., Zhang, G., & Song, Z. (2001). Comparison of different implementations of MFCC. *Journal of Computer Science and Technology*, *16*(6), 582–589. https://doi.org/10.1007/BF02943243