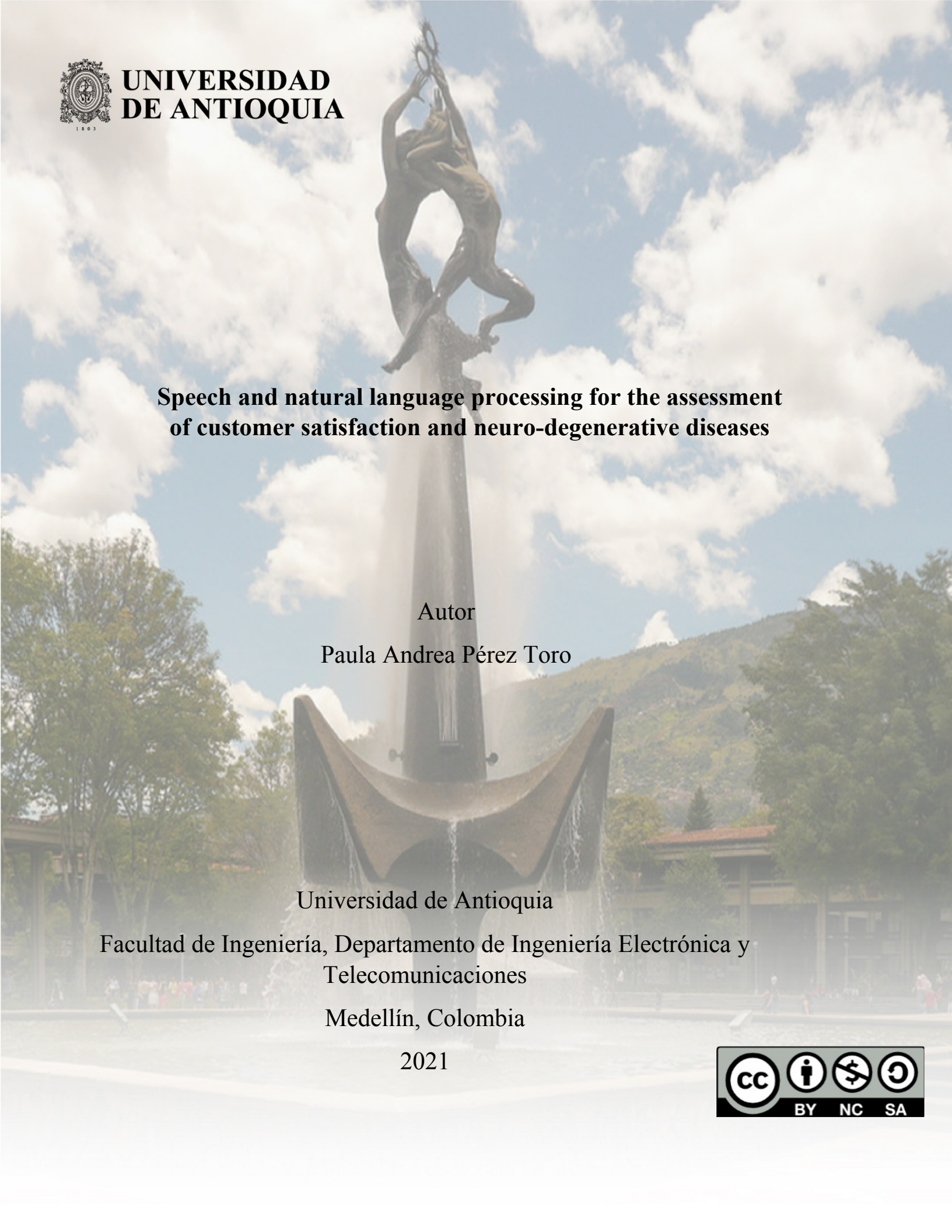

UNIVERSIDAD DE ANTIOQUIA

Speech and natural language processing for the assessment of customer satisfaction and neuro-degenerative diseases

Autor

Paula Andrea Pérez Toro

Universidad de Antioquia

Facultad de Ingeniería, Departamento de Ingeniería Electrónica y Telecomunicaciones

Medellín, Colombia

2021

Speech and natural language processing for the assessment of customer satisfaction and neuro-degenerative diseases

Paula Andrea Pérez Toro

Trabajo de investigación presentado como requisito para optar al título de:

Magíster en Ingeniería de Telecomunicaciones

Asesores (a):

Prof. Dr.-Ing. Juan Rafael Orozco Arroyave

Prof. Dr.-Ing. Elmar Nöth

Prof. Dr. Tobias Bocklet

Línea de Investigación:

Análisis de patrones

Grupo de Investigación:

Grupo de Investigación en Telecomunicaciones Aplicadas (GITA)

Universidad de Antioquia

Facultad de Ingeniería, Departamento de Ingeniería Electrónica y Telecomunicaciones

Medellín, Colombia

2021.

Speech and natural language processing for the assessment of customer satisfaction and neuro–degenerative diseases

Master's Thesis in Telecommunication Engineering

submitted
by

Paula Andrea Pérez-Toro

born 16.05.1994 in Itagüi, Colombia

Written at

Faculty of engineering
Department of Electronics and Telecommunications
University of Antioquia.

in Cooperation with the Lehrstuhl für Mustererkennung (Informatik 5),
Friedrich-Alexander-Universität Erlangen-Nürnberg.

Director: Prof. Dr. Ing. Juan Rafael Orozco-Arroyave

Co-Director: Prof. Dr. Ing. Elmar Nöth

Advisor: Prof. Dr. Tobias Bocklet

Started: 01.02.2019

Finished: 31.10.2020

Acknowledgments

First of all, I wish to express my gratitude to the directors of this study Prof. Dr.-Ing. Juan Rafael Orozco Arroyave and Prof. Dr.-Ing. Elmar Nöth, and also the co-advisor Prof. Dr. Tobias Bocklet. They supported me in this learning process, spending their time and dedication for making it possible. I am really proud to be a student of them, not only because they are successful academics but also for being excellent people. I want to express my deep gratitude to all the people that supported me from the start of this work and made it possible for me to accomplish several goals. I also want to thank the pattern recognition team of GITA in the University of Antioquia: Cristian Rios, Daniel Escobar, Luis Felipe Gómez, Felipe López, Felipe Parra, Cristian Rios, Guberney Muñeton, Reinel Castrillon, Helber Carvajal, Sebatian Roldán, among others, for supporting me in this process, giving me their friendship and their academic and personal support. Also, my friends Catalina Zapata, María Camila Mesa, Sebastian Arango, Juan David Robles, Nicanor Garcia, Mateo Lopéz, Andres Arango, Steven Martinez, Mateo Posada, and to "El combo de Metodología de la investigación". They urged me not to give up and helped me a lot in this step.

I want to express also my gratitude to all the people in the pattern recognition lab, especially the speech processing group, who supported me, gave me their friendship, helped me a lot in everything related to my master's degree, and always cheered me up: Philipp Klumpp, Camilo Vasquez, Tomas Arias, Sebastian Bayerl, and Martin Strauß. Also, I want to express my gratitude to two people from MAD lab Arne Kürderle and Nils Roth for their friendship and for supporting me in this process. Also to Prof. Korbinian Riedhammer that gave me the opportunity to keep going with my academic process in Germany. I would like to express my admiration, my deep gratitude, and love to my mother María Lucía Pérez Toro, she made that all of this was possible with her invaluable support, unconditional love, and understanding. She made her big effort for giving me the chances to study and she taught me to be a good person.

Finally, I would like to thank the University of Antioquia, the Friedrich Alexander Universität Erlangen-Nürnberg, "Grupo de Investigación en Telecomunicaciones Aplicadas" GITA and Pratech group in Colombia who supported my work these two years. To Bayerisches Hochschulzentrum für Lateinamerika-BAYLAT and Deutscher Akademischer Austauschdienst-DAAD, who gave me the chance to do a research stay in Germany and supported me also during this pandemic. Also, Fundalianza, a Parkinson's foundation, and "Grupo de Neurociencias de Antioquia" in Medellín helped me to obtain the necessary data to perform this research.

Abstract

Nowadays, the interest in the automatic analysis of speech and text in different scenarios have been increasing. Currently, acoustic analysis is frequently used to extract non-verbal information related to para-linguistic aspects such as articulation and prosody. The linguistic analysis focuses on capturing verbal information from written sources, which can be suitable to evaluate customer satisfaction, or in health-care applications to assess the state of patients under depression or other cognitive states. In the case of call-centers many of the speech recordings collected are related to the opinion of the customers in different industry sectors. Only a small proportion of these calls are evaluated, whereby these processes can be automated using acoustic and linguistic analysis. In the assessment of neuro-degenerative diseases such as Alzheimer's Disease (AD) and Parkinson's Disease (PD), the symptoms are progressive, directly linked to dementia, cognitive decline, and motor impairments. This implies a continuous evaluation of the neurological state since the patients become dependent and need intensive care, showing a decrease of the ability from individual activities of daily life. This thesis proposes methodologies for acoustic and linguistic analyses in different scenarios related to customer satisfaction, cognitive disorders in AD, and depression in PD. The experiments include the evaluation of customer satisfaction, the assessment of genetic AD, linguistic analysis to discriminate PD, depression assessment in PD, and user state modeling based on the arousal-plane for the evaluation of customer satisfaction, AD, and depression in PD. The acoustic features are mainly focused on articulation and prosody analyses, while linguistic features are based on natural language processing techniques. Deep learning approaches based on convolutional and recurrent neural networks are also considered in this thesis.

Contents

Chapter 1

Introduction

1.1 Motivation

Natural language is the most common system of symbols used by humans to create and communicate meaning, while speech is the physical production of sounds of spoken language. Natural Language Processing (NLP) aims to describe semantic, grammatical and syntactic aspects to address different problems related to the interaction between humans and machines. These methods are commonly used to assess the influence of verbal information in human interactions, language disorders, sentiment analysis, among others. Speech processing focuses on the analysis of non-verbal information related to para-linguistic aspects such as the articulatory, phonatory and prosodic system [1]. Currently, speech analysis is frequently used in applications related to automatic speech recognition, speaker verification, and pathological speech analysis. Although there are several works addressing acoustic and NLP methods, the automatic language understanding is a difficult task that requires a huge contextual knowledge due to the problems related to ambiguity, polysemy, sarcasm, or double sense. This thesis works towards improving the performance of the automatic language understanding and to reduce the impact of the aforementioned problems by using speech and natural language processing.

Additionally, speech and language analyses can be addressed in different applications that contain information related to emotions, mood, or affect. Research topics related to emotion modeling aim to discriminate cognitive processes of humans, such as memory, behavior, psychological hallmarks or decision-making [2]. For instance, customer satisfaction is linked to the heuristic affect, which is a mental shortcut that allows the decision-making and problem-solving in an efficient and faster way [3]. Patients with chronic diseases are often affected by mood swings due to psychiatric or physiological factors such as depression and those related to cognitive decline,

which decreases the quality of life of patients and in many cases increases symptoms, as in some neuro-degenerative diseases [4], [5].

This work proposes to extract features using classical acoustic and NLP methods, as well as concepts of emotions related to the arousal-valence plane [6]. The aim is to evaluate the suitability of using and combining acoustic and linguistic models to assess two different scenarios: (1) customer satisfaction, and (2) neuro-degenerative diseases

1.2 Context

Acoustic and linguistic analyses are carried out in this study to evaluate the customer satisfaction in call-centers, and to model communication and speech disorders in neuro-degenerative diseases.

1.2.1 Customer Satisfaction

Customer satisfaction is a scenario that makes use of technologies based on speech and language analyses. This is linked to the customer service, which corresponds to all actions related to provide a service, and it is the backbone of any business [7]. A large amount of collected information in the call-centers is related to the opinion of the customers, which is usually evaluated by conducting surveys at the end of the call. However, there are several limitations with this approach. For instance, the mildly dissatisfied or mildly satisfied customers often do not bother to take surveys. This behavior causes a bias in the satisfaction scores that would not reflect the real customer satisfaction based on taking only the extreme cases into consideration [8]. Since there is a large number of calls, only a small proportion of them (typically 2%) are evaluated. Customer satisfaction is evaluated from the speech recording of the customer opinion, whereby these processes can be automated using acoustic and linguistic analysis. The speech processing has shown good performance to characterize customer satisfaction levels in related applications [9], [10], as well as NLP methods [9], [11]. Thus, this thesis aims to assess the suitability of speech and language modeling to perform this task.

1.2.2 Neuro-degenerative Diseases

Another field of application in acoustic and linguistic analyses is the assessment of neuro-degenerative diseases. These are disorders of the nervous system that affect the neurons in the human brain. Particularly, neuro-degenerative diseases are progressive, having as predominating symptoms dementia, cognitive decline, and motor impairments. The loss of common

neurotransmitters and the death of the brain tissue cause communication deficits and emotional disturbances, associated with depressive signs, mood changes, sleep disorders, among others [12]. The two most common neuro-degenerative diseases, Alzheimer's Disease (AD) and Parkinson's Disease (PD), have symptoms related to the described disorders.

AD is the most prevalent neuro-degenerative disorder, and it is characterized by progressive dementia, degeneration, and death of the brain cells. AD symptoms include memory, behavioral, psychological impairments, and deterioration of cognitive functions related to communication deficits, i.e., the capability to produce coherent language [13]. AD is the most common form of dementia, affecting 2/3 of the total cases of dementia [14]. The total population of Colombia is approximately 50 million inhabitants, where there are around 221.000 people suffering from this disease [15]. The PSEN1-E280A or *Paisa mutation* is responsible for most of Early-Onset Alzheimer's (EOA) cases in Colombia. It is commonly diagnosed at a mean age of 49 years [16]. It affects a large kindred of over 5000 members that present the same phenotype. It is characterized by typical symptoms of the AD such as memory deficits in the third decade of life, development of progressive cognitive impairments such as verbal disfluency, changes in personality and behavior, among others [17].

In general, AD patients become dependent and need intensive care, showing a decrease of the ability from individual activities of daily life. The standard scales to evaluate the cognitive function of the patients are the Mini-Mental State Examination (MMSE) [18] and Montreal Cognitive Assessment (MoCA) [19], which are 30-point scales that contain items related to language production, immediate memory, naming, and spatial attention. Scores of over 24 and 26 indicate normal cognition for MMSE and MoCA respectively. Cognitive deficits and behavioral disorders also appear in AD. The patients tend to react aggressively, perceiving danger in common situation where none exists. These mood swings appear because of an alteration in perception of the reality. AD patients commonly present depression symptoms according to different studies [4], [20], [21]. According to [21] about 80% of AD patients can develop depression during the course of the disease. In addition, some studies suggest that the reduced ability to feel emotions is caused by the memory loss, which may induce the appearance of apathy and depression [22], [23]. Several studies have considered speech and NLP methods to assess the neurological state of AD patients [24]–[27].

PD is a neurological disorder characterized by the progressive loss of dopaminergic neurons in the mid brain [28], which are in charge to control movement and emotions. The most common motor symptoms are rigidity, bradikynesia, resting tremor, among others, which also affect the muscles involved in the speech production. Some of the voice impairments include pitch variation,

decreased loudness and hypokinetic dysarthria [29]. Non-motor complications also affect PD patients. The most common non-motor symptoms include sleep disorders, cognitive impairments, and depression [30]. If the depressive disturbances are not treated, they lead to other symptoms such as greater functional disability, faster physical and cognitive deterioration, increased caregiver distress, among others [31]. The depressive state of the PD patients is evaluated in a single item in the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [32]. Depressive symptoms in PD may experience fluctuation likewise as motor symptoms [33]. Prevalence of depression in PD varies from 20% to 50% and depressive symptoms are frequently associated with rapid progression of motor symptoms and cognitive impairments [5]. Communication deficits and impairments in the grammar production appear in 90% of the patients due to deficits in the production of the dopaminergic neurons [34]. Most of the studies in the literature have been focused on the speech analysis rather than in the language comprehension. Several studies suggest that, besides articulatory problems, impairments in grammar, verbal fluency, and semantic are also present in most of the patients [35], [36]. There are similar symptoms in both diseases. One of the most common symptoms is depression, which appears in up to 40% of patients with PD [37].

This work proposes to combine speech analysis and NLP methods to extract features from spontaneous speech recordings and their transcriptions. The challenge is to evaluate the suitability of these methods to assess neuro-degenerative diseases as well as mood states in the patients.

1.3 Hypothesis

Speech and natural language processing provide relevant information for the automatic assessment of the customer satisfaction, and neuro-degenerative diseases.

1.4 Objectives

1.4.1 General Objective

To design a methodology based on speech, natural language processing, and pattern recognition techniques in order to improve the performance in the automatic prediction of the customer satisfaction levels and monitoring the emotional and mood state of patients with neuro-degenerative diseases.

1.4.2 Specific Objectives

1. To analyze and evaluate the most suitable features, extracted from speech and natural language, to assess the customer satisfaction levels and mood states of the patients.

2. To design a robust algorithm to combine features from speech and language, to improve the performance when the modalities are applied individually.

3. To validate the performance of several feature sets considering the discrimination capability different classification tasks.

1.5 Contribution of this Study

This thesis proposes multimodal approach based on acoustic and linguistic analyses to assess two different scenarios: (1) customer satisfaction, and (2) neuro-degenerative diseases. The acoustic analysis includes classical features derived from the Short-Time Fourier Transform (STFT), Mel-Frequency Cepstral Coefficients (MFCCs), energy, duration, and the Fundamental Frequency (F_0). The linguistic approach includes classical features for inference such as Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), and word embedding methods such as Word2Vec (W2V), and Bidirectional Encoder Representations from Transformers (BERT). In addition, a deep learning based approach is proposed to extract information from the arousal-valence plane. The extracted features are used to evaluate different applications: (1) customer satisfaction analysis, (2) assessment of AD, and (3) depression in PD. All the approaches are compared with baseline features in the state-of-art. The algorithms were implemented in Python and Pytorch. The reported results in this study indicate that the combination of acoustic and linguistic information improved the performance considering different applications.

The following are the main outcomes of this thesis :

1. Collection of a multimodal corpus with speech, transcriptions, handwriting, and gait signals collected from PD patients since 2019, for different purposes by the GITA research group[1] in association with Fundalizanza[2].

2. A multimodal corpus of speech and transcriptions collected from 60 depressive and non-depressive PD patients in Colombia by the GITA research group[1] in association with Fundalizanza[2].

[1] https://gita.udea.edu.co/
[2] https://www.fundalianzaparkinson.org

3. A multimodal corpus of speech and transcriptions collected from different participants related to the *Paisa mutation* in Colombia by the GITA research group[1] and GNA research group[3].

4. An approach based on user modeling using acoustic and linguistic features together with a GMM-UBM models, for the assessment of depression in PD.

5. A novel approach to evaluate scenarios such as customer satisfaction and assessment of patients with neuro-degenerative diseases, using deep learning techniques and the arousal-plane information.

6. The development of WEBERT, which is a python toolkit designed for research purposes to automatically compute dynamic and static BERT embeddings based on the Hugging Face project[4]. WEBERT is available for English and Spanish (multilingual) models, as well as for base and large models, and cased and lower-cased options. BETO and SciBERT are also available here. BETO is a pre-trained BERT model from a Spanish corpus. SciBERT is a pre-trained model on English scientific text. The source code is available online[5].

7. I participated in the development of the mobile application Apkinson, which was developed to collect speech and movement data from PD patients, and to be used to monitor continuously the state of the patients using information from speech, hand movements, and fine-motor skills. The app was the main result of a Colombian-German project, financed by BMBF in which 16 young researchers from both countries participated. The source code is available online[6].

1.6 Outline

This work is divided into eight chapters. Chapter one contains the context, hypothesis, objectives and the contribution of this study. Chapter two includes the state-of-art. Chapter three describes the different methods addressed in this study. Chapter four contains the description of the different datasets. In Chapter five are presented the experiments, results and discussions. Finally, chapters six include the main conclusions, summary and further work.

[3] https://web.gna.org.co/en/us/
[4] https://github.com/huggingface/transformers
[5] https://github.com/PauPerezT/WEBERT
[6] https://github.com/jcvasquezc/SMA2

Chapter 2

State-of-the-art

2.1 Speech and Linguistic Analysis in Customer Satisfaction

Customer satisfaction is an important aspect for all industries. Automatic evaluation of customer satisfaction has been recently considered using speech and NLP strategies. In [11], the authors proposed an automated method to measure customer satisfaction by analyzing the transcriptions from 115 customer calls for a 2-point and 5-point satisfaction measurement. Phone calls were transcribed using the IBM Attila Speech Recognition toolkit [38]. The authors analyzed different aspects such as call duration, number of holds during the call, sentiments words, competitor mentions, talk speed and prosodic features, using the RapidMiner toolkit [39]. Different classifiers such as decision trees, logistic regression, Naïve Bayes, and Support Vector Machine (SVM) were implemented. The models were validated with a 10-fold cross-validation strategy. The authors reported accuracies with the SVM up to 89.42% in 2-point and 66.09% in 5-point.

In [40] the authors proposed the use of speech analysis in customer satisfaction for affective modeling according to the arousal-valence plane. The NTU_American dataset [41] was used in order to represent three emotions of anger, sadness, happiness and neutral serving as emotionless state. The analysis was performed based on identify the contribution of the neutral state in other emotional states. Acoustic features considered MFCCs, and for classification was considered an adaptive neuro-fuzzy inference system with substractive clustering. The authors reported results for measuring satisfaction of 40% accuracy, and neutral emotion showed higher performance with 58% of accuracy.

A text mining approach to explore customer satisfaction in airlines is presented in [42]. The authors performed an automatic customer satisfaction analysis in online customer reviews for airlines companies from the website Air Travel Review. NLP features based on Latent

Dirichlet Allocation (LDA) and part-of-speech tagging were obtained in order to identify different dimensions of satisfaction. The authors reported an accuracy of 80% for the two problem positive vs. negative reviews.

The combined analysis of speech recordings and their transcripts have shown promising results on different applications as well. In [9], the authors combined acoustic and language analysis to improve the performance on emotion recognition. The IEMOCAP database and data from a call-center were used in the experiments. The acoustic analysis was performed using Long Short Term-Memory (LSTM) networks trained with features extracted with OpenSMILE [43]. The language analysis was based on utterance-level embeddings generated from the transcriptions using multi-scale CNNs. Information from speech and text were merged and the final classification of emotions was performed with an SVM, which was optimized following a Leave-One-Speaker-Out (LOSO) cross validation strategy. The results indicated that the combination of acoustic features extracted from audio signals and language features extracted from their transcripts improve the accuracy up to 24% depending on the evaluated database.

Table 2.1: Summary of the state-of-the-art methods for customer satisfaction using acoustic and linguistic analysis

Related work	Acoustic Features	Linguistic Features	Datasets
Park, 2009 [44]	Call duration, number of holds during the call, talk speed	RapidMiner: Sentiment words, category-specific word, positive actitude and gratitude indicator, product name.	115 customer calls
Kamaruddin, 2016 [41]	MFCCs	-	NTU_American dataset
Lucini, 2020 [42]	-	Latent Dirichlet Allocation and part-of-speech tagging	Air Travel Reviews
Cho, 2018 [9]	OpenSMILE ComParE 2013: MFCCs, energy, F_0, entropy, zero crossing rate, shimmer, jitter, spectral-based features.	Word embeddings after one hot encodings input	DAIC-WOZ

State-of-the-art summary for customer satisfaction according to this work is shown in Table 2.1. Acoustic features mainly include MFCCs, energy, F_0, and prosody based features related with duration, among others. OpenSMILE has to be in consideration as baseline feature set for speech analysis. Linguistic features consider LDA, part-of-speech-tagging, and word embedding methods. In this approaches, indicators such as product name, competitor name, gratitude and positive answers take high relevance to perform customer satisfaction analysis.

2.2 Speech and Linguistic Analysis in Health-Care Applications

There are several studies focused on modeling speech deficits of PD patients related to cognitive decline [45]–[47]. In [47] the authors proposed a methodology to assess the cognitive decline of PD patients with a combination of clinical and acoustic features. The cognitive state of the patients was labeled with the Addenbrooke's Cognitive Examination (ACE-R) [48]. Data from 44 PD patients evaluated at baseline and two years after were considered for the analysis. The results indicated that the F_0 and the REM sleep behavioral disorder questionary (RBDSQ) explained 37.2% of the variability change of the ACE-R. In addition, the most correlated features with the cognitive decline were the disease duration, the speech index rhythmicity (SPIR), and the RBDSQ. The results of a linear regression showed that the SPIR was able to predict the cognitive decline with an accuracy up to 73.2%. There are studies that have considered NLP to evaluate PD patients. In [49], the authors extracted different components from transcriptions of spontaneous speech recordings. The authors used Latent Semantic Analysis (LSA), part-of-speech tagging and word-level repetitions via graph embedding tools to model the transcriptions. The study considered several classifiers including K-nearest neighbors, SVMs, Adaboost, and others, to discriminate between 50 PD patients and 50 HC subjects. Results of up to 77% of accuracy were reported. In [50] the authors aimed to predict motor, cognitive, and depressive symptoms of 35 PD patients, all of them English native speakers. The motor state was predicted based on the UPDRS score (UPDRS is a previous version of the MDS-UPDRS scale), and the global cognitive state was predicted based on the Montreal Cognitive Assessment (MoCA) scale. The depressive state was assessed based on the Geriatric Depression Scale (GDS). The clinical scales were predicted by computing articulation features such as formant frequencies and the derivatives of the MFCCs, and prosody features based on the phoneme rate. Acoustic features were used to train a Gaussian staircase regression algorithm to predict each neurological scale. The authors reported moderate Spearman's correlations in the prediction of the motor severity ($\rho = 0.42$) and global cognition ($\rho = 0.52$), however, the results on depression were not satisfactory ($\rho = -0.21$).

Cognitive deficits and behavioral disorders are more common in AD than in PD. Thus, besides the studies on PD, there are several works on AD where the impact of language impairments is studied [26], [51], [52]. In [52], the authors considered speech and language analyses to evaluate the depressive state of AD patients. The methodology was evaluated using Pitt Corpus from the Dementia Bank [53]. The language features include part-of-speech-tagging, parse constituents, psycho-linguistic measures, vocabulary richness, among others. On the other hand, acoustic

features considered such as fluency measures, MFCCs, voice quality features, energy intensity, shimmer, among others. The extracted features were used to classify HC subjects and AD patients with depression. The authors reported an accuracy up to 86.4% with an SVM classifier. The results classifying AD patients with and without depression achieved an accuracy up to 65.9%, using a logistic regression.

In [54], the authors investigated the relative changes in cognition and identification of non-verbal signals of emotion in AD patients. A group of 12 AD patients and 12 HC underwent facial and prosodic stimuli of five different emotions (happiness, sadness, anger, fear and neutral). The Florida Affect Battery (FAB) [55] was used to assess the emotional processing. An ANOVA was performed to quantify the difference between groups obtaining a significance value of 0.202 for the non-emotional prosody discrimination.

Recently in [27] the authors proposed an approach of counting word occurrences in transcriptions via BoW vectors. English transcriptions from the Pitt Corpus of the Dementia Bank [53] (168 AD patients and 94 HC subjects), were considered.The participants were asked to describe the cookie theft image [56]. BoW features were used to classify the AD patients and HC subjects using an artificial neural network. The authors followed a Leave-One-Speaker-Out (LOSO) cross-validation strategy, and reported accuracies up to 84.4%.

An n-gram based approach combined with LSTM cells is proposed in [57] to predict and classify AD patients and HC subjects. English transcriptions from the Pitt Corpus of the Dementia Bank [53] were considered. The participants were asked to describe the cookie theft picture [56]. The prediction of the MMSE score was proposed based on evaluating the perplexity of the transcriptions. The Spearman's correlation was higher for the AD ($\rho = 0.55$) than for HC subjects ($\rho = 0.11$). The classification between AD patients and HC subjects obtained 85.6% of overall accuracy.

In [58], the authors presented an automatic speech recognition based procedure for the extraction of a special set of acoustic features and a set of linguistic features extracted from transcripts of the same speech signals. The acoustic features were based on the Praat software [59] and linguistic features were based on the Magyarlanc toolkit [60]. The aim to discriminate between AD patients and healthy controls, and also AD patients from those with Mild Cognitive Impairments (MCI). The database for this purpose was recorded at the Memory Clinic at the Department of psychiatry of the University of Szeged, Hungary. This consists of 25 speakers for each group (75 speakers) and 225 recordings. The authors performed the classification followed a 4-fold cross-validation strategy using as classifier an SVM. The results showed that only using an acoustic-based feature set a high performance to classify various groups (accuracies ranging from

74% to 82%) was obtained. Similar accuracies were obtained using linguistic features. The fusion of the two set of features, the accuracy increases to 80-86%.

In Table 2.2 a summary of the state-of-the-art methods for neuro-degeneratives diseases is shown according to related work presented in this study. The predominant acoustic features consider F_0, Mel, duration, energy, formant frequencies. The linguistic features mostly are part-of-speech tagging, LSA, word frequency based features.

Table 2.2: Summary of the state-of-the-art methods for neurodegenerative diseases using acoustic and linguistic analysis

Parkinson's Disease			
Related work	**Acoustic Features**	**Linguistic Features**	**Datasets**
Rektorova, 2016 [47]	F_0, speech index rhythmicity	-	44 PD patients labeled according to ACER
Garcia, 2016 [49]	-	LSA, part-of-speech tagging, word level representations via graph embedding tools	PC-GITA
Smith, 2017 [50]	Formant frequency, MFCCs (static, delta), phoneme rate	-	35 PD patients
Alzheimer's Disease			
Fraser, 2016 [52]	MFCCs, voice quality features, energy, shimmer, jitter	Part-of-speech-tagging, parse constituents, psycho-linguistic measures, vocabulary richness	Pitt corpus from Dementia Bank
Buck, 2004 [54]	Prosody test conducted using Florida Affect Battery	-	12 AD patients and 12 HC subjects
Klumpp, 2018 [27]	-	Bag of words	Pitt corpus from Dementia Bank
Fritsch, 2019 [57]	-	N-Grams	Pitt corpus from Dementia Bank
Gosztolya, 2019 [58]	Articulation rate, speech tempo, utterance length, duration of silents, number of silence, hesitation rate, number of recurrences of a given phoneme	Part-of-speech-tagging, number and rate of words/phrases related to memory activity, number of negation words, number of of thematic words related to the content of the films.	75 Hungarian native speakers

Chapter 3

Theoretical Background

The proposed models are based on different acoustic and linguistic features to discriminate emotions, pathological mood and cognitive states. This chapter consists of four main sections: (1) pattern recognition and deep learning methods, in which is explained the basis of the performed methods for end-to-end analysis, (2) emotions modeling by the arousal-valence plane that contextualize the concept of emotion in the mentioned plane, (3) speech analysis methods consisting of all signal processing techniques performed in this study, and (4) linguistic analysis methods, where the NLP algorithms considered for the linguistic approach are explained.

3.1 Pattern Recognition and Deep Learning Methods

3.1.1 Pattern Recognition Methods

Principal Component Analysis

This is an unsupervised method that allows to simplify the spatial complexity of a feature space of dimensionality d onto a subspace of a lower dimension p. Principal Component Analysis (PCA) aims to find a linear projection that maximizes the variance in the data, obtaining non-correlated components [61].

Such a projection is defined in Equation 3.1, where $\boldsymbol{W}$ is a $d \times p$ eigenvector matrix used to transform the samples onto the new subspace, and $\boldsymbol{x}$ is a $d \times m$ dimensional vector with m as the number of samples.

$$\phi(\boldsymbol{x}) = \boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{x} \tag{3.1}$$

Figure 3.1 shows how PCA aims to find the line that projects the data in the direction of the maximum variance. PCA finds a new set of dimensions $\mathbb{R}^d \rightarrow \mathbb{R}^p$, such that all the p dimensions are orthogonal and ranked according to the variance of data along them. Thus, the eigenvector matrix $\boldsymbol{W}$ is computed in order to maximize the spread of the data, i.e., the variance. First, the d-dimensional mean vector μ is computed for every dimension of the whole dataset:

$$\mu = \frac{1}{m}\sum_{n=1}^{m} \boldsymbol{x}_n \tag{3.2}$$

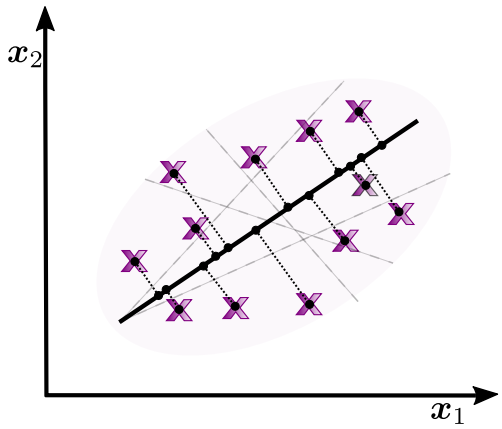

Figure 3.1: PCA projection (straight black line) with the maximum variance

Then, the covariance matrix Σ is computed using Equation 3.3. The covariance measures the total variation of two random variables, although is similar to the definition of correlation, in the covariance the values are not standardized.

$$\Sigma = \frac{1}{m-1}\sum_{n=1}^{m}(\boldsymbol{x}_n - \mu)(\boldsymbol{x}_n - \mu)^T \tag{3.3}$$

From the covariance matrix, the eigenvectors ν and eigenvalues λ are obtained using Equation 3.4. An eigenvector is a vector whose direction remains unchanged when a linear transformation is applied. The eigenvalues are defined as constants that multiply the eigenvectors in the linear transformations of a matrix, i.e., roots of the characteristic equation.

$$\Sigma\nu = \lambda\nu \tag{3.4}$$

The eigenvalue and eigenvector problem can be solved using singular value decomposition that projects data into a space of lower-dimensions preserving most of the variance, by releasing the singular vector of components associated with lower singular values. Then, the eigenvectors are

sorted in descending order according to their eigenvalues. The eigenvector defines the direction of the new axis as is shown in Figure 3.2, where it can be observed the two first eigenvectors.

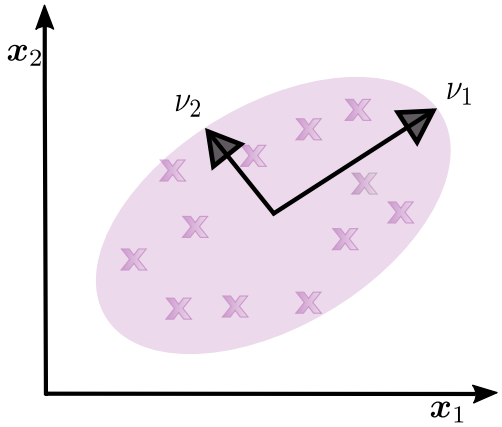

Figure 3.2: Directions of the eigenvectors that will define the new axis

The eigenvectors with the lower eigenvalues are ignored because it is assumed that those provide less information about the data distribution. The first p eigenvectors after sorting are used to construct $\boldsymbol{W}$ which is a $d \times p$ matrix. Finally, as is shown in Equation 3.1, the resulting eigenvector matrix is multiplied with the original features in order to project the data in a lower dimensional space. The p dimensional transformed features that now are uncorrelated, are known as Principal Components (PCs) that can be observed in Figure 3.3, where the new axes are these PCs and the data suffered a rotation to be adapted to these new axes.

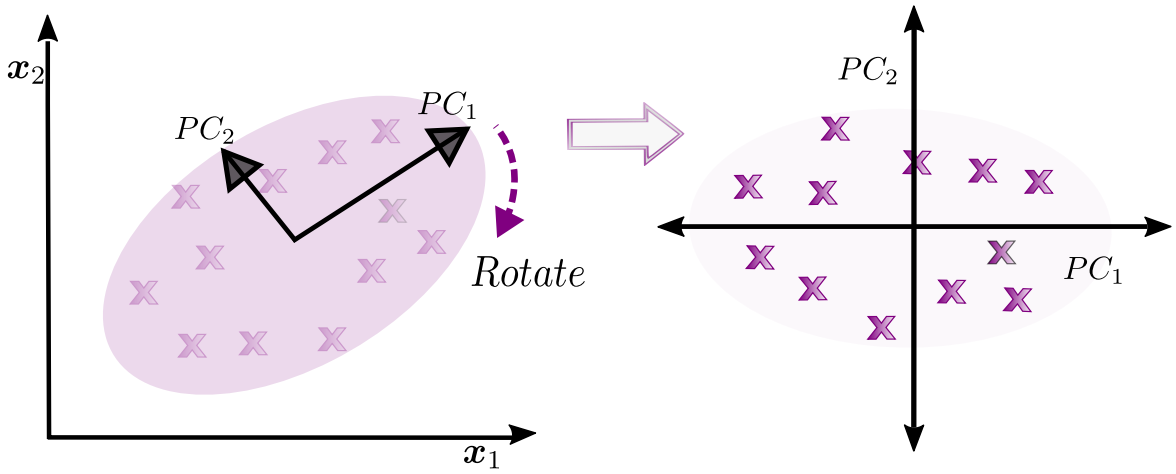

Figure 3.3: Transformed features with PCA

Support Vector Machines

SVM is a supervised classification algorithm proposed by Vapnik in [62]. The aim of this algorithm is to find the optimal hyperplane which maximizes the margin or width according to a training set S of m training samples as is shown in Equation 3.5. Each point $\boldsymbol{x}_i \, \epsilon \, \mathbb{R}^d$ belongs to a bi-class problem, thus a label is assigned for each one $y_i; \epsilon \, \Big\{-1, 1\Big\}$.

$$S = \Big\{\boldsymbol{x}_i, y_i\Big\} \;, i = 1, 2, ..., m \tag{3.5}$$

The hyperplane $\mathcal{H}$ defines the decision boundary of the SVM. In the simplest case it depends on a linear function expressed according to Equation 3.6, where $\boldsymbol{W}$ is the weight or perpendicular vector to the hyperplane, and b is the intercept of the line.

$$\mathcal{H} = \boldsymbol{W}^T \boldsymbol{x}_i + b \tag{3.6}$$

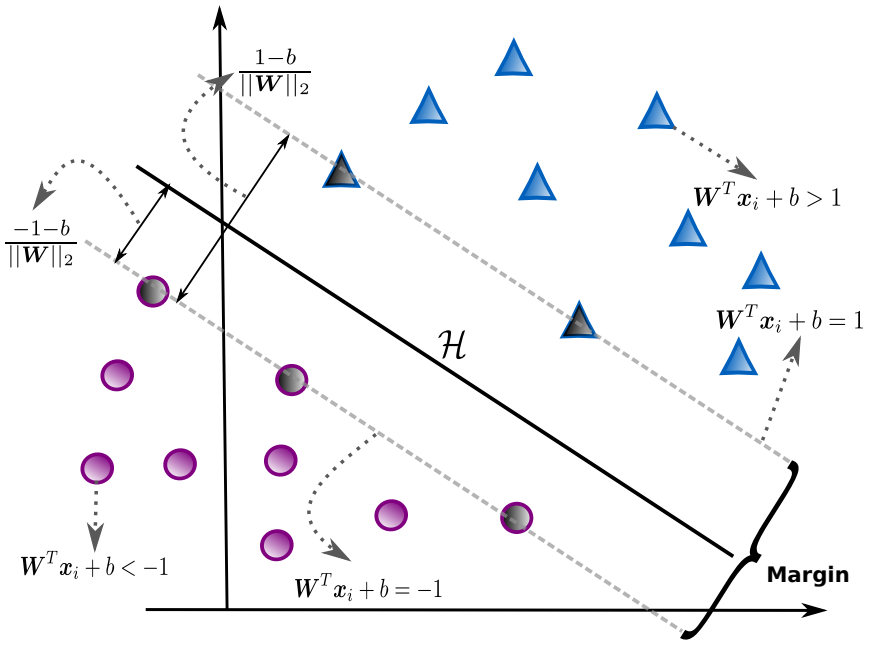

Figure 3.4: Best fitting hyperplane for the example training set S

As illustrated in Figure 3.4, the SVM predicts a point as positive target $y_i = +1$ when $\boldsymbol{W}^T\boldsymbol{x}_i - b > +1$, and as negative target $y_i = -1$ when $\boldsymbol{W}^T\boldsymbol{x}_i - b < -1$. Note that it only considers a perfectly linearly separable problem, i.e., that no point will be inside the margin. This SVM case is called hard margin. The goal is to find a $\boldsymbol{W}$ that maximizes the margin in Equation 3.7, which is equivalent to get a quadratic minimization problem according to

Equation 3.8.

$$margin = \frac{(1-b)}{||\boldsymbol{W}||_2} - \frac{(-1-b)}{||\boldsymbol{W}||_2} = \frac{2}{||\boldsymbol{W}||_2} \tag{3.7}$$

$$\underset{\boldsymbol{W}}{argmax}\ \frac{2}{||\boldsymbol{W}||_2} \equiv \underset{\boldsymbol{W}}{argmin}\frac{1}{2}||\boldsymbol{W}||_2^2 \tag{3.8}$$

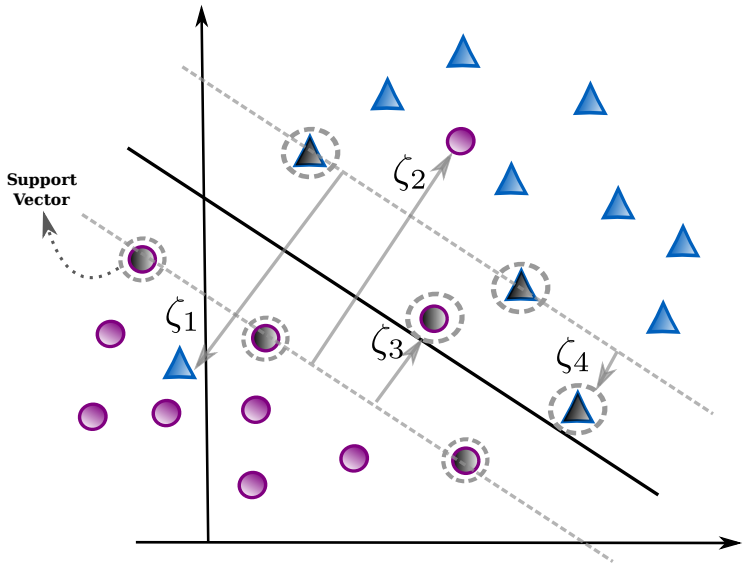

Figure 3.5: Best fitting hyperplane for SM-SVM

Hence, the data in this study cannot provide a perfectly linearly separable problem, another SVM case known as soft margin has to be considered. Soft Margin SVM (SM-SVM) introduces a cost in the objective function to penalize errors as is shown in Figure 3.5. A convex optimization problem is derived, where the primal objective function is represented by Equation 3.9, where such cost is introduced by using positive slack variables $\zeta_i \geq 0, \quad i = 1, 2, \ldots, m$. $C \geq 0$ is an adjustment parameter that controls how much is the penalty for the misclassified point. A larger C means a higher penalty to the errors and a small margin.

$$\underset{\boldsymbol{W}}{argmin}\ \frac{1}{2}||\boldsymbol{W}||_2^2 + C\sum_{i=1}^{m}\zeta_i \tag{3.9}$$

$\zeta_i = 0$ means that data points on the correct side and outside of the margin. Otherwise if a data point is on the decision boundary ($y_i = 0$), $\zeta_i = 1$, and when $\zeta_i > 1$, it will be misclassified. Then, these relaxed classification constrains will be replaced as in Equation 3.10.

$$\begin{aligned} y_i = +1 \quad &\rightarrow \boldsymbol{W}^T \boldsymbol{x}_i + b \geq 1 - \zeta_i \\ y_i = -1 \quad &\rightarrow \boldsymbol{W}^T \boldsymbol{x}_i + b \geq -1 + \zeta_i \end{aligned} \tag{3.10}$$

The optimization problem solution is solved by using Lagrange multipliers. The objective function for Lagrange primal optimization problem $\mathcal{L}$ in Equation 3.11, adds the associated Lagrange multipliers $\alpha_i \geq 0$, $i = 1, 2, 3, \ldots, m$. η_i is chosen to guarantee $\zeta_i \geq 0$.

$$\begin{aligned} \mathcal{L}(\boldsymbol{W}, b, \alpha_i, \eta_i, \zeta_i) &= \frac{1}{2}||\boldsymbol{W}||_2^2 + C\sum_{i=1}^{m}\zeta_i + \sum_{i=1}^{m}\alpha_i \underbrace{[-y_i(\boldsymbol{W}^T\boldsymbol{x}_i + b) + 1 - \zeta_i]}_{\leq 0} + \eta_i \sum_{i=1}^{m}\underbrace{(-\zeta_i)}_{\leq 0} \\ &= \frac{1}{2}||\boldsymbol{W}||_2^2 + C\sum_{i=1}^{m}\zeta_i - \sum_{i=1}^{m}\alpha_i[y_i(\boldsymbol{W}^T\boldsymbol{x}_i + b) - 1 + \zeta_i] - \eta_i\sum_{i=1}^{m}(\zeta_i) \end{aligned} \tag{3.11}$$

The constrained optimization is performed using the Karush Kuhn Tucker (KKT) conditions, which are given by primal, dual, and complementary slackness restrictions:

- Primal restrictions.

$$\begin{aligned} y_i(\boldsymbol{W}^T\boldsymbol{x}_i + b) - 1 + \zeta_i \geq 0 \\ \zeta_i \geq 0 \end{aligned} \tag{3.12}$$

- Dual restrictions.

$$\begin{aligned} \alpha_i \geq 0, \quad i = 1, 2, 3, \ldots, m \\ \zeta_i \geq 0, \quad i = 1, 2, 3, \ldots, m \end{aligned} \tag{3.13}$$

- Complementary slackness restrictions.

$$\begin{aligned} \alpha_i[y_i(\boldsymbol{W}^T\boldsymbol{x}_i + b) - 1 + \zeta_i] = 0 \\ \eta_i\zeta_i = 0 \end{aligned} \tag{3.14}$$

Then, the set of derivatives of the Lagragian respect to $\boldsymbol{W}$, b and ζ_i are zero.

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}} = 0 \quad \Rightarrow \quad \boldsymbol{W} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i \qquad s.t. \;\; \alpha_i \geq 0 \tag{3.15}$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \quad \Rightarrow \quad \sum_{i=1}^{m} \alpha_i y_i = 0 \qquad s.t. \;\; \alpha_i \geq 0 \tag{3.16}$$

$$\frac{\partial \mathcal{L}}{\partial \zeta_i} = 0 \quad \Rightarrow \quad C - \alpha_i - \eta_i = 0 \qquad s.t. \;\; \eta_i \geq 0, \;\; \alpha_i \geq 0 \tag{3.17}$$

There are additional constraints according to Equation 3.17 with respect to α_i, subject to:

$$0 \leq \alpha_i \leq C \Rightarrow \text{Box Constraints} \tag{3.18}$$

The data points enclosed with the dotted lien are the support vectors (see Figure 3.5). Those points define the decision boundary, i.e., $\alpha_i \geq 0$. Additionally, notice that:

- A subset of data points do not contribute to the model when $\alpha_i = 0$
- α has to be greater than zero to satisfy that $\boldsymbol{W}^T \boldsymbol{x}_i + b = 1 - \zeta_i$.
- The points lying on the margin occur when $\zeta_i = 0$. This implies that $\alpha_i < C$ and $\eta_i > 0$.
- Considering $\alpha_i = C$, the points inside the margin are misclassified if $\zeta_i > 1$, or correctly classified if $\zeta \leq 1$.

Thus, the intercept or independent term b in Equation 3.19, can be found when $\zeta_i = 0$, the data points are support vectors, and considering the box constraint.

$$y_i(\boldsymbol{W}^T \boldsymbol{x}_i + b) - 1 + \zeta_i = 0 \qquad b = \frac{1}{y_i} - \boldsymbol{W}^T \boldsymbol{x}_i \tag{3.19}$$

Finally, replacing the results and constraints into the primal Lagrange function, the wolfe dual problem is found as in Equation 3.20.

$$\mathcal{L} = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j$$
$$s.t.\ \ 0 \leq \alpha_i \leq C \tag{3.20}$$

The previous description is only to discriminate between two classes, however, the SVM can be adapted to solve multi-class classification problems. The multi-class classification in this study is performed by a method called ***"one-vs-the-rest"*** (OVR), that consists of fitting a classifier per class. The initial approach of OVR requires certain unanimity between all SVMs, i.e., a data point could be classified if and only if this SVM's class is accepted and the others are rejected. An advantage of this model is its interpretability since is possible to obtain some knowledge about the class inspecting its classifier.

Another consideration to take into account is that all the data dimensions are not linearly separable. In the SVM classifier, a kernel function is considered to transform the feature space into another that will be linearly separable. In this work, the kernel function is defined by $k(\boldsymbol{x}, \boldsymbol{x'}) = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x'}) \rangle$.

Kernel Trick

The kernel trick is a method that transforms a non-linear classification problem into a linearly separable [63]. It maps the original feature space into another space of higher dimensionality. The kernel function is defined as feature map $\phi(\boldsymbol{x})$ (basis function), which satisfied the Equation 3.21.

$$k(\boldsymbol{x}, \boldsymbol{x'}) = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x'}) \rangle \tag{3.21}$$

The kernel function corresponds to an scalar product in some feature space. This function is defined by $k : \mathbb{R}^d \times \mathbb{R}^d$, where $\mathbb{R}$ is symmetric and positive semi-definite.

In this study a kernel function called *"Radial Basis Function"* (RBF) is considered. This function is defined by Equation 3.22, where γ establishes the width of the bell-shaped curve. RBF has the property that each basis function depends only on the radial distance, most commonly used the Euclidean distance. Notice that the RBF kernel has a ready interpretation as a similarity

measure since it decreases with distance and ranges between zero and one.

$$k(\boldsymbol{x}, \boldsymbol{x'}) = \exp - \frac{||\boldsymbol{x} - \boldsymbol{x'}||^2}{2\gamma^2} \tag{3.22}$$

Random Forest

Random Forest (RF) is a supervised learning algorithm characterized by its precision and robustness against noisy data [64]. This algorithm consists of a set of individual decision trees from the randomly selected subset of the training set. Each tree contributes with a single vote in order to predict the most frequently class as is shown in Figure 3.6. If there are d input features, a number $n < d$ is specified such that at each node, n features are selected at random out of the d, i.e., n is the number of trees. The number of features in each subset is define as $s = d/n$. The large number of relatively uncorrelated models or trees can produce ensemble predictions that are more accurate than any of the individual predictions. The trees protect each other from their individual errors. In cases where some trees may be wrong, other trees may be right. Thus, it uses a combination of features at each node to grow a tree, instead of using the best variables, which reduces generalized error.

RF is actually an Ensemble-Bagging algorithm that generates random bootstrap samples from the training set. The main difference is that in RF selects only subset of features for training the individual trees, while in Bagging each tree is provided with the full set of features. The random feature selection allows the trees to be more independent of each other compared with regular Bagging, further, it is computationally faster because each tree learns only from a subset of features.

One of its advantages is that it does not suffer of over-fitting problems because RF takes the average of all predictions, which cancels out biases. The individual trees are generated by using an indicator of attribute selection. The most common in RF is the Gini Index also known as *"the Total Decrease in Node Impurity"*. The Gini Index considers a binary split for each attribute in order to compute the relative importance of the feature. This index which satisfies Equation 3.23, measures the impurity of a given element with respect to the other classes. T defines the training set, C_i is the class and $f(C_i, \boldsymbol{T})/|\boldsymbol{T}|$ is the probability that the selected case belongs to C_i. The more it decreases, the more important is the feature, i.e., that the mean decrease is an important parameter for feature selection. Thus, the forest is made to grow up until their maximum depth by using a given combination of features.

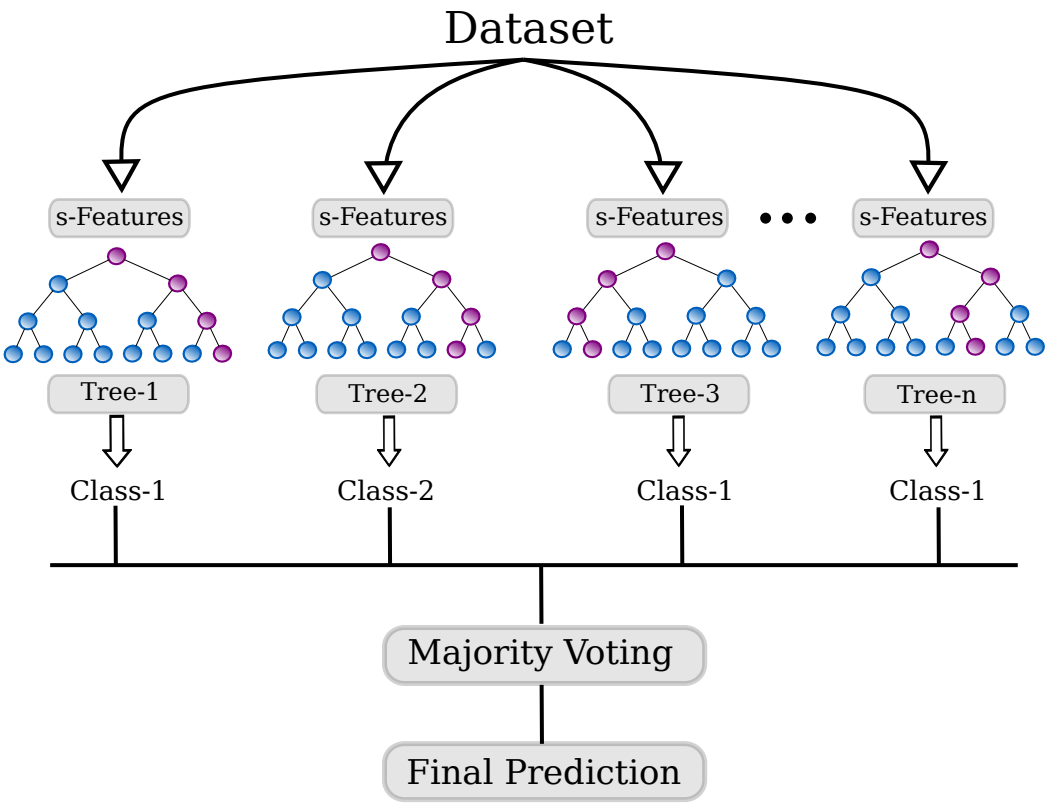

Figure 3.6: Architecture of the random forest model

$$\sum \sum_{j \neq i} (f(C_i, \boldsymbol{T})/|\boldsymbol{T}|)(f(C_j, \boldsymbol{T})/|\boldsymbol{T}|) \tag{3.23}$$

Finally, the RF algorithm is performed by the following steps:

1. Random samples from the dataset are selected.

2. A decision tree for each sample (Tree-n) is constructed. Thus, the prediction from every decision tree is obtained.

3. Each tree will grow to its maximum extension without pruning in which the features form each node. The maximum extension or depth is manually set and optimized.

4. The final prediction is performed by majority voting of all individual trees.

Gaussian Mixture Model-Universal Background Model

Gaussian Mixture Model-Universal Background Model (GMM-UBM) was first proposed by Reynolds in [65]. This model is commonly used in topics related to speaker verification. In this thesis the aim is to find differences between PD patients with depression (D-PD) and PD patients without depression (ND-PD), and as reference point an UBM composed by HC subjects. The algorithm consists of three main steps: (1) the Gaussian Mixture Model (GMM), (2) the Universal Background Models (UBM), and (3) Adaptation of the GMM with respect to the UBM.

Gaussian Mixture Model

GMMs are parametric models capable of representing a probability densities as a weighted sum of M Gaussian distributions (see Figure 3.7).

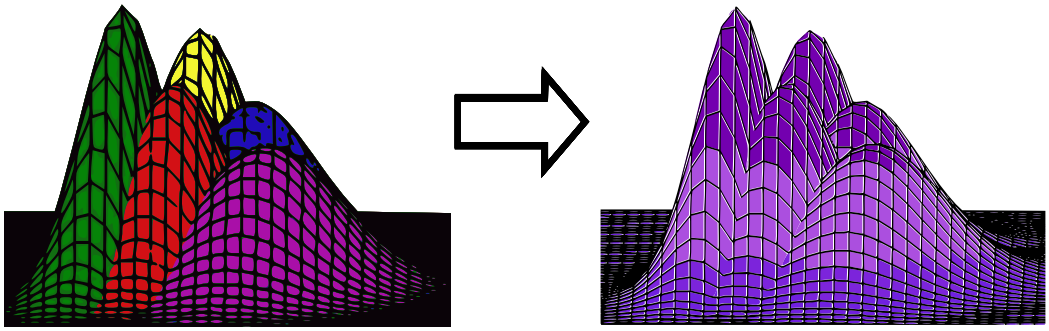

Figure 3.7: Gaussian Mixture Model representation

Let $\boldsymbol{X}$ be a set of d-dimensional feature vectors $\boldsymbol{x_i}$, the mixture density used for the likelihood estimation is given by Equation 3.24.

$$P(\boldsymbol{x_i}|\boldsymbol{\Theta}) = \sum_{k=1}^{M} \omega_k P_k(\boldsymbol{x_i}|\boldsymbol{\theta_k}) \tag{3.24}$$

ω_k is the weight of the mixture, M is the number of Gaussian components or clusters in the $\boldsymbol{\Theta}$ model. $\boldsymbol{x_i}$ is the i-th feature vector of $\boldsymbol{X}$. $\boldsymbol{\theta_k}$ is the set of Gaussian parameters given by the mean $d \times 1$ vector $\boldsymbol{\mu_k}$, and the covariance $d \times d$ matrix $\boldsymbol{\Sigma_k}$. P_k is the Probability Density Function (PDF), which is estimated in Equation 3.25. The weights of the mixture models must satisfy the constrain $\sum_{k=1}^{M} \omega_k = 1$.

$$P_k(\boldsymbol{x_i}|\boldsymbol{\theta_k}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{x_i}-\boldsymbol{\mu_k})^T(\boldsymbol{\Sigma_k^{-1}})(\boldsymbol{x_i}-\boldsymbol{\mu_k})\right)}{2\pi^{d/2}|\boldsymbol{\Sigma_k}|^{1/2}} \tag{3.25}$$

The parameters of the Gaussian Mixture are estimated using the maximum likelihood, according to the Expectation Maximization algorithm (EM). In the first step of this algorithm, $\boldsymbol{\theta_k}$ and ω_k are initialized randomly. Commonly, the k-means clustering algorithm is used for that. The second step is called the E-step. Here, $P_k(\boldsymbol{\theta_k}|\boldsymbol{x_i})$ is computed using Equation 3.25 for every feature vector $\boldsymbol{x_i}$ given every $\boldsymbol{\theta_k}$. Then, the posterior probability is computed according the Bayes theorem [66] which satisfies Equation 3.26.

$$P_k(\boldsymbol{\theta_k}|\boldsymbol{x_i}) = \frac{P_k(\boldsymbol{x_i}|\boldsymbol{\theta_k})P_k(\boldsymbol{\theta_k})}{\sum_{k=1}^{M} P_k(\boldsymbol{x_i}|\boldsymbol{\theta_k})P_k(\boldsymbol{\theta_k})} \tag{3.26}$$

In the third step or M-step, the objective function (see Equation 3.27) aims to maximize $P_k(\boldsymbol{\theta_k}|\boldsymbol{x_i})$, thus the parameters of the Gaussian mixtures are computed.

$$\boldsymbol{\theta_{k+1}} = \underset{\boldsymbol{\theta_k}}{argmax} P_k(\boldsymbol{\theta_k}|\boldsymbol{x_i}) \tag{3.27}$$

The weights of the Gaussian components given by $P_k(\boldsymbol{\theta_k})$ are defined as:

$$P_k(\boldsymbol{\theta_k}) = \hat{\omega}_{k+1} = \frac{\sum_{i=1}^{d} P_k(\boldsymbol{\theta_k}|\boldsymbol{x_i})}{d} \tag{3.28}$$

Then, $\boldsymbol{\mu_k}$ and $\boldsymbol{\Sigma_k}$ are updated until convergence using Equation 3.29 and 3.30, respectively.

$$\hat{\boldsymbol{\mu}}_{k+1} = \frac{\sum_{i=1}^{d} P_k(\boldsymbol{\theta_k}|\boldsymbol{x_i})\boldsymbol{x_i}}{\sum_{i=1}^{d} P_k(\boldsymbol{\theta_k}|\boldsymbol{x_i})} \tag{3.29}$$

$$\hat{\boldsymbol{\Sigma}}_{k+1} = \frac{\sum_{i=1}^{d} P_k(\boldsymbol{\theta_k}|\boldsymbol{x_i})(\boldsymbol{x_i}-\hat{\boldsymbol{\mu}}_{k+1})^T(\boldsymbol{x_i}-\hat{\boldsymbol{\mu}}_{k+1})}{\sum_{i=1}^{d} P_k(\boldsymbol{\theta_k}|\boldsymbol{x_i})} \tag{3.30}$$

Then, after the k-th iteration $P_k(\boldsymbol{x_i}|\boldsymbol{\theta_k})$ is increased, i.e., $P_{k+1}(\boldsymbol{x_i}|\boldsymbol{\theta_{k+1}}) > P_k(\boldsymbol{x_i}|\boldsymbol{\theta_k})$. In this algorithm the feature vectors $\boldsymbol{x_i}$ are assumed independent, so the log-likelihood of a given model $\boldsymbol{\Theta}$ may be computed.

Universal Background Models

The GMM trained with a large sample of speakers is called UBM. In this case, given a spontaneous speech transcript and a hypothesized D-PD or ND-PD, the goal is to determine how

close is the PD patient to the UBM given trained using HC subjects. The UBM model is less well defined since it must potentially represent the entire space of possible alternatives to the patient. No objective measure is defined to establish how many data is needed to train the UBM. There are two main approaches to build the UBM. The first approach is to train individually a set of sub-population models related to the data, this to cover the space of the alternatives. The second approach consists of only pool all data to train the UBM via EM algorithm. This thesis considers the second approach to train the UBM using the HC subjects.

Adaptation of the GMM

The basic idea is to derive the patient model by updating the parameters in the UBM via adaptation. This allows to get a better coupling between the patient model and the UBM. The adaptation is a two step estimation process, similar to the EM algorithm. The first step is related to the E-step computed for each mixture in the UBM. The second step is related to the adaptation, in which a data-independent coefficient is used to combine the old sufficient statistics given by the UBM parameters with the new sufficient statistics estimates. This coefficient relies more on the new sufficient statistics in mixtures with larger data, and in smaller data relies more the old sufficient statistics, both related to the final parameter estimation. The probabilistic alignment for the training vector $\boldsymbol{x}$ is determined into the UBM mixture component. For the k-th mixture in the UBM $P(k|\boldsymbol{x_i})$ is computed as in Equation 3.31.

$$P(k|\boldsymbol{x_i}) = \frac{\omega_k P_k(\boldsymbol{x_i})}{\sum_{j=1}^{M} \omega_j P_j(\boldsymbol{x_i})} \tag{3.31}$$

$P(k|\boldsymbol{x_i})$ and $\boldsymbol{x_i}$ will be used to compute the statistics for ω_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ as in Equations 3.32, 3.33, and 3.34.

$$\mathrm{n}_k = \sum_{i=1}^{D} P(k|\boldsymbol{x_i}) \tag{3.32}$$

$$\mathrm{E}_k(\boldsymbol{X}) = \frac{1}{\mathrm{n}_k} \sum_{i=1}^{d} P(k|\boldsymbol{x_i})\boldsymbol{x_i} \tag{3.33}$$

$$\mathrm{E}_k(\boldsymbol{X}^2) = \frac{1}{\mathrm{n}_k} \sum_{i=1}^{d} P(k|\boldsymbol{x_i})\boldsymbol{x_i}^2 \tag{3.34}$$

The adapted parameters for the k-th mixture are obtained by the new sufficient statistics from

the training data (ω_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$) as in Equations 3.35, 3.36, and 3.37.

$$\hat{w}_k = \left[\alpha_k^{\omega}\frac{\mathrm{n}_k}{d} + (1-\alpha_k^{\omega})w_k\right]\gamma \tag{3.35}$$

$$\hat{\boldsymbol{\mu}}_k = \alpha_k^{m}\mathrm{E}_k(\boldsymbol{X}) + (1-\alpha_k^{m})\boldsymbol{\mu}_k \tag{3.36}$$

$$\hat{\boldsymbol{\Sigma}}_k = \alpha_k^{v}\mathrm{E}_k(\boldsymbol{X}^2) + (1-\alpha_k^{v})(\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k^2) - \hat{\boldsymbol{\mu}}_k^2 \tag{3.37}$$

Where $\{\alpha_k^{\omega}, \alpha_k^{m}, \alpha_k^{v}\}$ are the adaptation coefficients that control the balance between new and old estimates for the weight, the mean and the variance respectively. γ is the scale factor, that is computed over all adapted weights to fulfill the constraint $\sum_{k=1}^{M}\omega_k = 1$. The adaptation coefficient α_k^{ρ} is defined in Equation 3.38, where r^{ρ} is a fixed relevance factor for parameters ρ.

$$\alpha_k^{\rho} = \frac{\mathrm{n}_k}{\mathrm{n}_k + \mathrm{r}^{\rho}} \tag{3.38}$$

The updating parameter can be derived from the general Maximum A-posteriori estimation known as MAP.

3.1.2 Deep Learning Methods

Deep learning has been frequently used nowadays in different applications related to speech, image, video and NLP. The performance varies depending on the application and the structure of the Deep Neural Network (DNN) architecture and the availability of data [67].

DNNs are formed with several layers, in which the output of a layer corresponds to the input of a deeper one. The layers are used in order to control the information to feed the DNN model, improve the robustness, and the performance.

A Feed-Forward DNN (FF-DNN) is an artificial neural network that the information moves in only one direction, forward, i.e., the outputs of the model do not feed back on themselves. The goal of a FF-DNN is to approximate a function f^*. A FF-DNN defines a mapping $y = f^*(\boldsymbol{x};\theta)$ by learning the value of the parameters θ that result in the best function approximation. The output of an FF-DNN is expressed according to Equation 3.39, where j is the number of abstraction layers, x is the input and ϕ is the nonlinear transformation. In addition, Figure 3.8 shows the general scheme of a DNN.

$$y = \phi_j(f_j(\cdots(\phi_2(f_2(\phi_1(f_1(\boldsymbol{x})))))))) \tag{3.39}$$

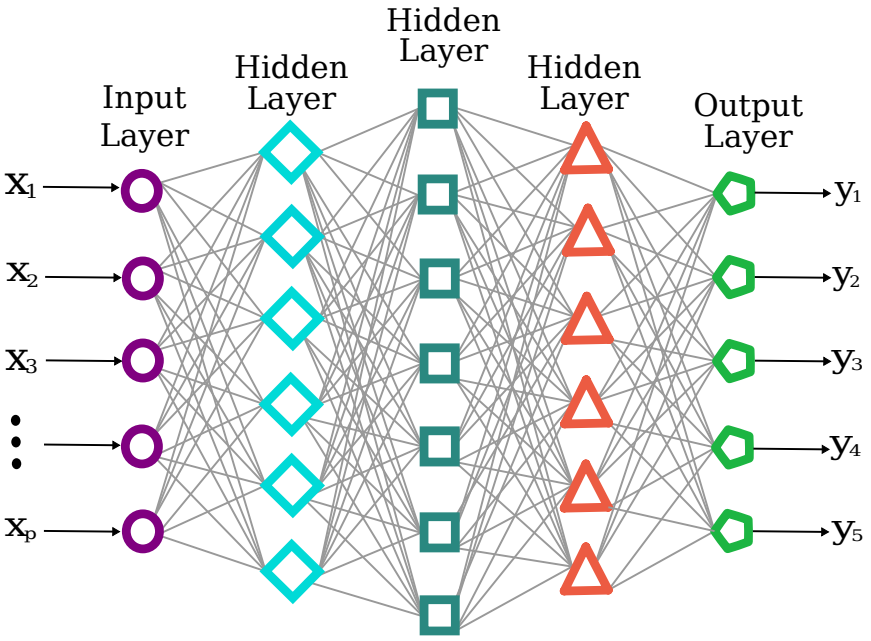

Figure 3.8: General scheme of a feed-forward DNN

In the training process an objective function is computed to measure the performance in order to update its internal structure in form of adjustable parameters. The Stochastic Gradient Descend (SGD) algorithm is used to train deep learning methods. It is an iterative optimization algorithm to find minimum values in convex and differential functions. This method (see Equation 3.40) uses a small set of the training data as input to adjust the weight parameters $\boldsymbol{w}$ of a DNN. The gradient g^k is based on a loss function $g^k = L(\boldsymbol{w}^{(k)})$. The step size defined by learning rate η is used to converge to a local minimum. This process is repeated many times with several small data sets or batches.

$$\boldsymbol{w}^{(k+1)} = \boldsymbol{w}^{(k)} - \eta \bigtriangledown L(\boldsymbol{w}^{(k)}) \tag{3.40}$$

An Extension of SGD is the Stochastic optimization method called Adam. The name Adam is derived from Adaptive Moment Estimation. It provides more robustness, since SGD maintains a single learning rate, and Adam computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients.

Adam is a combination between two optimization techniques: Adaptive Gradient Algorithm (AdaGrad) and Root Mean Square Propagation (RMSProp). Adam computes and exponential moving average for the gradient v^k and the square gradient r^k, which satisfies Equation. The parameters β_1 and β_2 control the decay rates of these moving averages. ϵ is a floating number to prevent divisions by zero. $\odot$ is the Hadamard product.

$$
\begin{aligned}
v^k &= \beta_1 v^{k-1} + (1-\beta_1) g^k \\
r^k &= \beta_{12} v^{k-1} + (1-\beta_2) g^k \odot g^k \\
\boldsymbol{w}^{(k+1)} &= \boldsymbol{w}^{(k)} - \eta \frac{v^k}{\sqrt{r^k} + \epsilon}
\end{aligned}
\tag{3.41}
$$

Convolutional Neural Networks

There are other DNN architectures easier to train than the fully connected DNNs. These are called Convolutional Neural Networks (CNN), which are frequently used in image processing. In speech analysis, CNNs are used in different applications such as detection of events in audio, speech recognition and ER [68]. CNNs are designed to process multiple arrays. For instance, the analysis of two dimensional arrays from a time-frequency representation (spectrograms) from speech signals. These networks are formed by a structure of alternating convolutional filters and pooling layers instead of the fully connected layers of a DNN.

The input of a CNN is a tensor $\boldsymbol{X} \in \mathbb{R}^{v \times d \times c}$, where v and d can be the time-frequency axes from spectrograms with c channels. A weight tensor $\boldsymbol{W} \in \mathbb{R}^{m \times m \times h}$ is convolved with the input element $i,\ j$ of matrix $\boldsymbol{X}$ in each convolutional layer according to the Equation 3.42, that produces a hidden representation $\boldsymbol{H} \in \mathbb{R}^{v-m+1 \times d-m+1 \times h}$ of the extracted features. m is the order of the convolutional filter, and h is the number of hidden units in the convolutional layer.

$$
\boldsymbol{H}(i,\ j,\ h) = \text{conv}(\boldsymbol{X}, \boldsymbol{W_h})(i,\ j) = \sum_{j=1}^{m} \sum_{n=1}^{m} \boldsymbol{X}(i+m,\ j+n) \boldsymbol{W_h}(m,\ n) \tag{3.42}
$$

Usually, after the convolution operation a pooling layer is used to resample the hidden representation $\boldsymbol{H}$, removing non-relevant information that may appear because speech accent or distortion. The last layer of a CNN corresponds to a fully connected layer that groups all the features with a non-linear activation function to make the final decision. Figure 3.9 shows a typical architecture of a CNN formed by convolution layers followed by a fully connected layer to predict the output.

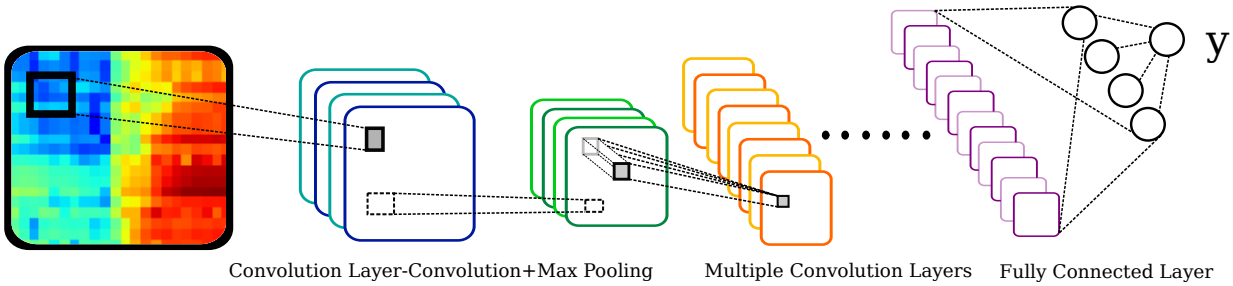

Figure 3.9: General scheme of a convolutional neural network with a fully connectd layer

Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are related to process sequential data in DNNs. They incorporate a feed-back allowing that the network to have "memory". The decision in a time instant $t-1$ affects the decision of an RNN in a time instant t. The networks are connected to their past decisions (feedback loop), where the sequential information is stored in the hidden state of a network, which manages to span many time steps.

RNNs seek an existing correlation between the events that are separated in time. These correlations are called "long-term dependency". For instance, an utterance contains several words that depend on the previous words to put them into context, as a feedback loop. The process of carrying memory forward is defined in Equation 3.43, where the hidden layer (h^t) in an instant t is computed.

$$\boldsymbol{h}^{(t)} = f(\boldsymbol{h}^{(t-1)};\ \boldsymbol{x}^{(t)};\ \theta) \tag{3.43}$$

The feedback occurs at every time step in the series, each hidden layer state contains also traces of all those that preceded $\boldsymbol{h}^{(t-1)}$ for a long memory. Figure 3.10 shows the general scheme of a simple RNN, where it can be observed that the RNN has a chain form of repeating modules of neural networks.

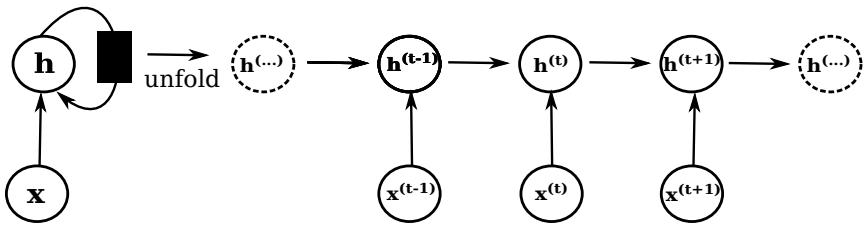

Figure 3.10: General scheme of a RNN

In the RNNs, the repeating module consists of a very simple structure of just a single $tanh$

layer, as is shown in Figure 3.11.

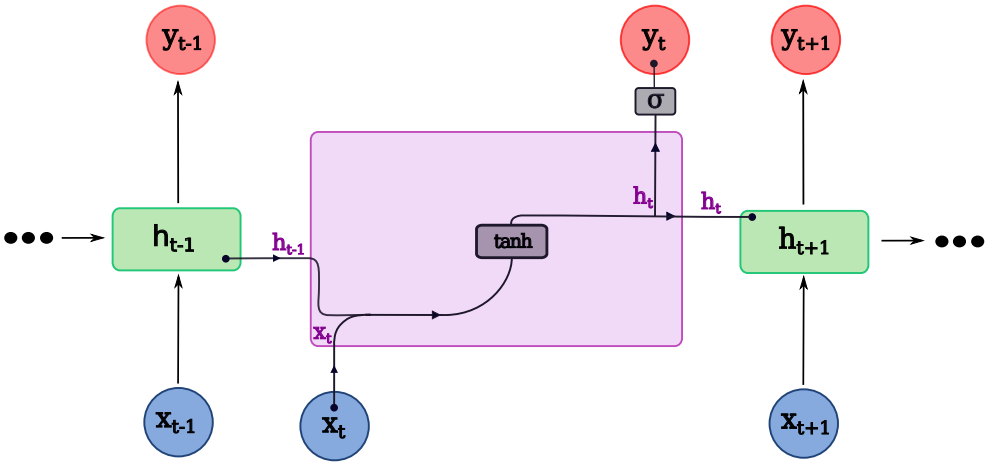

Figure 3.11: RNN cell unit

Long Short-Term Memory

The LSTM is a RNN-based architecture, capable to learn long-term dependencies. This neural network can remember information for long periods of time and it is designed to solve the problem of vanishing gradients in the typical RNN. The main idea is to introduce gates that control writing and accessing memory in an additional cell state. The LSTM cell unit is illustrated in Figure 3.12.

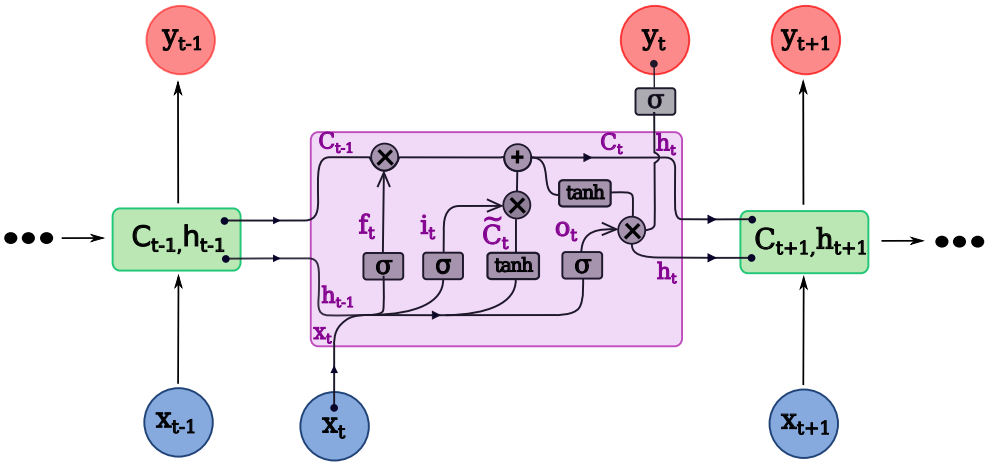

Figure 3.12: LSTM cell unit

The LSTM has a chain structure as the RNN with a memory called hidden state, but a new memory is added. This memory is knwon as cell state (C_t). The cell state goes straight down the

entire chain. It undergoes only linear changes and it is time dependent. An LSTM consists of the interaction of four different layers to update the internal state in multiple steps, as is shown in Figure 3.12. These layers are called gates.

The first gate, also known as the forget gate, controls which details from the previous cell state are forgotten. The main idea is to forget and memorize information in separate states by looking at the previous hidden state ($\boldsymbol{h}_{t-1}$) and the input $\boldsymbol{x}_t$, which satisfies Equation 3.44. σ is the sigmoid activation function and $\boldsymbol{W}_f$ is the weights of the neural network that corresponds to the forget gate.

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t]) \tag{3.44}$$

Then, the second one is the input gate. It contributes to the decision of which values will be updated. This gate is represented by $\boldsymbol{i}_t$ in Equation 3.45. New candidate values $\widetilde{\boldsymbol{C}}_t$ (see Equation 3.46) are created by the layer with a $\tanh$ activation.

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t]) \tag{3.45}$$

$$\widetilde{\boldsymbol{C}}_t = \tanh\left(\boldsymbol{W}_C[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t]\right) \tag{3.46}$$

The cell state ($\boldsymbol{C}_t$) and the hidden state ($\widetilde{\boldsymbol{h}}_t$) are updated separately. A new cell state appears by summing the remaining information from $\boldsymbol{C}_{t-1}$, and new information from the input and the previous hidden state as is shown in Equation 3.47, where $\odot$ is an element-wise multiplication.

$$\boldsymbol{C}_t = \boldsymbol{f}_t \odot \boldsymbol{C}_{t-1} + \boldsymbol{i}_t \odot \widetilde{\boldsymbol{C}}_t \tag{3.47}$$

Finally, the output is computed based on the input and the cell state using Equation 3.48. The hidden state is obtained by multiplying the output and the cell state passed through a $\tanh$ layer (Equation 3.49). The sigmoid function in $\boldsymbol{o}_t$ decides which values from the input will pass. The $\tanh$ is used to weight the values from the cell state, and thus decide levels of importance. In Equation 3.50, it can be observed that the output $\boldsymbol{y}_t$ directly depends on the hidden state $\boldsymbol{y}_t$.

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t]) \tag{3.48}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh\left(\boldsymbol{C}_t\right)) \tag{3.49}$$

$$\boldsymbol{y}_t = \sigma(\boldsymbol{h}_t) \tag{3.50}$$

Gated Recurrent Unit

Gated Recurrent Unit (GRU) can be considered as a variation of the LSTM, due to the similarity in their designs, both based on gating mechanisms. This variation was proposed in order to reduce the number of parameters in the LSTM and for easier training. The main difference is that the memory operates directly via the hidden state. The number of gates are reduced to just two, i.e., the number of layers and parameters are fewer. The GRU cell unit is shown in Figure 3.13.

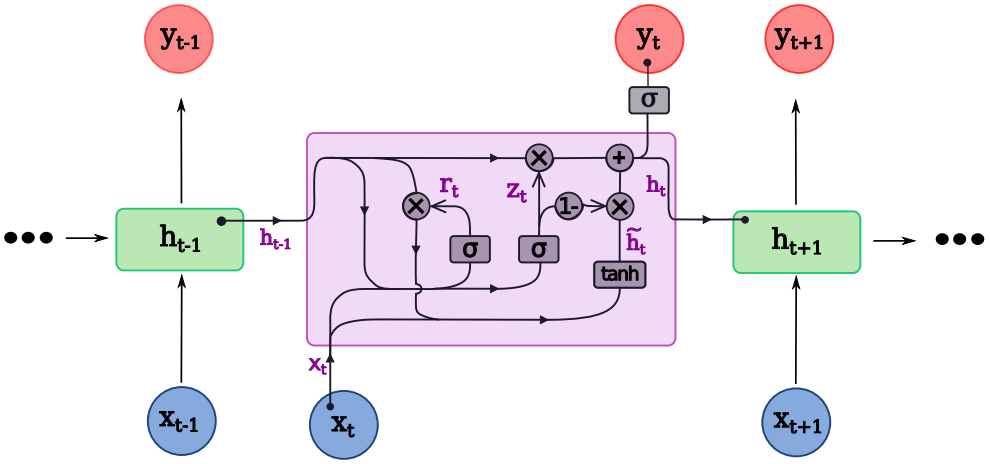

Figure 3.13: GRU cell unit

The first gate is called reset gate and determines the influence of the previous hidden states. The relation with the LSTM is that the forget and the input gate are combined in this gate. The reset gate is computed similar as the gates in the LSTM as is shown in Equation 3.51.

$$\boldsymbol{r}_t = \sigma(\boldsymbol{W}_r[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t]) \tag{3.51}$$

The update gate in Equation 3.52 is the second one. The update gate determines the influence of past information in the current state, i.e., the influence of newly computed update.

$$\boldsymbol{z}_t = \sigma(\boldsymbol{W}_z[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t]) \tag{3.52}$$

In Equation 3.53, a current memory content is computed by using the reset gate. An update will be proposed, where the input and the reset are combined. It allows to observe how the gates

will affect the final output. As the reset gate is computed by using a sigmoid activation function, it will have a low influence of previous hidden state when $\boldsymbol{r}_t$ is close to zero.

$$\widetilde{\boldsymbol{h}}_t = \tanh\left(\boldsymbol{W}_h[\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}, \boldsymbol{x}_t]\right) \tag{3.53}$$

Finally, the updated hidden state is computed, as is shown in Equation 3.54. The update gate is needed in this step, as it controls the combination of the old state and the proposed update. This will determine what to keep from the current memory content $\widetilde{\boldsymbol{h}}_t$ and what from the hidden states. The output node $\boldsymbol{y}_t$ is defined in the same way as in the LSTM, which satisfies Equation 3.50.

$$\boldsymbol{h}_t = \boldsymbol{z}_t \odot \boldsymbol{h}_{t-1} + (1 - \boldsymbol{z}_t) \odot \widetilde{\boldsymbol{h}}_t \tag{3.54}$$

Deep learning in speech analysis: there are several approaches to analyze speech signals by DNNs [69]. For instance, DNNs could be trained using previously extracted acoustic features. In speech processing, a spectrogram is a graphical representation that allows to visualize the frequency content in an audio. Spectrograms have been widely used in CNNs based schemes to achieve emotion recognition. On the other hand, RNNs are commonly used to process time series of speech frames, among several other possibilities.

Deep learning in text analysis: the aim is to represent words as vectors through neuronal networks. There are several methods based on deep learning techniques, but this study will be focused on a novel method called "Word embeddings". The aim of the word embeddings is to redefine the high dimensionality word features into low dimensional feature vectors, preserving the contextual similarity in the corpus. These methods are widely used in deep learning models as RNN and CNN. The most popular models to build word embeddings are Word2Vec [70], Glove [71] and BERT [72]. Word2Vec is an alternative of word representation that pretends to represent the words as a multidimensional space vector, where similar or related words are represented by nearby points, i.e. a FF-DNN that learns vectors to improve the predictive ability. Glove is a count-based model that learns words from their co-occurrence information from a corpus. BERT is one of the most recent word embedding methods. This consists of a model of pre-trained language representation, with a more complex scheme, by jointly conditioning on both left and right context in all layers.

3.1.3 Performance Metrics

The performance metrics are used to evaluate the accuracy and generalization capability of the outcomes of a model. They are defined as a process that quantifies the effectiveness and efficiency of past actions. In this study, several metrics are used to evaluate different pattern recognition algorithms. Commonly, it is necessary to introduce various measures in the context of a classification problem, where the labels for a bi-class classification are defined as positives or negatives [73]. These sets of measures come from a four-cell contingency table known as confusion matrix.

Confusion Matrix

The confusion matrix allows to visualize the algorithm performance of a supervised learning method. It is a table that displays and compares original labels with the ones predicted by the model. Table 3.1 shows the structure of the confusion matrix. The number of class predictions are represented by rows, while the instances of the original labels are representing by columns. Based on a binary classifier each cell in the matrix is defined as follows:

- True positive (TP): the number of cases correctly identified as the positive class.
- True negative (TN): the number of cases correctly identified as the negative class.
- False positive (FP): the number of misclassified cases for the positive class.
- False negative (FN): the number of misclassified cases for the negative class.

Table 3.1: Confusion matrix

		Original labels	
		1	0
Predicted labels	1	**TP**	FP
	0	FN	**TN**

The confusion matrix is used to guide the computation of the different performance matrix, such as: accuracy, specificity, sensitivity, recall, F-score, precision and the Receiver Operating Characteristic curve (ROC).

Unweighted Average Recall

The Unweighted Average Recall (UAR) in Equation 3.55 is the ability to discriminate different classes, i.e, the fraction of predictions that were correctly classified by the model. Table 3.2 shows the taken cells from the confusion matrix to calculate the UAR.

Table 3.2: Confusion matrix cells used to computed the accuracy

		Original labels	
		1	0
Predicted labels	1	**TP**	FP
	0	FN	**TN**

$$\text{UAR} = Average\left(\frac{\text{TP}}{\text{TP}+\text{FN}}, \frac{\text{TN}}{\text{TN}+\text{FP}}\right) \times 100 \tag{3.55}$$

Sensitivity and Specificity

The sensitivity in Equation 3.56, is the proportion of TP correctly identified as such. Table 3.3 shows the necessary cells of the confusion matrix to calculate the sensitivity.

Table 3.3: Confusion matrix cells used to computed the sensitivity

		Original labels	
		1	0
Predicted labels	1	**TP**	FP
	0	**FN**	TN

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{3.56}$$

The specificity in Equation 3.57, is the ability to discriminate the negative class correctly. It is the proportion of TN correctly identified as such, i.e, the rate of TN. Table 3.4 shows the necessary cells from the confusion matrix to calculate the specificity.

Sensitivity and specificity ranges from 0 to 100%

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} \tag{3.57}$$

Table 3.4: Confusion matrix cells used to computed the specificity

		Original labels	
		1	0
Predicted labels	1	TP	**FP**
	0	FN	**TN**

Receiver Operating Characteristic Curve

The ROC curve is derived from a Gaussian distribution constructed from the classification scores. The classification score indicates a measure of confidence in the prediction. A graphical representation is performed to show the binary classifier performance, while its discrimination threshold is varied. The True Positive Rate (TPR) known as the Sensitivity, and False Positive Rate (FPR) define as $1-$Specificity are computed to obtain the distribution. The ROC curve (Figure 3.14) is created by plotting TPR and FPR for all possible threshold values, where this curve is defined in the x-axis by FPR, and in the y-axis by TPR.

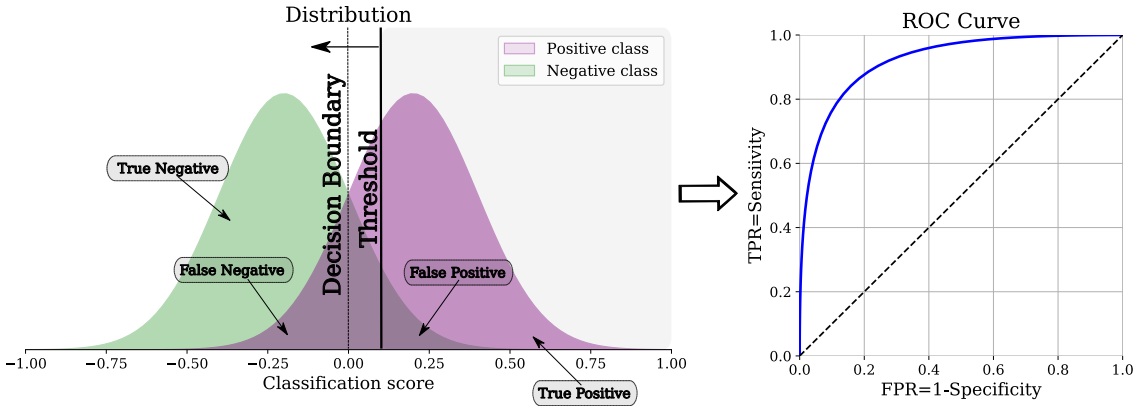

Figure 3.14: ROC curve derived from a Gaussian distribution

The model performance is determined by looking at the Area Under the ROC Curve (AUC). The AUC represents degree or measure of separability. It ranges from 0 to 1, in which 1 indicates that the model is able to perfectly discriminate between the two classes. Four different cases are considered as examples: (1) in Figure 3.15, is considered two distribution (at the left) with non-overlapping that produce a ROC curve (at the right) with and AUC of 1, i.e., that the model has an ideal measure of separability, (2) in Figure 3.16, the prediction is less accurate, since the distribution are more overlapped, with an AUC of 0.80, (3) in Figure 3.17, the AUC is approximately 0.5. It indicates that model has no discrimination capacity to distinguish between the classes, and (4) in Figure 3.18, an AUC of 0 is obtained, when the predictions are wrong,

confusing a negative class as a positive class and vice versa.

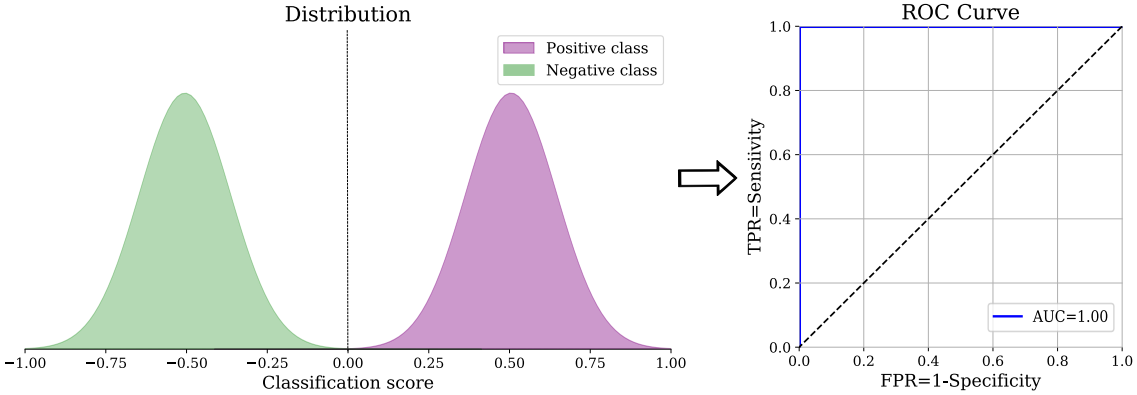

Figure 3.15: ROC curve from a non-overlapping distribution

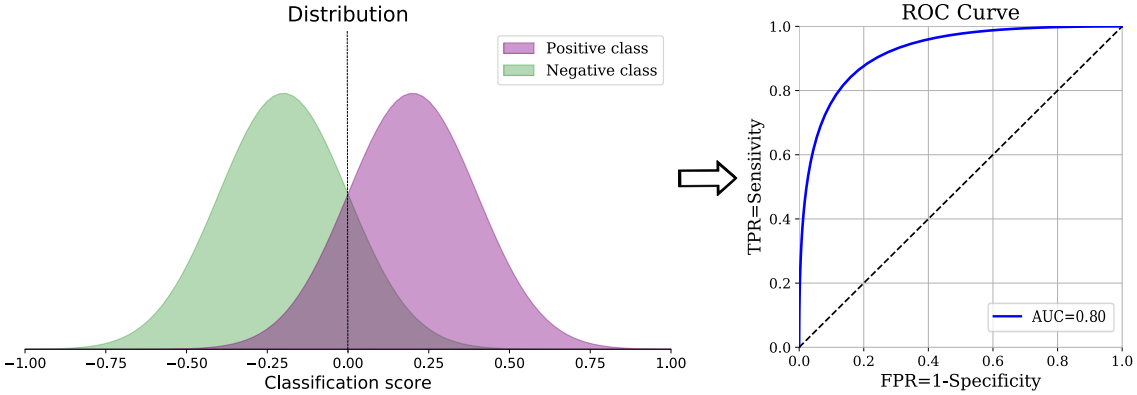

Figure 3.16: ROC curve from a overlapping distribution

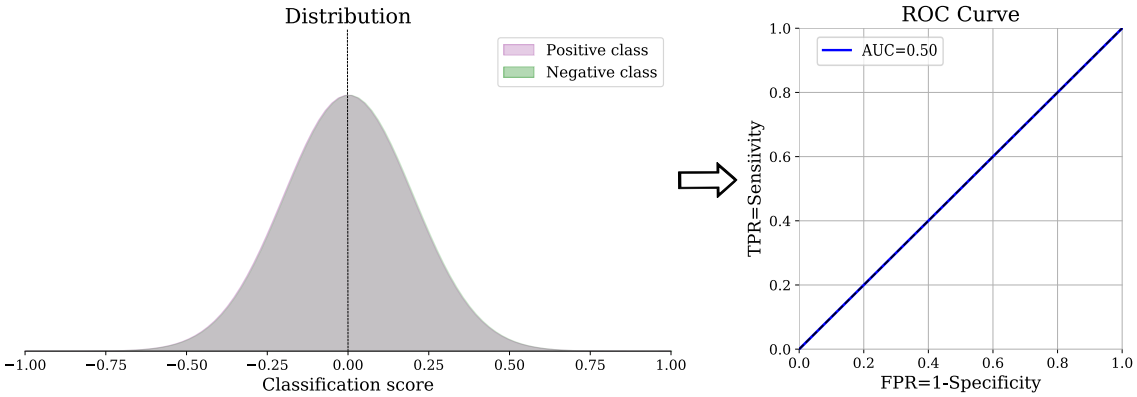

Figure 3.17: ROC curve from a totally overlapping distribution

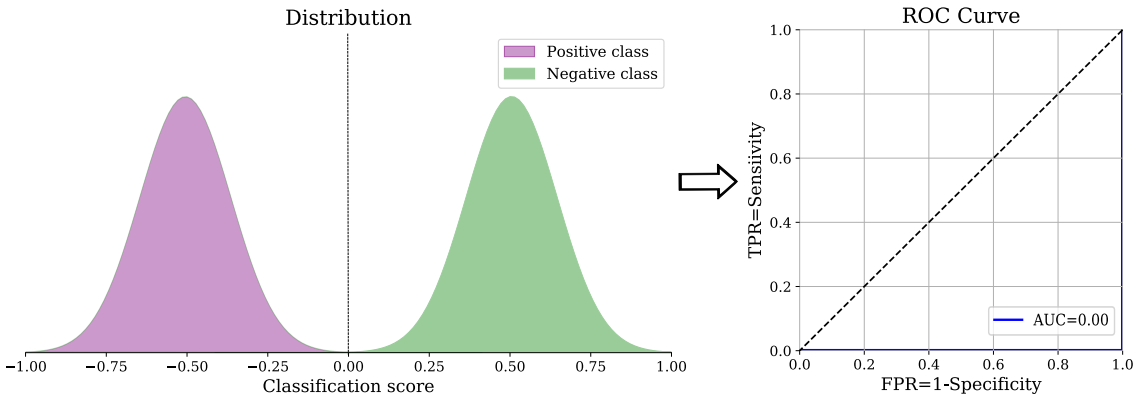

Figure 3.18: ROC curve from a non-overlapping distribution but with all misclasified predictions

Precision and Recall

Precision or positive predictive value in equation 3.58, is the proportion of positive identification that were actually correct. It defines the ability of a classifier not to predict a positive identification which is originally negative. Table 3.5 shows the necessary cells from the confusion matrix to calculate the specificity.

Table 3.5: Confusion matrix cells used to computed the precision

		Original labels	
		1	0
Predicted labels	1	**TP**	**FP**
	0	FN	TN

$$Precision = \frac{TP}{TP + FP} \tag{3.58}$$

Recall has the same definition that the sensitivity for a binary classification, but it is a more general definition that can be extended for multi-class classification problems. Precision and recall ranges from 0 to 1.

F-score

F-score in equation 3.59 is a measure to test the general performance of the classifier by computing the harmonic mean of the precision and recall. This measure penalizes harder the extreme values. The value of the F-score varies between 0 (worst possible value) and 1 (best possible value).

Table 3.6: Confusion matrix cells used to computed the f-score

		Original labels	
		1	0
Predicted labels	1	**TP**	**FP**
	0	**FN**	TN

$$F\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.59}$$

3.2 Emotions Modeling by the Arousal-Valence Plane

Emotions are conscious experiences characterized by a high mental activity, being a subjective reaction to certain events. These are generated as a lower response occurring in the subcortical regions of the brain and the neocortex. Darwin conceptualized the notion of emotions as separate entities, arguing that all humans show emotions through similar behaviors. He defines these separate entities into six emotional states: happiness, sadness, fear, anger, surprise and disgust [74]. Nowadays, it is considered that the human being expresses 28 different emotions, since these six discrete emotions are considered insufficient to capture the emotional richness that humans present and which influence humans' decisions to such a great degree [6]. Wundt proposed a different conceptualization of emotions in a continuum of pleasantness and activity/intensity. This notion of emotions was adopted by Russell who proposed a continuous 2-dimensional arousal-valence model [6]. Emotion modeling in this thesis is based on the Russell's model which considers valence as a representation of positive-negative hedonic tone, and arousal as level of calmness or excitement. The hedonic tone is a property of a sensory or another experience related to the characteristic ability to feel pleasure.

The heuristic affect underlies the concepts of feeling, mood and emotion. It is related to the basic sense of feeling as a non-conscious reaction to a stimulus, that occurs before the cognitive processes necessary for the formation of a more complex emotion. Conversely, feelings are the conscious experience of emotional reactions [3]. These processes affect the mood for a short period of time. Moods are not as intense as emotions and have less specific or immediate cause, and these influence the emotion you feel. Emotions last from seconds to minutes, while moods could last days or even months, such as depression, anxiety, among others.

Affective states are not independent, they have certain relation to one another. Human emotions are constantly changing, and it is really challenging to define a standard value to distinguish

emotions. It is either pleasant or unpleasant (valence), or whether you are feeling calm or agitated (arousal). "The arousal-valence plane" [6] is commonly used to model emotional states in a multidimensional space (see Figure 3.19). This leads to represent different emotions in two dimensions called "arousal" and "valence". Since it is really challenging to analyze these emotional changes in a qualitative form, this plane allows to perform a quantitative analysis of emotions. The vertical dimension, also called arousal, corresponds to the excitation-relaxation. The arousal dimension defines the intensity of emotion, that ranges between high arousal (active) and low arousal (passive). The horizontal dimension known as valence lies to make a range between pleasure-displeasure emotions that refers to the level of physiological activation associated with emotions. The left side corresponds to a negative valence and the right to a positive, this regarding the level of satisfaction or dissatisfaction towards emotion. According to the aforementioned, emotions are defined as a linear combination of valence and arousal within circumflex model. User state modeling, in general, aims to capture similar aspects related to the emotions and mood in several applications, where the challenge is how to extract the information according to the modality [75], [76].

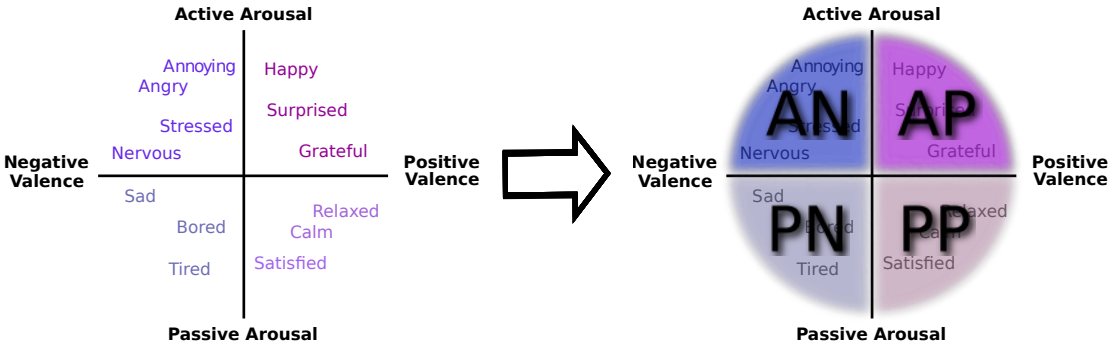

Figure 3.19: Bi-dimensional emotion representation in the arousal-valence plane

In this study, the arousal-valence plane is divided into four quadrants, where AP is Active Positive, AN is Active Negative, PN is Passive Negative, and PP is Passive Positive. This is proposed in order to get a wide representation of emotions to address different applications.

3.3 Speech Analysis Methods

In the speech production process, the vocal folds generate complex sounds composed by F_0 and the integer multiples of this frequency (harmonics). This complex sound is then passed and filtered in the vocal tract, producing different phonetic structures in the speech.

The vocal tract is formed by oral and nasal cavities, pharynx and glottis. Inside of these cavities there are the articulatory organs, which are divided into active and passive. Active articulators are actively involved in the the speech production process and include organs such as the lower lip, tongue, uvula, and glottis. Passive articulators are involved in the speech production but do not move. These consider organs such as upper lip, teeth, palate, velum, and pharynx. The vocal tract adopts several configurations through the different positions of the articulatory organs, acting as an acoustic filter for the produced sounds in the glottis.

Changes in the vocal tract are based on the length of the tract and the different diameters along its length. This filter is defined by the formant frequencies, which appear as resonances in the power spectrum of the speech signal.

3.3.1 Pre-Processing

Speech is greatly affected by different factor such as type of microphone, noise, volume, among other perturbations caused by the recording condition. Pre-processing of speech signals is considered an important step for the development of a robust and efficient acoustic analysis. The pre-processing mainly includes several steps: normalization, perturbation removal, and segmentation. In non-controlled conditions some existing algorithms to remove background noise are used [77] in this study.

The speech signals consist of: voiced and unvoiced segments (see Figure 3.20). Voiced segments are related to the vibration of the vocal folds due to the glottis closure that produces quasi-periodic behaviors. Unvoiced segments are mostly produced by different aspects such as the release or closure in the vocal tract or turbulent airflow at the constriction.

Besides voiced/unvoiced segments, the transitions from unvoiced to voiced segments (onset) and from voiced to unvoiced segments (offset) are obtained as is shown in Figure 3.21. These transitions are detected based on the presence of the $\boldsymbol{F}_0$ using Praat [59]. The application of the onset and offset transitions have proven to be useful in ER [78], [79].

3.3.2 Prosodic Analysis

Prosody analyzes the supra-segmental and melodic aspects involved in the word production. Supra-segmental aspects are related to duration, intensity, and tone. Melodic considers aspects such as rhythm, accentuation, and co-articulation phenoms. Prosody features are derived from physiological parameters based on $\boldsymbol{F}_0$, duration, and the energy content.

The emotion of a speaker affects the energy content in the speech. For instance, in [80],

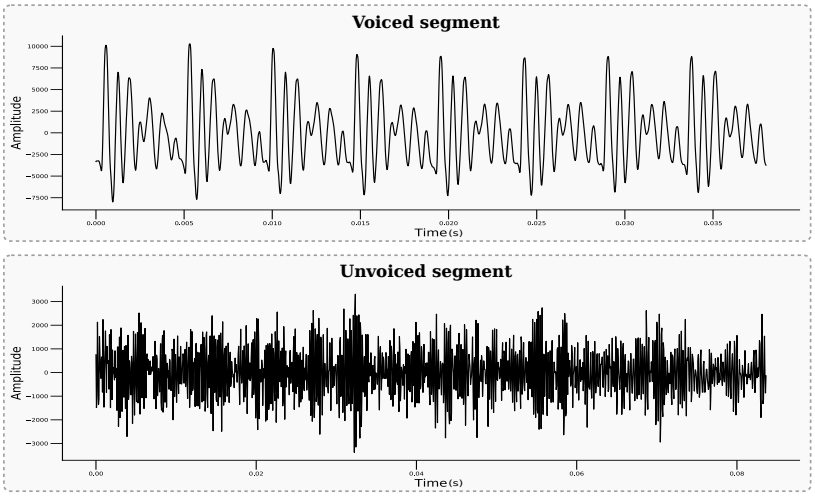

Figure 3.20: Example voiced and unvoiced segment of a female speaker

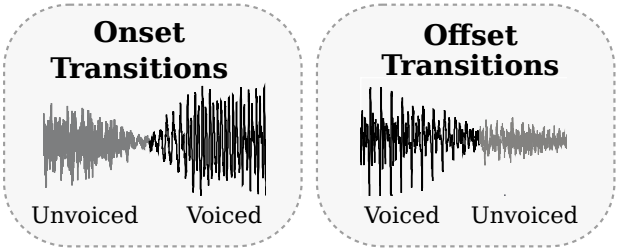

Figure 3.21: Onset and offset transitions

[81] it was reported that happiness and anger exhibit a higher energy envelope, while sadness is associated with a decreased energy.

The F_0 measures how low or high sounds the voice frequency of a person. The contour of the F_0 is a useful marker to discriminate emotions from speech. For instance, neutral speech produces a narrower F_0 range than the emotional speech. Fear has a high median, wide range, and a moderate rate of change [82]. Conversely, angry speech has a high median, wide range, and high rate of change in comparison with other emotions [83]. The vowels in the angry speech exhibit the highest F_0, and have downward slopes in relation with other emotions [80]. F_0 related to low arousal emotions such as disgust or sadness, exhibits a lower mean and a narrower range for sadness, and a low median, wide range, and lower rate of change for disgust [81].

Features based on different Duration Ratios (DR) based on the duration of pauses, voiced and unvoiced segments are also considered, which satisfy Equations 3.60, 3.61, 3.62, 3.63, 3.64, and 3.65.

$$DR_1 = Pause/(Voiced + Unvoiced) \tag{3.60}$$

$$DR_2 = Pause/Unvoiced \tag{3.61}$$

$$DR_3 = Unvoiced/(Voiced + Unvoiced) \tag{3.62}$$

$$DR_4 = Voiced/(Voiced + Unvoiced) \tag{3.63}$$

$$DR_5 = Voiced/Pause \tag{3.64}$$

$$DR_6 = Unvoiced/Pause \tag{3.65}$$

Dynamic features related to prosody are considered using only the voice segments. These features are obtained following the methodology presented in [84]. This method consists of obtaining of the six coefficients α_i from the $M = 5$-degree Legendre polynomials $P_i(t)$ (Equation 3.66) that model the pitch and the energy contour, separately. These coefficients models different aspects of the contours such as the mean, slope, curvature, and the inflection points that model the fine detail. Additionally, these dynamic features also include the duration of each voice segments in order to get 13 prosody descriptors per voiced segment.

$$f(t) = \sum_{i=0}^{M} \alpha_i P_i(t) \tag{3.66}$$

3.3.3 Articulation Features

Mel-frequency Cepstral Coefficients

Commonly, in speech analysis the data is processed by computing a compressed representation of the signal that can not capture all of the dynamic information. However, it is possible to obtain a representation to observe how the energy varies in the frequency domain with respect to the time. It can be achieved by using the Fourier transform that converts the signal into a time-frequency representation. In this work a variation known as the Short-Time Fourier Transform (STFT) is used. Mel spectrogram is based on the spectrogram obtained by computing the STFT and the spectral power, but in addition the frequencies are converted into Mel scale. These coefficients are a smoothed representation of the speech spectrum taking into account information of the scale of the human hearing.

The first step to compute the Mel spectrogram is the application pre-emphasis filter to the raw signal x. The pre-emphasis filter is a first order filter that aims to balance the frequency spectrum from the high frequencies that usually have smaller amplitudes than lower frequencies. It is computed using Equation 3.67, where α is the filter coefficient, t is the time step, and the negative term is the pre-emphasis.

$$y(t) = x(t) - \alpha x(t-1) \tag{3.67}$$

The frequencies in a signal changes across the time. It is assumed that the frequencies are stationary of a short period of time. Thereby, the signal is splitted into frames and passed across of a Hamming window using Equation 3.68, where N is the window length and $0 \leq n \leq N-1$. The Hamming window helps to avoid discontinuities.

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \tag{3.68}$$

Then, the STFT and next the power spectrum (P) are computed in each frame i.

$$P_i = \frac{|STFT(\boldsymbol{x_i})|^2}{N} \tag{3.69}$$

In Equation 3.70, the Frequencies (f) in Hertz are converted into Mel scale (m). The Mel scale aims to simulate the non-linear perception of the human auditory with respect to the sounds, since the energy in a critical band of a frequency has influence in the human hearing. This critical band bandwidth varies with the frequency, being this scale linear below $1kHz$ and logarithmic upper this threshold. The information is transformed into the Mel scale domain in order to extract

a set of critical frequency bands by applying a bandpass filter adjusted around the center frequency, i.e., triangular filters (see Figure 3.22). After applying the filter bank to the power spectrum, it is converted to a log scale. The number of filters was set on 64 in this thesis.

$$m = 2595 \log_{10} \left(1 + \frac{f}{700}\right) \tag{3.70}$$

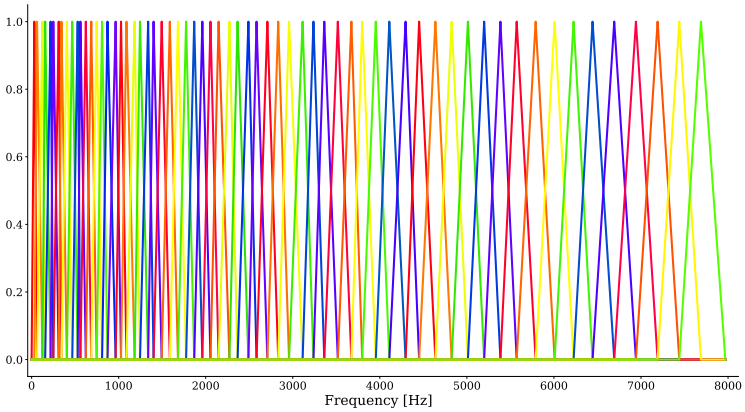

Figure 3.22: Filter bank on a Mel scale

This study considers two approaches using Mel. First approach considers the real and the imaginary part from the STFT in order to compute a 2-Dimensional Mel spectrogram that captures more information related to the dynamic in the signal. The signal was framed in a 40 ms window with a step size of 10 ms, and then a 500 ms sequence was formed. Finally, the second approach considers MFCCs. Thus, the Discrete Cosine Transform (DCT) is applied on the filter bank coefficients to decorrelate them. Usually, these are represent by the 12 first coefficients, and their first and second derivate.

Bark Band Energies

Bark bands are defined as the critical bands of human hearing [85]. Commonly, the human hearing system is modeled by a bank of 24 critical bands. These critical bands are used to quantify the capability of the human ear to distinguish between individual frequency tones. Each critical band is able to simulate the same quantity of cells from the basilar membrane, producing a proportional displacement in relation to the frequency distribution in Equation 3.71.

$$Bark(f) = 13 \arctan\left(0.00076 f\right) + 3.5 \arctan\left(\frac{f}{7500}\right)^2 \tag{3.71}$$

This relation can be observed in Figure 3.23. Note, that the bandwith of critical band are constant at 100 Hz for frequencies that are below of 500 Hz, while the increment is proportional to the logarithm of frequency for medium and high frequencies. The Bark scale frequency bands are almost linear below $1kHz$, while from frequencies superior to $1kHz$ the scale grows exponentially, which yields a perceptual filter-bank.

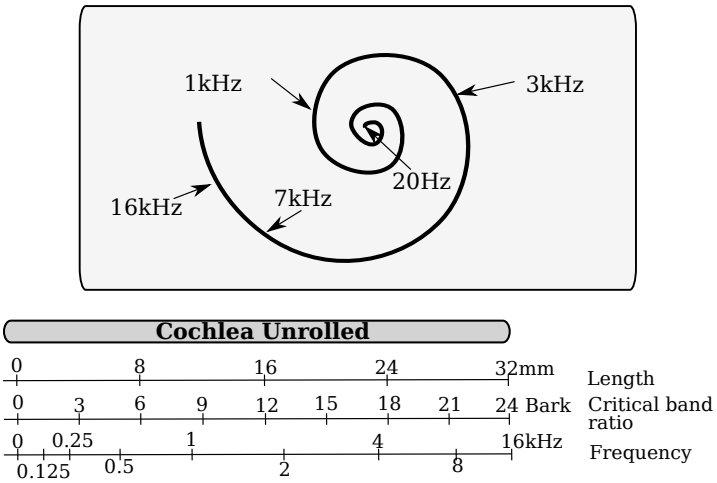

Figure 3.23: Critical bands of human hearing according to the Bark scale

In this work the log-energy of the speech signal distributed only considering 22 critical bands is computed. The process to obtain these Bark Band Energies (BBE) consist of calculating the corresponding Fourier spectrum in 22 frequency bands according to the Bark scale, and thus the log-energy of each band is computed.

Formant Frequencies

A formant frequency is an acoustic resonance of the vocal tract. When the sound waves pass through the supraglottic cavities, they modify the amplitude of the harmonics due to the resonance phenomenon. The formants are defined as each of the preferred resonating frequencies corresponds to the relevant bump in the frequency response curve. These formants vary for each emotion, especially the two first formants F_1 and F_2. F_1 is higher in emotions such as *anger* or *happiness* than *sadness* and *boredom*.

3.3.4 Speech Features to Model Emotions

There are physiological changes related to emotions that influence aspects of speech production such as breathing, phonation, resonance, prosody, and articulation. Emotions such as happiness,

anger, and fear, induce an increase in sub-glottal pressure, a dryness of the mouth, and occasional muscle tremor. In addition, high arousal emotions produce louder and faster speech, which is characterized by strong high frequency energy, a higher average pitch, and a wider pitch range [86]. On the other hand the low arousal emotions such as sadness and boredom affect the parasympathetic nervous system, producing speech with slow rate, low pitch, and lower high frequency energy [87].

Table 3.7 shows the relationship between the speech parameters and emotions. Note that F_0 and the energy content are the parameters that are more related to the emotional content.

Table 3.7: Relationship between emotions and speech parameters. Table adapted from [2]

Feature	Happiness	Anger	Sadness	Fear	Disgust
Articulation	Normal	Tense	Slurring	Precise	Normal
Rate	Slower/Faster	Slightly Faster	Slightly Slower	Much faster	Very much faster
F_0	Much higher	Very much higher	Slightly Lower	Very much higher	Very much lower
F_0 **Range**	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Energy content	Higher	Higher	Lower	Normal	Lower
Voice Quality	Breathy	Breathy chest blaring tone	Resonant voicing	Irregular chest tone	Grumble
F_0 **Changes**	Smooth, upward	Abrupt on stressed	Downward inflections	Normal	Wide, downward inflections

3.4 Linguistic Analysis Methods

NLP is a set of techniques responsible for understanding, interpreting, and manipulating human written language. It aims to extract information from natural language via machine learning methods. Some of the most common applications are: text classification and categorization, language generation, multi-document summarization, machine translation, and sentiment analysis. Particularly for sentiment analysis, there are word clusters related to certain types of emotions, e.g., words as “Happy”, “Smile” or “Love” are related to positive emotions, while “Cry”, “Sadness” or “Boring” are related to negative emotions. The context is another fundamental part to identify certain physiological events and can be more exhaustive when recognizing some emotions [88].

The common pipeline in NLP (see Figure 3.24) consists of four main steps: data pre-processing, vectorization, transformation, and model training. Data pre-processing helps to reduce the noise in the data. Vectorization converts a collection of text documents into a numerical vector. Transformation aims to extract features by different machine and deep learning techniques such as BoW, TF-IDF, W2V, and BERT. Model training consists of applying pattern recognition methods in order to analyze content based on particular categories.

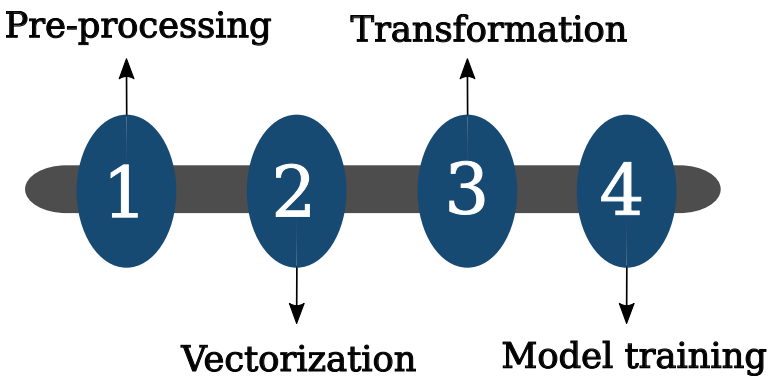

Figure 3.24: Text pre-processing scheme

3.4.1 Pre-Processing

Several types of noise are present in the text data. Text pre-processing aims to clean and standardize the text, making it noise-free and ready for analysis as it is shown in Figure 3.25.

Figure 3.25: Text pre-processing scheme

The following two steps are implemented for this purpose:

- *Noise removal:* non relevant information for the context, such as stopwords, accents or punctuation, are removed from the text. A Spanish dictionary of noise entities from the Natural Language Toolkit (NLTK) [89] is considered to remove stopwords such as *in*, *the*, *of*, among others. Punctuation and special characters (numbers, URLs) are removed by preparing a dictionary of noisy entities and iterate the text object by tokens, thus removing those tokens which are present in the dictionary.

- *Lexicon normalization:* there are multiple representations for a single word. To standardize the words in an equal representation, all words were transformed via stemming to remove the suffixes. Another implemented method was lemmatization, which transforms the words into their root form.

The pre-processing step is not applied for all the considered NLP methods. Noise removal and lexicon normalization are applied for BoW, TF-IDF, and W2V, however, the stemming process is

not applied for W2V. In the case of BERT, the removal of punctuation and special characters as text pre-processing is considered.

3.4.2 Bag of Words

BoW is an inference method that creates a vocabulary of all words that appear in the whole document. First, the sentences are represented as a collection of words that will be represented in a feature vector with fixed size. Each sentence is pre-processed and tokenized. Thus, the vocabulary is created, and multiple occurrences of the same word are removed by an iterative process. The words of the entire corpus are counted and stored in a vector with a length of the total number of words in the corpus. An example of how the feature vector looks like is shown in Figure 3.26, where three different sentences with similar words are presented. Each row in the feature vector corresponds to one sentence, and to each column to the different words in the whole corpus. Note that each position contains the number of occurrences of an specific word in the sentences or document.

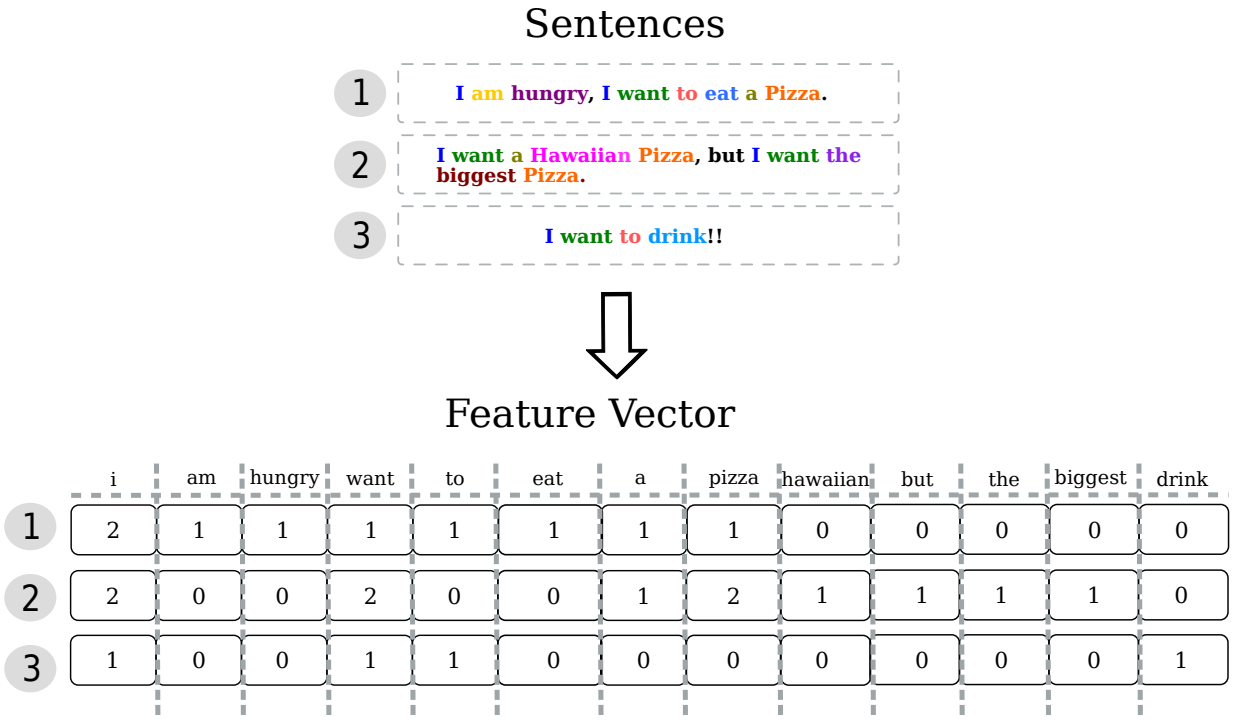

Figure 3.26: Example of a generated feature vector using BoW modeling

BoW only considers if a known word occurs in a document or not. The more similar words in two documents, the more similar the documents can be. There are two main limitations in BoW

modeling: (1) this method disregarding the order in which the words appear, i.e., it does not take into account the embedded contextual information, and (2) for large documents, it may result in a vector with lots of null values, known as sparse vector.

3.4.3 Term Frequency-Inverse Document Frequency

TF gives the relative frequency of a specific term, word or combination of words in a document. These values are compared to the relative frequency of other terms in a text or document. Then, to each term in a document a weight for that term is assigned, that depends on the number of occurrences t of it in the document. Thus, a simple to compute a score based on the ratio is assigned between t and the number of words in the whole corpus d using Equation 3.72, where $TF_{i,j}$ is the term frequency of a term i in a document j.

$$TF_{i,j}(t,d) = \frac{t_{i,j}}{d} \tag{3.72}$$

The limitation of TF is that it only considers all the query terms with equal importance. In fact, certain terms have little or no discriminating power to determine relevance. IDF acts as TF corrector factor for this issue. It attenuates the effect of the query terms that occur frequently in the corpus to be meaningful for relevance determination. IDF scales down the weights of the term with high frequency by comparing the number of all available documents n and the number of documents that contain the term to analyze df_i. In other words, IDF determines the relevance of the text with respect a specific word as:

$$IDF_i(df) = \log\left(\frac{n}{df_i}\right) \tag{3.73}$$

TF-IDF is based on a heuristic intuition in which a query term that occurs in many documents is not the best discrimination and should be given less weight than one which occurs in a few documents. Then, the definitions of TF and IDF are combined using a simple product in order to produce a composite weight for each term:

$$TFIDF = TF \times IDF \tag{3.74}$$

TF-IDF will be higher when the term occurs within a small number of documents, an it will be lower when it occurs less frequently in a document or in many documentsdocuments. In the case the word occurs in all documents will be the lowest one. Note that as same as BoW the exact ordering of the terms in a document is ignored but the number of occurrences of each term is

important.

3.4.4 Word2Vec

An alternative for word representation is vector space models. These models pretend to represent the words as a vector in a multidimensional space, where similar or related words are represented by nearby points. A well-known method is W2V, which consists of a Neural Network (NN) with one hidden layer. There are two main W2V models: Continuous Bag of Words (CBoW) and Skip-Gram (SG). This thesis only considers CBoW, which aims to predict a word based on a given context.

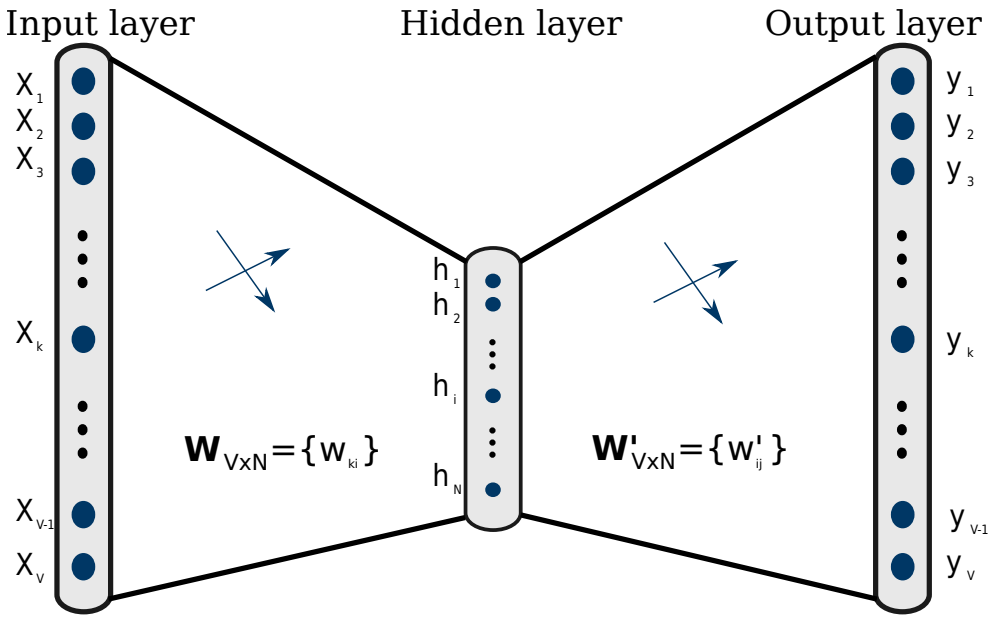

Figure 3.27: W2V-CBoW model using one word for context

W2V-CBoW is implemented with a fully connected neural network with a single hidden layer, which only takes one word for the context as shown in Figure 3.27. The input words are represented as one-hot encoding $\boldsymbol{X} = \{x_1, x_2, x_3, ..., x_v\}$, i.e., as binary vectors. An example of one-hot encoding representation is shown in Figure 3.28, where a vocabulary of all the words in the text is created and a word is encoded as a vector of the same dimensions of the vocabulary. The vocabulary in this example is only the words in the sentence “I want a Hawaiian Pizza”. Then, each dimension corresponds to a word in the vocabulary, in order to have a vector with all zeros and a 1, which represents the corresponding word (“Hawaiian”).

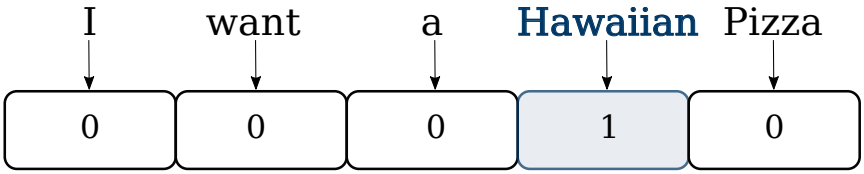

Figure 3.28: Example of one-hot encoding representation

Then, the size of the hidden layer is set to the dimensionality of the resulting word vectors. The vocabulary size V and the number of the linear neurons N are the hyper-parameters related with the hidden layer size. The connections from hidden layer to output layer can be described by a weight matrix $\boldsymbol{W}_{V\times N}$ with V rows, and N columns (see Equation 3.75). The input to the network considers only one word for context and it is encoded using one-hot encoding representation meaning that only one input row is set to one and rest of the input rows are set to zero. Hence, a vector $\boldsymbol{w}_{ki}$ is obtained which represents the associated word i in the input layer, and other rows from the weight matrix are ignored. Then, $\boldsymbol{w}_{ki}$ is passed through a softmax activation. The activations of the hidden layer are stored, being those "the word vectors".

$$\begin{aligned}\boldsymbol{h} = \boldsymbol{X}^T\boldsymbol{W} = \begin{bmatrix}x_1, x_2, x_3, \cdots, x_k, \cdots, x_v\end{bmatrix} \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1N} \\ w_{21} & w_{22} & \cdots & w_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ w_{k1} & w_{k2} & \cdots & w_{kN} \\ \vdots & \vdots & \vdots & \vdots \\ w_{v1} & w_{v2} & \cdots & w_{vN} \end{bmatrix} \\ = \begin{bmatrix}x_k w_{k1}, x_k w_{k2}, x_k w_{k3}, \cdots, x_k w_{kN}\end{bmatrix} = \begin{bmatrix}w_{k1}, w_{k2}, w_{k3}, \cdots, w_{kN}\end{bmatrix} := \boldsymbol{W}_{ki}\end{aligned} \tag{3.75}$$

Multiple context words can be used to do the same, thus neighbor words in this representation are considered to train the NN. Figure 3.29 shows the model that takes C context words that is similar to a "window size".

The selection of the neighbor words depends on the C, which is necessary to model the temporal context of each word. The temporal context is obtained by computing the average of the activations of the hidden layer over all of the words in the the context. C was set at 7, because it is the average number of words per sentence in the transcripts in this thesis.

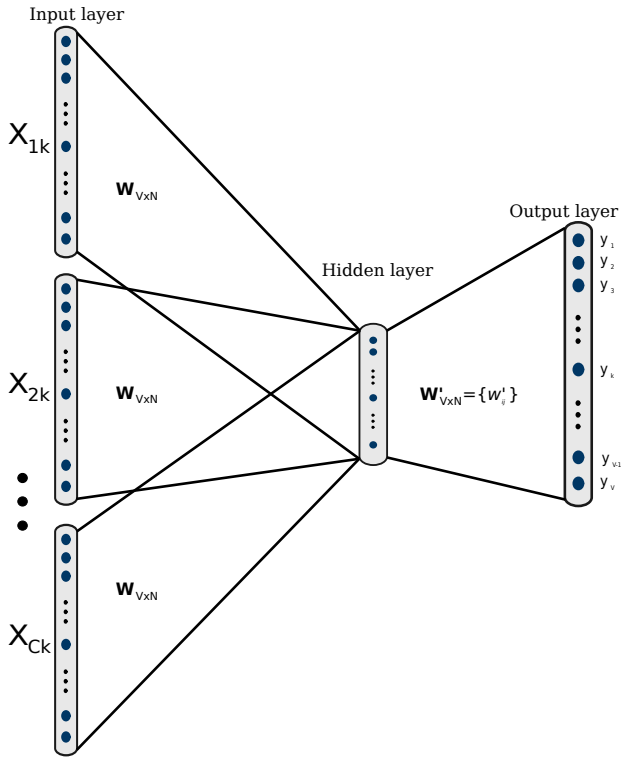

Figure 3.29: W2V-CBoW model using multiple words for context

3.4.5 Bidirectional Encoder Representations from Transformers

BERT is an unsupervised and deeply bidirectional pre-trained model proposed in [72]. Unidirectional models consider previous words to predict a target word, unlike bidirectional models which use the previous and the following words. This method is based in transfer learning and transformers models. The idea of transfer learning because the model is first trained on two unsupervised tasks. The first one is based on Masked Language Modeling (MLM), which aims of predicting a missing word in a sentence. The second one is Next Sentence Prediction (NSP), where the system is trained to predict whether a sentence follows another. BERT allows the words in the corpus to be represented into lower dimensional feature vectors using "Transformers", which is an attention-based encoder-decoder type first proposed in [90].

The transformer processes all elements simultaneously by forming direct connections between individual elements through a process known as attention. The transformer is composed by encoder and decoder layers with connections in between. BERT only considers the encoder layer that a high level maps an input sequence into an abstract continuous representation that holds and comprises all the learned info in the input.

Figure 3.30 shows the overall architecture used by the MLM tasks in BERT, where 15% of the

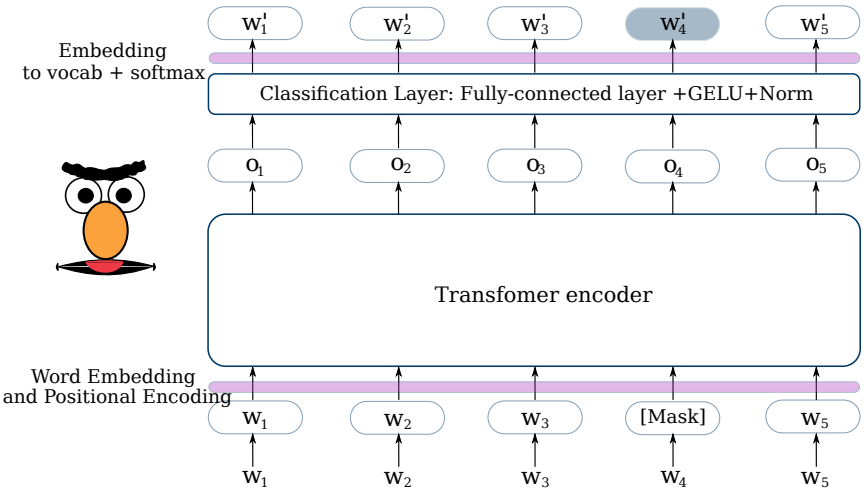

Figure 3.30: Masked language modeling architecture for BERT

words in each sequence are replaced with a [MASK] token, in order to predict the original value of the masked words, based on the context provided by the other words that are not masked. Each of the special tokens and words have a specific predefined ID giving by each model, whether the model does not know an specific word represented by ## at the beginning, the unknown token ([UNK]) is added. Since a RNN is not used and the transformer encoder reads the entire sequence of words at once, the system needs to know the positions of each word in order to perform the relations between the words. Thus, a positional encoding layer is considered for this task, which injects positional information into the embeddings by sine and cosine functions. For odd time steps a cosine function satisfies Equation 3.76, where pos refers to the position, i the dimension and $d_{model} = 512$ is the maximum length of a sequence, and for even time steps a sine function using Equation 3.77. Cosine and sine functions were chosen because they have linear properties that allow to the model an easily learn.

$$PE(pos, 2i + 1 = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{3.76}$$

$$PE(pos, 2i = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{3.77}$$

The Transformer encoder in BERT is a stack of 12 encoder layers, each of which operates on the output of the layer that came before. The encoder layer (see Figure 3.31) is composed by a multi-head attention module, followed by a residual connection with a layer normalization, a feed forward network composed by 2 fully connected layer with a ReLu in between, to finally apply a residual connection with a layer normalization.

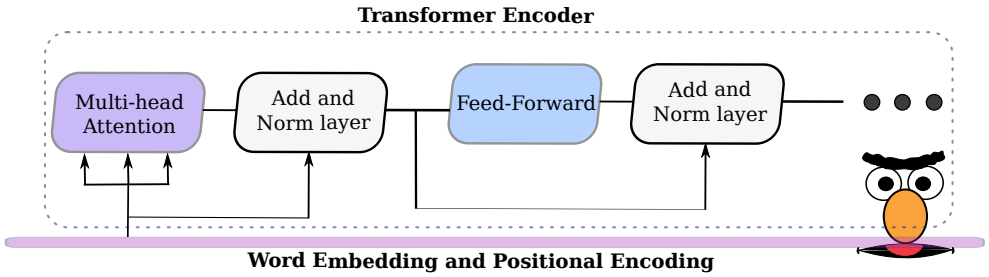

Figure 3.31: Transformer encoder for BERT

BERT actually learns multiple attention mechanisms, called heads, which operate in parallel to one another. The model can capture a broader range of relationships between the words via multi-head attention. In Figure 3.32, multi-head attention applies multiple self-attention mechanisms. This allows to look at each individual word in different windows for a better encoding. The first step to compute self-attention is to feed the input into three fully connected layers in order to create a query, key, and value vector for each word. The matrices Queries (Q), Keys (K) and Values (V) are the set of the aforementioned vectors, where the rows correspond to the words and the columns to the weight dimension (64). It is necessary to score each word of the input sequence against the actual word. The score matrix is calculated by taking a dot product matrix multiplication between Q and K. It allows to know how much focus should a word put on the other words. Next, the scores are scaled down by dividing the square root of the Q and K dimension, which allows for more stable gradients, and then the result is passed through a softmax to normalize the scaled scores. The above process can be summarized by Equation 3.78. The obtained scores are multiplied by the Values, for finally passing the result across a fully connected layer.

$$Z = Softmax\left(\frac{Q \times K^T}{\sqrt{64}}\right) V \tag{3.78}$$

In multi-head attention, it is necessary to split the matrices into adding vectors before applying self-attention. Each of these vectors goes to the same self-attention process separately, known as head. Note that attention heads do not share parameters, each head learns a unique pattern attention. The used BERT model is known as BERT-Base, that consists of 12 encoder layers and 12 heads, where it has 144 attention mechanisms.

Additionally, for NSP task, BERT accepts as special tokens [CLS] and [SEP]. [CLS] marks the beginning of the sequence and [SEP] marks the sentence boundary. For instance, in Figure 3.33 is shown an example of a compound sentence. The compound sentence starts with the [CLS] token and is delimited by [SEP], which indicates the end of each sentence. The difference with respect to the MLM task in the word embedding and positional encoding process, it is that the

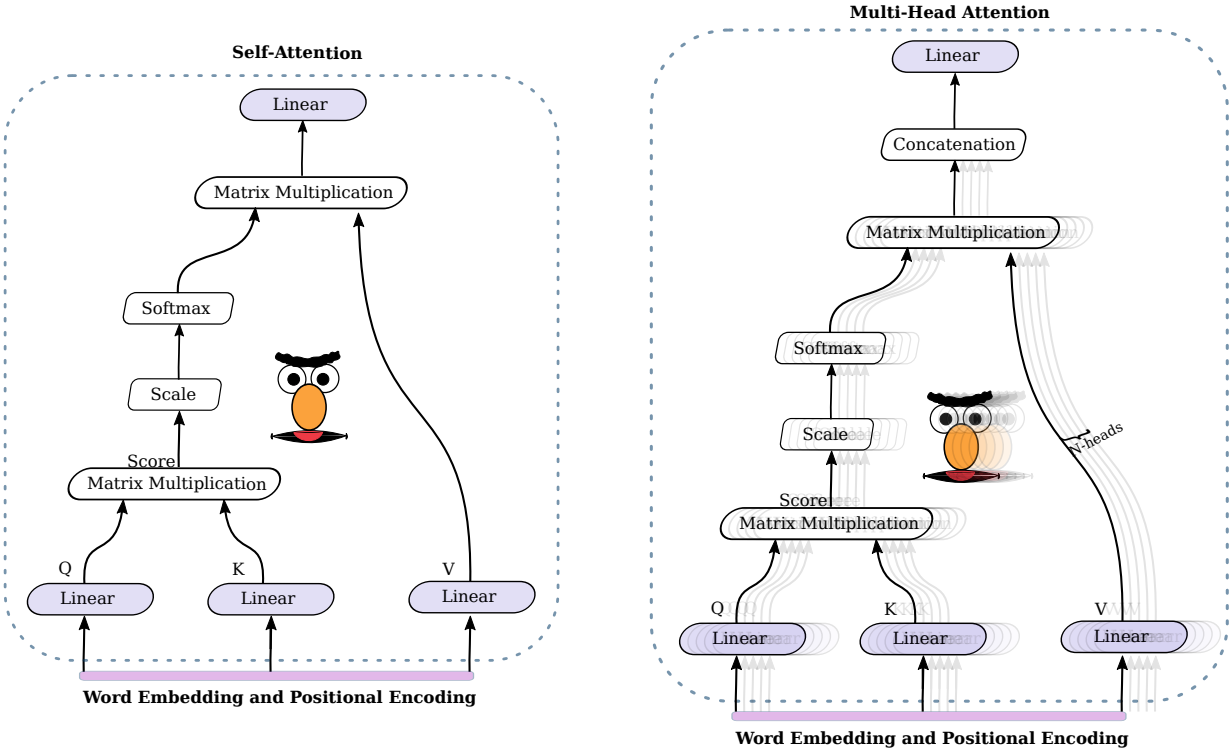

Figure 3.32: Self attention to the left and Multi-Head attention to the right

sentence positional encoding is added to indicate which sentence each word belongs to. In this thesis, BERT is used in order to extract the feature embeddings of the transcriptions.

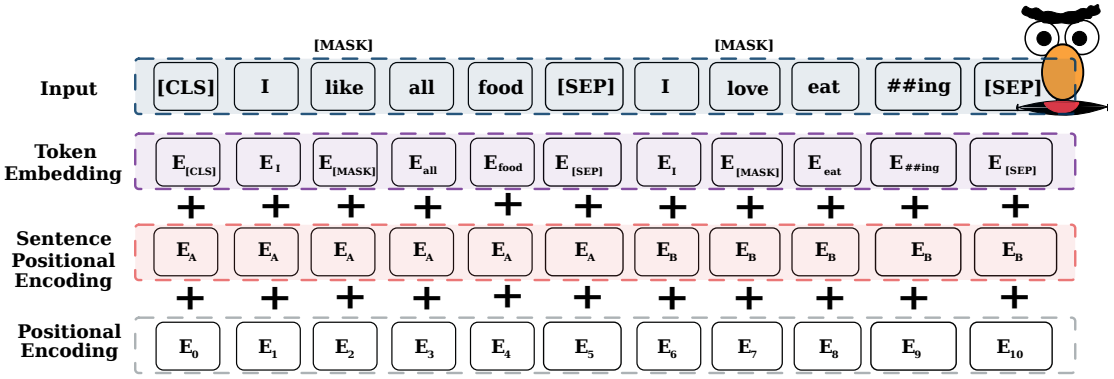

Figure 3.33: Word embedding and positional encoding process in Next Sentence Prediction task for BERT.

Chapter 4

Datasets

4.1 Call-center Datasets

These datasets consist of recordings from Colombian banking and insurance call-centers. The interactions were acquired with a sampling rate of 8 KHz. They were annotated according to two different emotion labels: positive (satisfied) and negative (unsatisfied). The calls were transliterated following the verbatim protocol, using headphones to maximize the transcription accuracy. The banking call-center dataset consists of a total of 2363 recordings, 1285 male and 1078 female subjects. The customer opinions were recorded in voicemail. After labeling, 1093 utterances for Satisfied Customers-Banking Call Center (SC-BC) and 1270 for Dissatisfied Customers-Banking Call Center (DC-BC) were obtained. The whole vocabulary of the lemmatized transcriptions without stopwords (e.g., the, of, in, on, etc.) contains 4757 words. The average duration of the recordings is 14$\pm$9 seconds for SC-BC and 29$\pm$21 for DC-BC. The insurance call-center dataset contains 283 recordings, 111 male and 172 female subjects. After labeling, 229 interactions for Satisfied Customers-Insurance Call Center (SC-IC) and 54 for Dissatisfied Customers-Insurance Call Center (DC-IC) were obtained. The whole vocabulary of the lemmatized transcriptions without stopwords (e.g., *the*, *of*, *in*, *on*) contains 4926 words. The average duration of the interactions is 1136$\pm$481 seconds for SC-IC and 1257$\pm$662 for DC-IC.

4.2 Genetic Alzheimer's Dataset

The E280A or *Paisa mutation* [16] that is the most common cause of genetic Early-Onset Alzheimer's EOA in Colombia. The most common symptoms of this mutation are memory deficits in the third decade of life, develop of progressive cognitive impairments such as verbal

disfluency, changes in personality and behavior, among others. Nevertheless, there are some phenotype variations such as epilepsy, cerebellar ataxia, verbal impairment, gait difficulties, Parkinsonism, among others. This mutation affects 25 extended families with more than living 5,000 members who historically lived in isolated regions of Andes mountains in the Colombian state of Antioquia. Nowadays, approximately one half the living members of this kindred live in Medellín (the capital of Antioquia) and the rest live in nearby towns in Antioquia. This population is remarkable for its unusual size and for the high level of participation of longitudinal studies. The members of this kindred can be Genetic Carriers (GC) or Non-Genetic Carriers (NGC). GCs inherit this mutation but they do not show any symptom of AD, however, GCs are able to pass the allele onto their offspring. This database is being recorded since 2018 in the University of Antioquia by Grupo de Neurociencias de Antioquia and GITA Lab. The data consist of spontaneous speech recordings and their transliterations from 114 Spanish speakers from Colombia, 28 asymptomatic subjects belonging to families with the *Paisa mutation* that are GC and 36 that are NGC, 23 MCI patients with EOA, and 27 HC subjects. The task consisted of asking the participants to describe the cookie theft picture [56]. The average duration of the recordings is 84±48 seconds for the GC subject, 83±42 seconds for the NGC subject, 53±25 for the MCI patients, and 42±19 for HC subjects. The transcriptions were produced by a professional for linguistics following the verbatim protocol, using headphones to maximize the transcription accuracy. The whole vocabulary of the lemmatized transcriptions without stopwords (e.g., the, of, in, on, etc.) consists of 922 words. The data was labeled by expert listeners according to the MMSE and MoCA scales. Additional information of the participants including age, gender, and their neurological state are included in Table 4.1.

Table 4.1: General information of the subjects in the Genetic Alzheimer's Dataset

	MCI	HC	GC	NGC
Gender [F/M]	12/11	15/12	16/12	22/14
Age [F/M]	48.1(5.5)/51.0(7.9)	49.5(7.7)/53.2(7.7)	31.5(5.6)/31.8(5.0)	32.1(6.4)/33.5(4.8)
Education [F/M]	5.8(3.4)/7.3(6.2)	7.3(3.6)/8.8(4.6)	10.6(2.8)/11.0(3.3)	13.5(2.6)/12.4(2.7)
MMSE [F/M]	25.0(4.3)/25.7(2.3)	28.5(1.4)/28.8(1.2)	29.4(0.8)/29.5(0.9)	29.5(0.9)/29.6(0.8)
MoCA [F/M]	14.9(5.2)/15.7(4.0)	20.7(4.2)/22.3(5.4)	24.5(2.3)/24.3(2.7)	25.5(2.3)/25.9(2.7)

MCI patients: AD patients with mild cognitive impairment. **GC** subjects: asymptomatic genetic carriers. **NGC** subjects: asymptomatic non-genetic carriers. **HC** subjects: healthy control subjects. Values are expressed as mean (standard deviation). F: female. M: male. Age and education are given in years.

4.3 Alzheimer's Dementia Recognition through Spontaneous Speech Dataset

The Alzheimer's Dementia Recognition through Spontaneous Speech (The ADReSS) dataset was created for the Alzheimer's speech classification task in the Interspeech ADReSS challenge 2020 [91]. The participants were native English speakers, matched for age and gender. It consist of spontaneous speech recordings and transcripts describing the cookie theft picture [56]. The recordings consider full enhanced audio and normalized sub-chunks, which were segmented using voice activity detection algorithm based on a signal energy thresholding. They were normalized across all speech segments to control the variation caused by the recording conditions. This thesis uses the same training (108 recordings) and test (48 recordings) set provided by the challenge. Additional information of the HC subjects, and AD patients considering age, gender, and MMSE are included in Table 4.2.

Table 4.2: General information of the subjects in the ADReSS dataset.

	AD patients	HC patients
Number of subjects [F/M]	43/35	43/35
Age [F/M]	66.7(6.0)/66.5(7.6)	66.6(6.0)/65.9(7.3)
MMSE [F/M]	16.8(5.0)/19.0(5.8)	29.0(1.3)/29.0(1.0)

AD: Alzheimer's disease patients. **HC**: healthy control subjects.
Values are expressed as mean (standard deviation). F: female. M: male.
Age is given in years.

4.4 PC-GITA

The transcriptions from the recordings of spontaneous speech of 50 PD patients and 50 HC subjects are considered in this thesis to analyze the suitability of the NLP methods to discriminate the disease. The speech recordings are part of the PC-GITA corpus [92]. For this thesis only recordings of monologues are considered. The task consisted of asking the participants to talk about their daily routines. The average duration of the monologues is 48$\pm$29 seconds for the patients and 45$\pm$24 for the healthy subjects. The transcriptions were produced following the verbatim protocol, using headphones to maximize the transcription accuracy and to minimize possible human errors. The whole vocabulary of the lemmatized transcriptions without stopwords (e.g., the, of, in, on, etc.) consists of 1182 words. The patients were evaluated by an expert

neurologist and labeled according to the MDS-UPDRS-III score. Table 4.3 shows additional information of the subjects.

Table 4.3: General information of the subjects in the PC-GITA dataset. Time since diagnosis, age and education are given in years. ^{a}p calculated through Chi-square test. ^{b}p calculated through t test.

	PD patients	HC subjects	Patients vs. controls
Gender [F/M]	25/25	25/25	p=1.00^{a}
Age [F/M]	60.7(7.3)/61.3(11.7)	61.4(7.1)/60.5(11.6)	p=0.98^{b}
Education [F/M]	11.5(4.1)/10.9(4.5)	11.5(5.2)/10.6(4.4)	p=0.88^{b}
Time since diagnosis [F/M]	12.6(11.5)/8.7(5.8)		
MDS-UPDRS-III [F/M]	37.6(14.0)/37.8(22.1)		

PD patients: Parkinson's disease patients. **HC** subjects: Healthy control subjects. Values are expressed as mean (standard deviation). F: female. M: male. Time since diagnosis, age and education are given in years. ^{a}p calculated through Chi-square test. ^{b}p calculated through t-test.

4.5 Depression in Parkinson's Disease

The data considered transliterations of spontaneous speech from 60 Spanish speakers from Colombia, 25 D-PD patients and the remaining 35 ND-PD patients. They were labeled according to the depression item of the MDS-UPDRS-I. D-PD are the patients with the item higher than zero and ND-PD patients are those with the item equal to zero. The participants were requested to talk about their daily routines. After performing the tasks, the patients were immediately evaluated by the neurologist. The average duration of the monologues is 84±34 seconds for the D-PD patients and 80±37 for the ND-PD patients. The transcripts were obtained from audio files that were manually transcribed following the verbatim protocol, using headphones to maximize the transcription accuracy. The whole vocabulary of the lemmatized transcriptions without stopwords contains 1240 words. Additional information of the PD patients including age, gender, education level, time since diagnosis, and their neurological state are included in Table 4.4.

Table 4.4: General information of the subjects in the Depression in Parkinson's Disease dataset

	D-PD patients	ND-PD patients	D-PD vs. ND-PD
Number of subjects [F/M]	15/10	17/18	p =0.99^a
Age [F/M]	66.1(10.3)/68.8(9.1)	60.3(13.2)/66.3(10.3)	p =0.24^b
Education [F/M]	9.8(3.8)/11.1(5.1)	11.2(5.1)/13.8(4.1)	p =0.05^b
Time since diagnosis [F/M]	6.4(5.3)/8.1(4.5)	13.7(12.7)/7.3(4.8)	p =0.26^b
MDS-UPDRS-III [F/M]	33.3(18.8)/35.3(16.6)	26.5(10.2)/36.3(1.7)	p =0.81^b
MDS–UPDRS–Depression [F/M]	1.4(0.7)/1.3(0.7)	0.0/0.0	p =0.05^b

D-PD patients: Depressive Parkinson's disease patients. **ND-PD** patients: Non-Depressive Parkinson's disease patients. Values are expressed as mean (standard deviation). F: female. M: male. Time since diagnosis, age and education are given in years. $^a p$ calculated through Chi-square test. $^b p$ calculated through Mann-Whitney U-test.

4.6 Interactive Emotional Dyadic Motion Capture

The IEMOCAP [93] is an acted, multi-speaker and multimodal dataset collected by the University of Southern Carolina. It contains 12 hours of audios, dialog transcripts, video and motion capture of dyadic interactions between ten different English native speakers. This consists of five sessions, which contain a total of 10039 utterances and an original sampling rate of 16kHz. Each session is displayed by a pair of speakers (male and female) in scripted and improvised scenarios. The utterances are transliterated by 3 human annotators to reduce the transcription error. They are discretely labeled according to 9 categorical attributes: anger, happiness, excitement, sadness, frustration, fear, surprise, other and neutral state. Additionally, this dataset contains dimensional scaled attributes from 0 to 5 such as valence, arousal, dominance. In this thesis, they are centered in 2.5 in order to divide the quadrants of the arousal-valence plane. Notice in Table 4.5 that the data is highly imbalance especially for classes related to negative valences. This problem is addressed by performing data augmentation in both modalities. For the speech recordings, the amount of data was augmented by adding Gaussian noise with different SNR (20 dB, 30 dB, 40 dB) to every signal. For the transcripts, the data augmentation was performed by means of translate the text from English to Spanish and then from Spanish to English.

Table 4.5: Number of samples after data augmentation for speech and text in IEMOCAP dataset.

Task	**Arousal**		**Valence**		**Quadrants**			
	AA	**PA**	**PV**	**NV**	**AP**	**AN**	**PN**	**PP**
Original size	3427	1288	4082	633	2990	437	196	1092
DA for Speech	26472	13684	35032	5124	22968	3504	1620	12064
DA for Text	13236	6842	17516	2562	11484	1752	810	6032

DA: Data Augmentation. **AA**: Active Arousal. **PA**: Passive Arousal. **PV**: Positive Valence. **NV**: Negative Valence. **AP**: Active Positive. **AN**: Active Negative. **PN**: Passive Negative. **PP**: Passive Positive.

Chapter 5

Experiments and results

This thesis proposes multimodal analysis based on acoustic and linguistic information to assess different scenarios: (1) customer satisfaction, and (2) neuro-degenerative diseases. This section is divided into five different experiments: (1) Evaluation of customer satisfaction, (2) Assessment of Genetic AD, (3) Linguistic analysis to discriminate PD, (4) Depression in Parkinson's disease, and (5) User state modeling based on arousal-valence plane for customer satisfaction and health-care.

5.1 Evaluation of Customer Satisfaction

In this experiment, customer satisfaction analysis is performed considering two different datasets related to insurance and banking call-centers separately. The datasets were presented in Section 4.1. The analysis is based on the discrimination of satisfied and dissatisfied customer interactions.

5.1.1 Methodology

The general methodology addressed in this experiment is illustrated in Figure 5.1. Articulation and prosody features are considered for speech, while W2V, BERT, and the Spanish version of BERT (BETO) are extracted for linguistic analysis.

Acoustic Analysis

The acoustic analysis in this study is based on the articulation and prosodic features proposed previously [94], [95]. The onset and offset transitions are detected based on the presence of the fundamental frequency. Chunks of chunks of 40 ms a shift of 10 ms of speech are taken to the left and to the right of each border to form offsets and onsets.

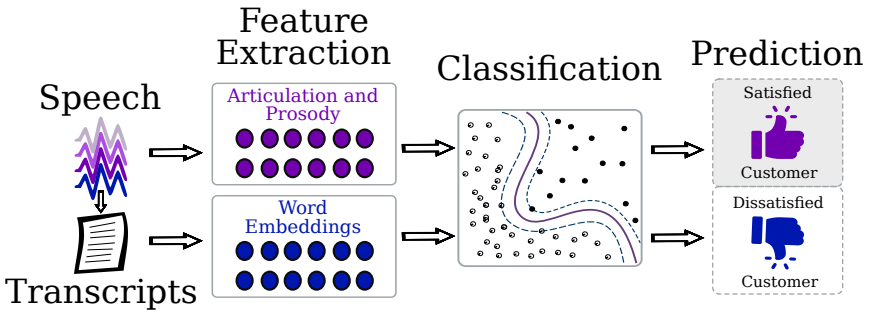

Figure 5.1: Scheme of the methodology addressed in this thesis for the evaluation of customer satisfaction

Articulation analysis is based on the energy content and the formant frequencies. The energy content is modeled considering 22 frequency bands according to the Bark scale, and 12 MFCCs along with their derivatives. The feature set is completed with the first two formant frequencies and their derivatives, computed from the voiced segments. A total of 122 descriptors are extracted (see Table 5.1). Four statistical functionals are computed for all articulation descriptors: mean, standard deviation, kurtosis and skewness, forming a 488-dimensional feature vector per utterance. Prosody analysis is based on the fundamental frequency contour, the energy, and the duration. F_0 and energy contour are modeled considering the tilt, the mean square error (MSE), and the first and last voiced and unvoiced segments. Features based on duration consider voiced rate, duration of pauses and ratios. A total of 103 descriptors are extracted (see Table 5.1).

Linguistic Analysis

The linguistic analysis in this study is based on word-embeddings such as W2V, BERT, and BETO. Four functionals (mean, standard deviation, kurtosis and skewness) are computed over all word-embeddings for each method to form a static vectors for each speaker. Table 5.2 shows the number of computed NLP features for each method.

5.1.2 Optimization and Classification

The classification is performed using a Radial Basis Function-Support Vector Machine (RBF-SVM). The validation process for the banking call-center dataset follows a bootstraping strategy of 70%-15%-15% as is shown in Figure 5.2.A., where 70% of the data is used for training, 15% to optimize the hyper-paramaters of the SVM, and 15% for testing. The experiments in the insurance call-center dataset follows a 5-fold cross-validation strategy as illustrated in Figure 5.2.B. The data is divided into $k = 5$ subsets, one subset is used to test the model, and the rest ($k - 1 = 4$)

Table 5.1: List of computed acoustic descriptors

	Articulation features
1-22	Bark band energies in onset transitions
23-34	MFCCs in onset transitions
34-46	First derivative of the MFCCs in onset transitions
47-58	Second derivative of the MFCCs in onset transitions
59-80	Bark band energies in offset transitions
81-92	MFCCs in offset transitions
93-104	First derivative of the MFCCs in offset transitions
105-116	Second derivative of the MFCCs in offset transitions
117	First formant Frequency
118	First Derivative of the first formant frequency
119	Second Derivative of the first formant frequency
120	Second formant Frequency
121	First derivative of the Second formant Frequency
122	Second derivative of the Second formant Frequency
	Prosodic features
1-6	F_0 contour
7-12	Tilt of a linear estimation of F_0 for each voiced segment
13-18	MSE of a linear estimation of F_0 for each voiced segment
19-24	F_0 on the first voiced segment
25-30	F_0 on the last voiced segment
31-34	Energy contour for the voiced segments
35-38	Tilt of a linear estimation of energy contour for voiced segments
39-42	MSE of a linear estimation of energy contour for voiced segment
49-54	Energy on the last voiced segment
55-58	Energy-contour for unvoiced segments
59-62	Tilt of a linear estimation of energy contour for unvoiced segments
63-66	MSE of a linear estimation of energy contour for unvoiced segments
67-72	Energy on the first unvoiced segment
73-78	Energy on the last unvoiced segment
79	Voiced rate
80-85	Duration of voiced segments
86-91	Duration of unvoiced segments
92-97	Duration of pauses
98-103	Duration of ratios

Table 5.2: List of computed linguistic descriptors

	Linguistic features
1-100	W2V
100-868	BERT
868-1636	BETO

are used to train the model and to perform an internal 4-fold cross-validation to optimize the hyper-paramaters of the SVM. The optimal parameters of the RBF-SVM are found through a grid search where $C \in \{10^{-4}, 10^{-3}, ..., 10^{4}\}$ and $\gamma \in \{10^{-4}, 10^{-3}, ..., 10^{4}\}$. The optimization criterion is the f-score obtained in the development set. The classification considers each feature set and the combination using an early fusion strategy to merge linguistic and acoustic features before performing the classification.

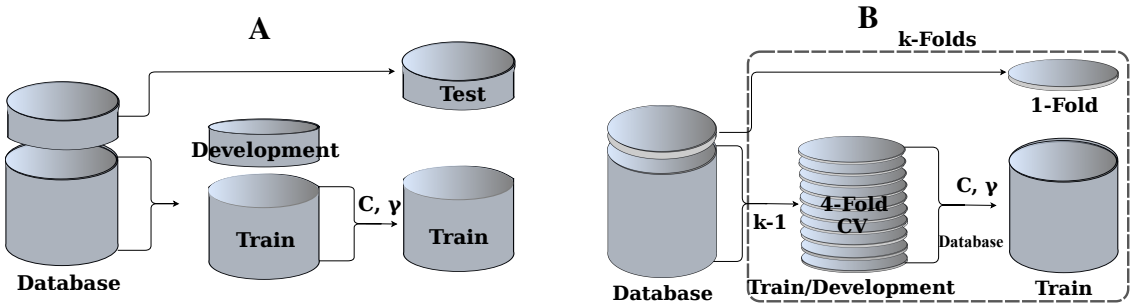

Figure 5.2: Database distribution for the evaluation of customer satisfaction: A) Bootstraping strategy for customer satisfaction in banking call-centers. B) Cross-validation strategy for customer satisfaction in insurance call-centers. **CV:** cross-validation. **N:** number of samples.Database distribution. **CV:** cross-validation. **N:** number of samples.

5.1.3 Results and Discussion

Two different experiments are performed related to customer satisfaction: (1) automatic classification of customer satisfaction in banking call-centers, (2) automatic classification of customer satisfaction in insurance call-centers. The same procedure to extract features is considered for both datasets. The performance of the classifiers is evaluated according to their F-score, Unweighted Average Recall (UAR), Sensitivity (Sens), Specificity (Spe), and the Area Under the receiver operating characteristic Curve (AUC).

Banking Call-Center

This experiment is performed using the banking call-center dataset presented in Section 4.1. Five different feature sets are considered for acoustic and linguistic analyses: (1) articulation, (2) prosody, (3) W2V, (4) BERT, and (5) BETO. The results for each feature set separately are shown in Table 5.3. Acoustic features show the highest results with 0.70 of F-score for articulation, and 0.72 for prosody. Linguistic results are the less accurate with an average F-score of 0.51.

Table 5.3: Results for the banking call-center dataset using each feature set separately

Features	Experiment	F-score	UAR	Sens	Spe	AUC	C	γ
Acoustic	Articulation	0.70	70.4	59.4	80.5	0.76	10e-2	10e-4
	Prosody	0.72	72.4	72.4	72.4	0.81	10e+2	10e-5
Linguistic	W2V	0.52	52.7	45.9	58.9	0.52	10e0	10e-4
	BERT	0.51	51.5	39.4	62.7	0.51	10e+1	10e-5
	BETO	0.51	51.3	40.6	61.1	0.50	10e0	10e-5

Notes: **UAR**: unweighted average recall. **Sens**: sensitivity. **Spe**: specificity. **AUC**: Area under the ROC curve. **C** and γ are the optimal meta-parameters obtained in development for RBF-SVM. Unweighted average recall, Sensitivity and Specificity are given in [%].

Table 5.4 shows the results using early fusion to combine the different feature sets, without including BETO. Even though BERT and BETO produce a similar performance, in the fusion of features BERT obtained higher results in this experiment. Notice that highest performance in this table is obtained for the combination of articulation, prosody, and W2V, indicating not improvement in comparison to only using prosody.

The score distributions for each feature set, and the combination of articulation, prosody and W2V are shown in order to observe more information related to the behavior of the classifiers (see Figures 5.3 and 5.4). The scores correspond to the distance to the hyperplane in the RBF-SVM. The dark gray bars correspond to the scores for DC-BC, the white bars are the scores computed for the SC-BC, and the light gray bars correspond to the intersection between both sets, and reflect the classification errors. Note that for prosody the score range is wider in comparison with the other results, while when the early fusion between articulation, prosody and W2V is performed the score range is narrower.

The ROC curves are illustrated in Figure 5.5. These curves allow to show the results more compactly. The features related to acoustics obtained the highest performances, while the linguistics for this task have to be improved.

Table 5.4: Results for the banking call-center dataset using early fusion of the different feature sets.

Experiment	F-score	UAR	Sens	Spe	AUC	C	γ
Art-Pro	0.69	69.3	62.4	75.7	0.69	10e-2	10e-4
Art-W2V	0.69	69.0	58.2	78.9	0.75	10e-1	10e-4
Art-BERT	0.64	64.5	52.9	75.1	0.70	10e-1	10e-5
Pro-W2V	0.67	67.3	63.5	70.8	0.74	10e-1	10e-4
Pro-BERT	0.64	63.9	55.9	71.4	0.67	10e+2	10e-5
W2V-BERT	0.51	52.1	38.8	64.3	0.51	10e0	10e-5
Art-Pro-W2V	0.70	70.1	63.5	76.2	0.78	10e-1	10e-5
Art-Pro-BERT	0.67	67.6	57.1	77.3	0.74	10e-1	10e-5
Art-W2V-BERT	0.64	63.9	54.1	73.0	0.69	10e-1	10e-5
Pro-W2V-BERT	0.63	63.1	55.9	69.7	0.66	10e+1	10e-4
Art-Pro-W2V-BERT	0.67	67.6	59.4	75.1	0.73	10e-1	10e-5

Notes: **UAR**: unweighted average recall. **Sens**: sensitivity. **Spe**: specificity. **AUC**: Area under the ROC curve. **C** and γ are the optimal meta-parameters obtained in development for RBF-SVM. Unweighted average recall, sensitivity, and specificity are given in [%].

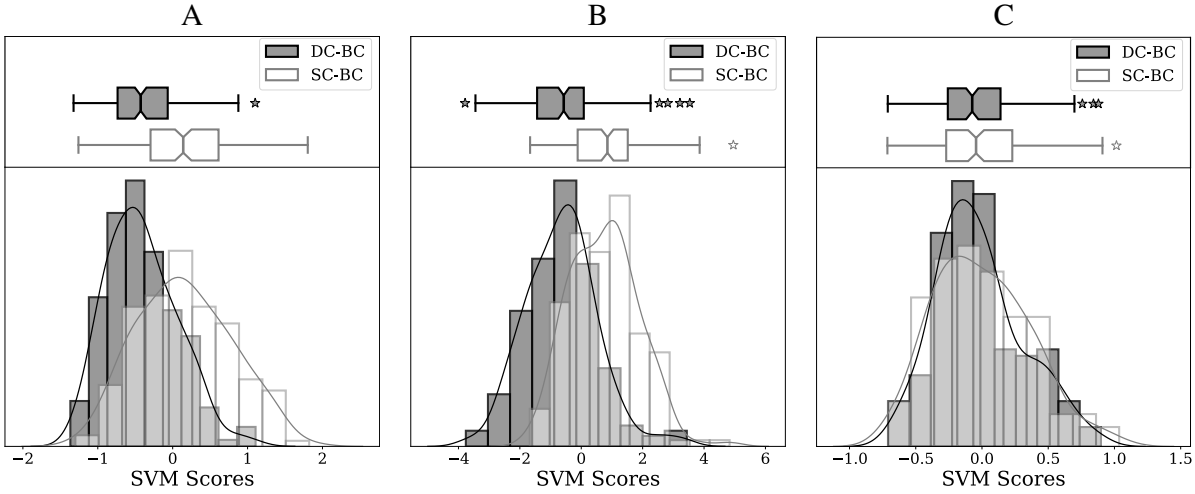

Figure 5.3: Scores for the banking call-center dataset obtained for: A) Articulation. B) Prosody. C) W2V.

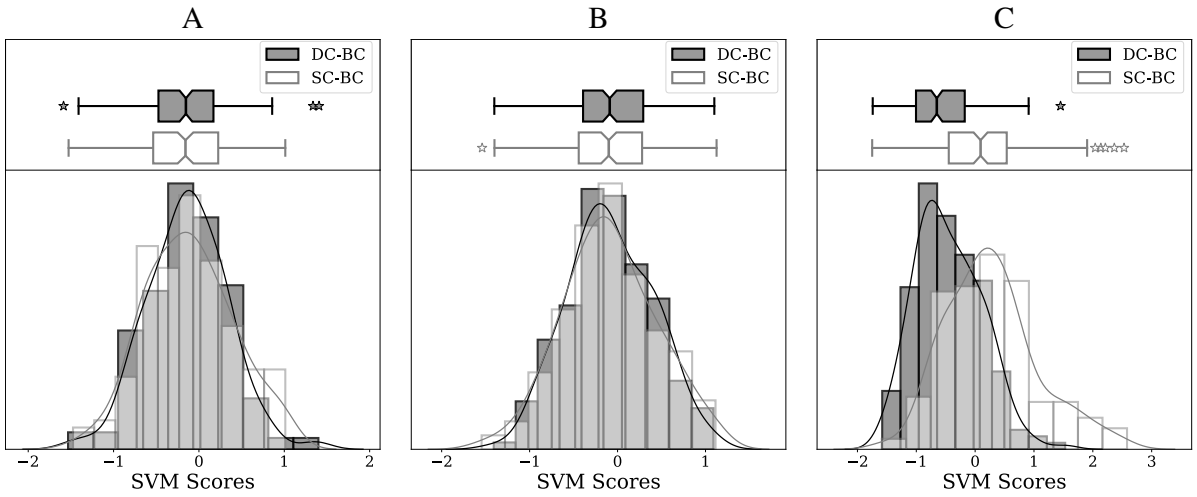

Figure 5.4: Scores for the banking call-center dataset obtained for linguistic features: A) BERT. B) BETO. C) Early fusion between articulation, prosody and W2V.

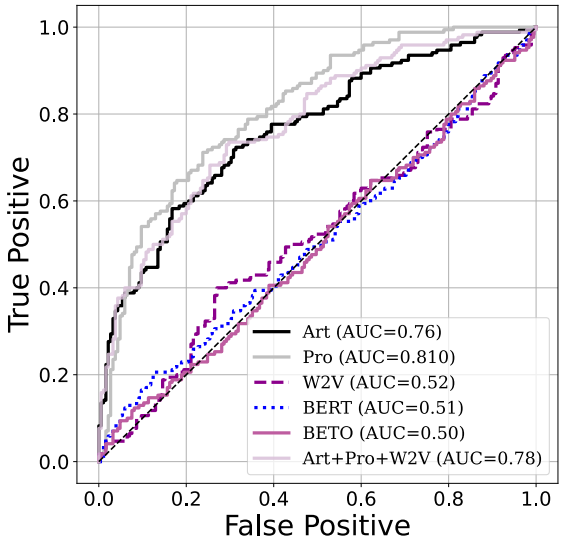

Figure 5.5: ROC Curve for the banking call-center dataset obtained for different feature sets. Art: articulation. Pro: prosody.

Insurance Call-Center

The same five feature sets and the insurance call-center (see Section 4.1) are considered for this experiment. Due to the unbalance in the data, a sub-set of the SC-IC class is chosen to match the size of the DC-IC set. The experiment is performed five times with different random sub-sets of SC-IC, reporting the average of these experiments. The results using each feature set separately are shown in Table 5.5. Unlike the experiments performed with the banking call-center dataset, linguistic features obtained the most accurate results, and BETO produce an slightly improvement (1%) in performance in comparison with BERT.

Table 5.5: Results for the insurance call-center dataset of each feature set separately

Features	Experiment	F-score	UAR	Sens	Spe	AUC	C	γ
Acoustic	Articulation	0.54	54.3	52.6	55.9	0.57	10e0	10e-4
	Prosody	0.56	56.1	60.7	51.5	0.58	10e1	10e-4
Linguistic	W2V	0.70	70.6	74.1	67.0	0.75	10e0	10e-5
	BERT	0.70	70.6	63.3	77.8	0.76	10e-1	10e-5
	BETO	0.71	71.3	65.9	76.7	0.76	10e0	10e-5

Notes: **UAR**: unweighted average recall. **Sens**: sensitivity. **Spe**: specificity. **AUC**: Area under the ROC curve. **C** and γ are the optimal meta-parameters obtained in development for RBF-SVM. Unweighted average recall, Sensitivity and Specificity are given in [%].

Table 5.6 shows the results using the early fusion strategy without considering BERT. The combination of features using BETO obtained higher performance for BERT in this experiment. The combination of prosody and BETO produce the most accurate results (F-score=0.73), where the sensitivity increased in comparison with using only BETO.

Figures 5.6 and 5.7 show the scores obtained for each feature set, and the combination of prosody and BETO. Similar as in the banking call-center experiment, the dark gray bars correspond to the scores for DC-IC, the white bars are the scores computed for the SC-IC. The combination of prosody and BETO improves the performance and decreases the number of outliers in the distribution in comparison with only using BETO.

The ROC curves obtained are shown in Figure 5.8.Note that similar results are not obtained for both approaches even though both are classifying satisfaction. It may be produced because the information is derived from two different sources, for the insurance call-center dataset the interaction between client and advisor, and for the banking call-center dataset customer opinions recorded in a voicemail.

Table 5.6: Results for the insurance call-center dataset using early fusion of the different feature sets.

Experiment	F-score	UAR	Sens	Spe	AUC	C	γ
Art-Pro	0.55	55.2	53.7	56.7	0.57	10e0	10e-4
Art-W2V	0.65	65.0	63.7	66.3	0.72	10e+1	10e-5
Art-BETO	0.72	72.0	67.8	76.3	0.79	10e-1	10e-5
Pro-W2V	0.70	69.8	71.1	68.5	0.75	10e0	10e-4
Pro-BETO	0.73	73.5	71.5	75.6	0.80	10e-1	10e-5
W2V-BETO	0.71	70.8	67.1	74.5	0.79	10e-1	10e-5
Art-Pro-W2V	0.67	67.0	65.9	68.1	0.72	10e+1	10e-5
Art-Pro-BETO	0.73	73.1	71.9	74.4	0.79	10e-1	10e-5
Art-W2V-BETO	0.72	71.9	70.7	73.0	0.79	10e-1	10e-5
Pro-W2V-BETO	0.71	70.8	69.0	72.7	0.79	10e-1	10e-5
Art-Pro-W2V-BETO	0.72	72.0	71.1	73.0	0.78	10e0	10e-5

Notes: **UAR**: unweighted average recall. **Sens**: sensitivity. **Spe**: specificity. **AUC**: Area under the ROC curve. **C** and γ are the optimal meta-parameters obtained in development for RBF-SVM. Unweighted average recall, sensitivity, and specificity are given in [%].

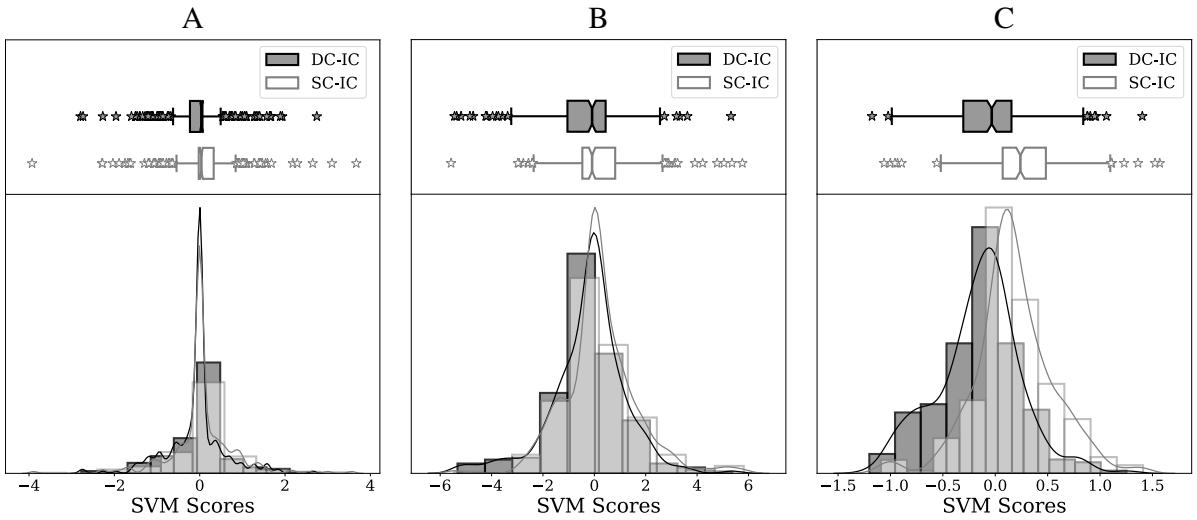

Figure 5.6: Scores for the insurance call-center dataset obtained for: A) Articulation. B) Prosody. C) W2V.

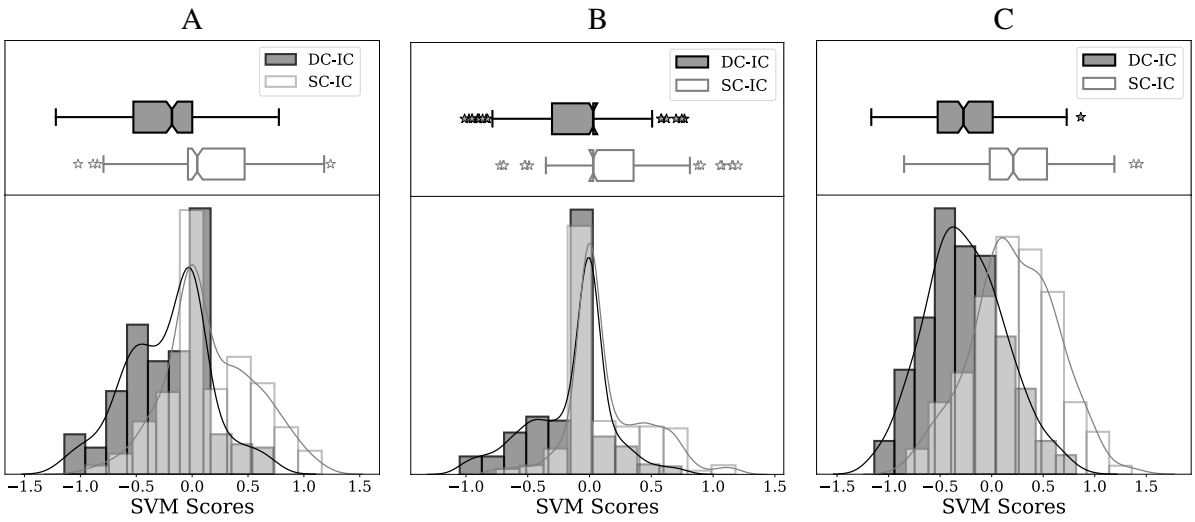

Figure 5.7: Scores for the insurance call-center dataset obtained for: A) BERT. B) BETO. C) Early fusion between prosody and BETO.

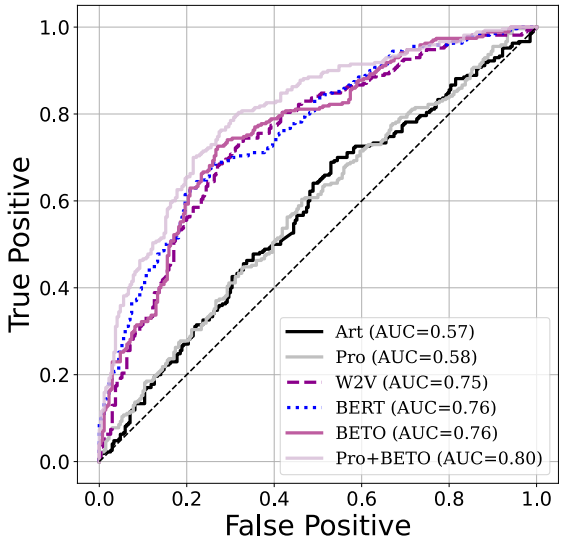

Figure 5.8: ROC Curve for the insurance call-center dataset obtained for different feature sets. Art: articulation. Pro: prosody.

5.2 Assessment of Genetic Alzheimer's disease

AD is highly characterized by the deterioration of the capability to produce coherent language that affects lexical, grammatical and semantic processes. In addition to language disorders, different studies have shown abnormalities in language production, characterized by the difficulty to access semantic information intentionally, which affects the speech fluency of the patients [25]. The dataset considered in this experiment is presented in Section 4.2. This thesis proposes the use of acoustic and NLP methods to extract features from transcriptions to discriminate between asymptomatic subjects belonging to families with AD that are Genetic Carriers (GC), that are not Genetic Carriers (NGC), and EOA patients with MCI.

5.2.1 Methodology

Figure 5.9 shows the general methodology addressed in this study. This experiments adopted the same methodology as in Section 5.1 regarding the extraction of acoustic and linguistic features. Four functionals (mean, standard deviation, kurtosis and skewness) are computed over all the acoustic and linguistic features sets to form a static vectors for each speaker.

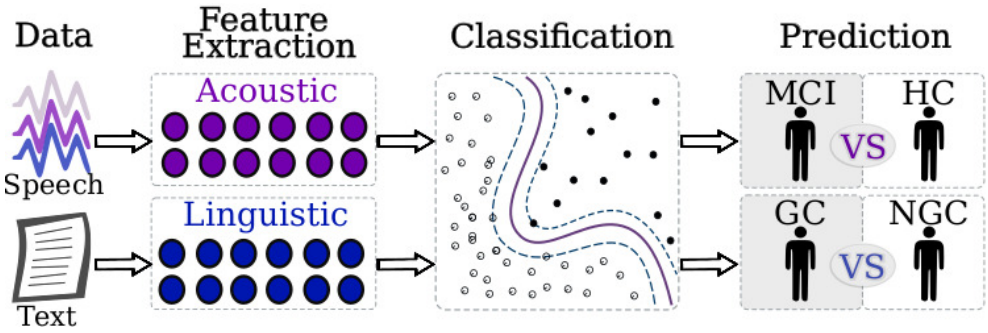

Figure 5.9: Scheme of the methodology addressed in this thesis for the assessment of Alzheimer's disease

5.2.2 Optimization and Classification

The classification is performed using an RBF-SVM. The validation process is a modification of the regular Leave-One-Speaker-Out (LOSO) strategy (see Figure 5.10). During the regular LOSO, the meta-parameters of the classifier have to be decided on the best parameters in the development for all the N speakers. However, these results are optimistic, since up to N parameter sets are found, i.e., N different classifiers. The proposed validation in this study uses the regular strategy with an

internal 6-fold cross-validation to optimize the hyper-paramaters of the SVM. N different optimal hyper-parameters will be obtained and stored. The found settings were sorted and the median of all of these values was obtained in order to have only one C and one γ. Finally, the LOSO strategy is performed again with the fixed parameters. The optimal parameters of the RBF-SVM are found through a grid search where $C \in \{10^{-6}, 10^{-3}, ..., 10^{5}\}$ and $\gamma \in \{10^{-6}, 10^{-3}, ..., 10^{5}\}$. The optimization criterion was the F-score obtained in development, and as a tiebreaker method Area Under the Curve (AUC).

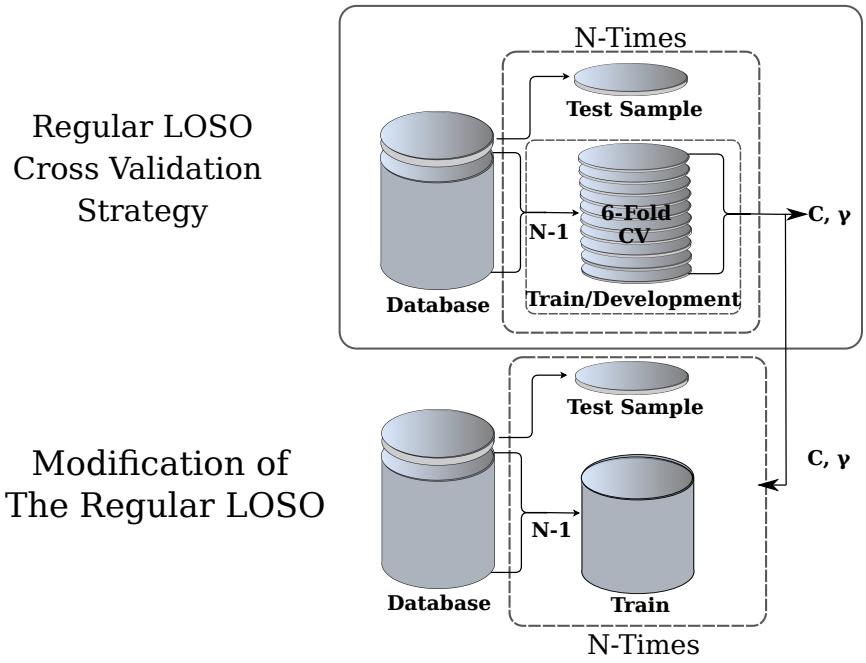

Figure 5.10: Database distribution for the assessment of Alzheimer's disease. **CV:** cross-validation. **N:** number of samples.

5.2.3 Results and Discussion

Linguistic differences between AD patients and HC subjects can be shown via word cloud representations (see Figures 5.11 and 5.12). The texts are pre-processed using noise removal and lexicon normalization, in order to build the word clouds. This representation allows the viewer to see which words are used more or less frequently. The bigger the word in the cloud, the more frequently it is used. Most of the words are similar since all groups performed the *cookie theft* task, where the attention is focused in words such as kid ("niño"), cat ("gato"), mother ("mamá", "señor"), and cookie ("galleta"). Note that the word "mamá" is in male form due to the lemmatization process. The GC and NGC groups mention the mother as "mamá" and "señor",

while MCI patients and HC subjects recognize the mother easier as "señor" (Mrs). The MCI patients more frequently use the Colombian crutch "pues" followed by the HC subjects. This may denote usually a lack of fluency in the speech. Additionally, the age may influence the lexicon, since on average the MCI and HC are 20 years older than GC and NGC subjects.

Figure 5.11: Word cloud representation for the assessment of genetic carries in Alzheimer's disease: A) GC subjects. B) NGC subject.

Figure 5.12: Word cloud representation for the assessment of Alzheimer's disease patients with MCI: A) MCI patients. B) HC subjects.

The experiments consider two classification tasks: (1) MCI vs. HC, and (2) GC vs. NGC. Other classification tasks are not considered since the corpora are not balanced with respect to the age. Kruskal-Wallis test with Bonferroni correction was performed to evaluate whether there

was a significant difference between groups. The null hypothesis of the medians coming from the same distribution was rejected ($p \ll 0.05$) in all cases.

GC vs. NGC

Table 5.7 shows the classification results considering each feature set individually. Prosody and BERT produce the highest results for GC vs. NGC. Despite the fact that BERT obtained a higher F-score than prosody, sensitivity and specificity are more balanced for prosody.

Table 5.7: Results for the assessment of genetic carries in Alzheimer's disease using each feature set separately

Features	Experiment	F-score	UAR	Sens	Spe	AUC	C	γ
Acoustic	Articulation	0.47	46.9	39.3	52.8	0.47	55e-1	55e-5
	Prosody	0.67	67.2	60.7	72.2	0.70	10e1	10e-5
Linguistic	W2V	0.53	53.1	39.3	63.9	0.52	10e0	10e-5
	BERT	0.68	68.8	50.0	83.3	0.74	10e0	10e-5
	BETO	0.65	65.6	53.6	75.0	0.72	55e-1	10e-5

Notes: **UAR**: unweighted average recall. **Sens**: sensitivity. **Spe**: specificity. **AUC**: Area under the ROC curve. **C** and γ are the optimal meta-parameters obtained in development for RBF-SVM. Unweighted average recall, Sensitivity and Specificity are given in [%].

Table 5.8 shows the results using the early fusion strategy. The combination of features does not improve the results of classification using each feature set separately.

Figures 5.13 and 5.14 show the score distributions for each feature set and the combination of W2V and BERT. The dark gray bars correspond to the scores for NGC subjects, the white bars are the scores computed for the GC subjects. Note that for prosody the range of the scores for the NGC subjects is wider, and for the GC most of the subjects are concentrated close to the decision boundary of the RBF-SVM.

The ROC curves obtained are shown in Figure 5.15. Prosody, BERT and BETO obtained the highest performances. BERT and BETO produce similar results, which concludes that for our approach the translation to Spanish did not show a strong impact on the results.

Table 5.8: Results for the assessment of genetic carries in Alzheimer's disease using early fusion of the different feature sets.

Experiment	F-score	UAR	Sens	Spe	AUC	C	γ
Art-Pro	0.52	51.6	53.6	50.0	0.55	10e0	10e-5
Art-W2V	0.49	50.0	32.1	63.9	0.49	10e0	10e-5
Art-BERT	0.60	60.9	42.9	75.0	0.69	10e0	10e-5
Pro-W2V	0.53	54.7	32.1	72.2	0.52	10e-1	10e-4
Pro-BERT	0.63	64.1	46.4	77.8	0.68	10e0	10e-5
W2V-BERT	0.65	67.2	42.9	86.1	0.62	10e-1	10e-5
Art-Pro-W2V	0.53	54.7	32.1	72.2	0.53	10e0	10e-5
Art-Pro-BERT	0.62	62.5	46.4	75.0	0.70	10e0	10e-5
Art-W2V-BERT	0.53	54.7	35.7	69.4	0.64	55e-1	10e-5
Pro-W2V-BERT	0.61	60.9	50.0	69.4	0.65	10e0	10e-5
Art-Pro-W2V-BERT	0.62	64.1	39.3	83.3	0.63	10e-1	10e-5

Notes: **UAR**: unweighted average recall. **Sens**: sensitivity. **Spe**: specificity. **AUC**: Area under the ROC curve. **C** and γ are the optimal meta-parameters obtained in development for RBF-SVM. Unweighted average recall, sensitivity, and specificity are given in [%].

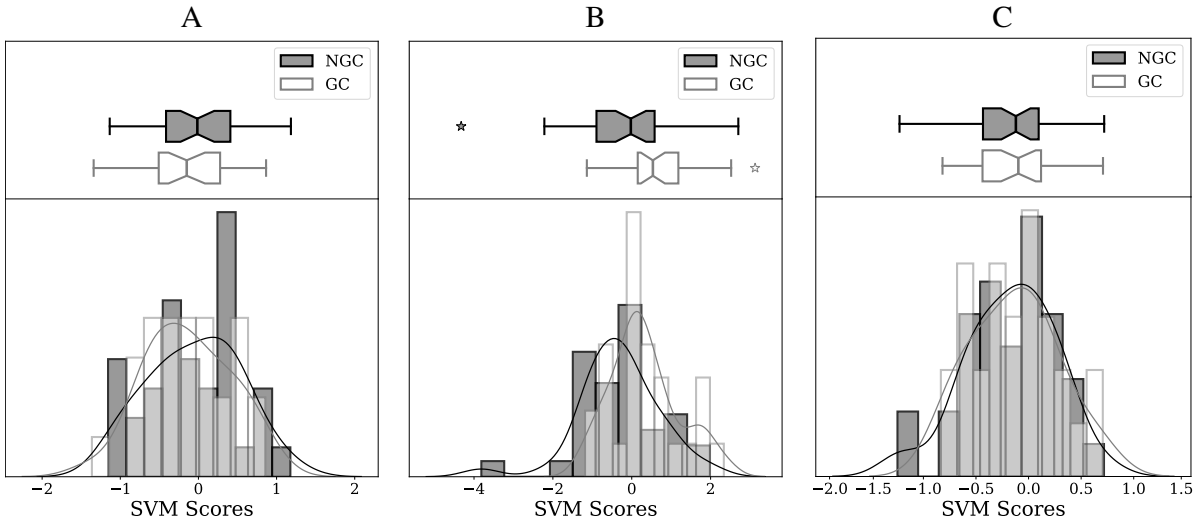

Figure 5.13: Scores for the assessment of genetic carries in Alzheimer's disease obtained for: A) Articulation. B) Prosody. C) W2V.

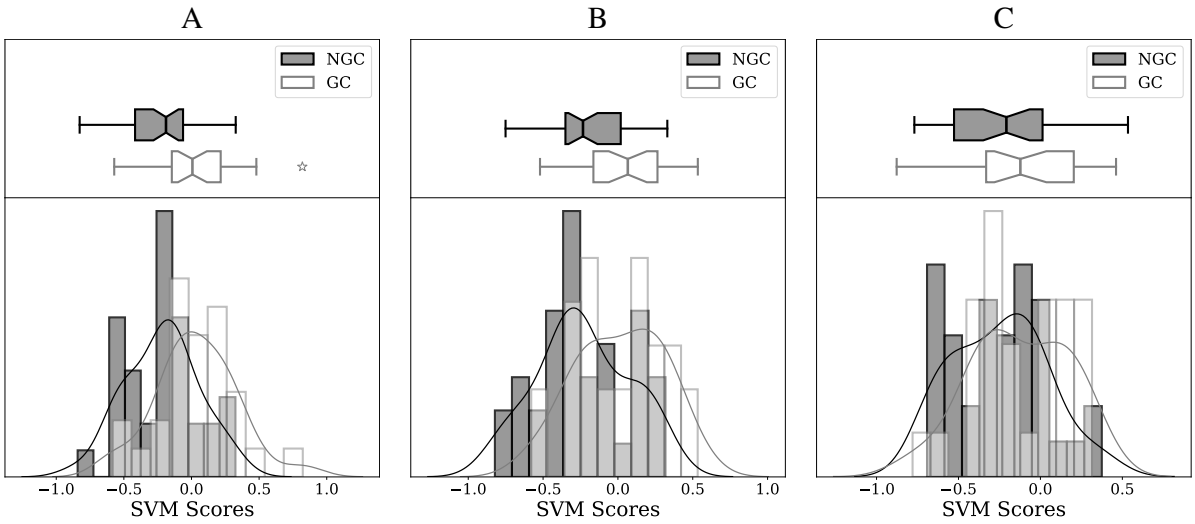

Figure 5.14: Scores for the assessment of genetic carries in Alzheimer's disease obtained for: A) BERT. B) BETO. C) Early fusion between W2V and BERT.

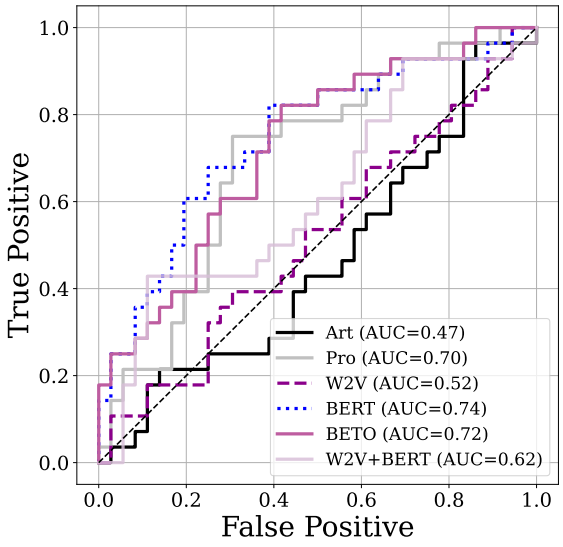

Figure 5.15: ROC Curve for the assessment of genetic carries in Alzheimer's disease obtained for different feature sets. Art: articulation. Pro: prosody.

MCI patients vs. HC subjects

Table 5.9 shows the classification results considering each feature set individually to discriminate AD patients with MCI. The most accurate results for MCI vs. HC are obtained using acoustic features, while linguistics do not produce good performance.

Table 5.9: Results for the assessment of Alzheimer's disease patients with MCI using each feature set separately

Features	Experiment	F-score	UAR	Sens	Spe	AUC	C	γ
Acoustic	Articulation	0.66	66.0	69.6	63.0	0.70	55e-1	55e-5
	Prosody	0.66	66.0	56.5	74.1	0.70	10e1	10e-5
Linguistic	W2V	0.48	48.0	39.1	55.6	0.46	55e0	10e-5
	BERT	0.50	50.0	43.5	55.6	0.45	10e0	10e-5
	BETO	0.50	50.0	39.1	59.3	0.55	55e-1	10e-5

Notes: **UAR**: unweighted average recall. **Sens**: sensitivity. **Spe**: specificity. **AUC**: Area under the ROC curve. **C** and γ are the optimal meta-parameters obtained in development for RBF-SVM. Unweighted average recall, Sensitivity and Specificity are given in [%].

Table 5.10 shows the results using the early fusion strategy. In general, when articulation and prosody features are combined, the performance of the classifier improves in comparison with each feature set separately.

Table 5.10: Results for the assessment of Alzheimer's disease patients with MCI using early fusion of the different feature sets.

Experiment	F-score	UAR	Sens	Spe	AUC	C	γ
Art-Pro	0.74	74.0	65.2	81.5	0.77	10e0	10e-5
Art-W2V	0.58	58.0	52.2	63.0	0.63	10e0	10e-5
Art-BERT	0.54	54.0	47.8	59.3	0.57	10e0	10e-5
Pro-W2V	0.48	48.0	39.1	55.6	0.50	10e0	10e-5
Pro-BERT	0.48	48.0	39.1	55.6	0.50	10e0	10e-5
W2V-BERT	0.50	50.0	47.8	50.0	0.52	10e0	10e-5
Art-Pro-W2V	0.66	66.0	60.9	70.4	0.68	10e0	10e-5
Art-Pro-BERT	0.58	58.0	56.5	59.3	0.59	10e0	10e-5
Art-W2V-BERT	0.50	50.0	543.5	60.0	0.53	55e-1	10e-5
Pro-W2V-BERT	0.43	44.0	30.4	55.6	0.44	10e0	10e-5
Art-Pro-W2V-BERT	0.54	54.0	43.5	63.0	0.55	10e0	10e-5

Notes: **UAR**: unweighted average recall. **Sens**: sensitivity. **Spe**: specificity. **AUC**: Area under the ROC curve. **C** and γ are the optimal meta-parameters obtained in development for RBF-SVM. Unweighted average recall, sensitivity, and specificity are given in [%].

Figures 5.16 and 5.17 show the scores obtained for each feature set, and the combination of prosody and BETO. The dark gray bars correspond to the scores for HC subjects, the white bars are the scores computed for the MCI patients. The combination of articulation and prosody improves the performance and decreases the number of outliers in the distribution in comparison with only using articulation.

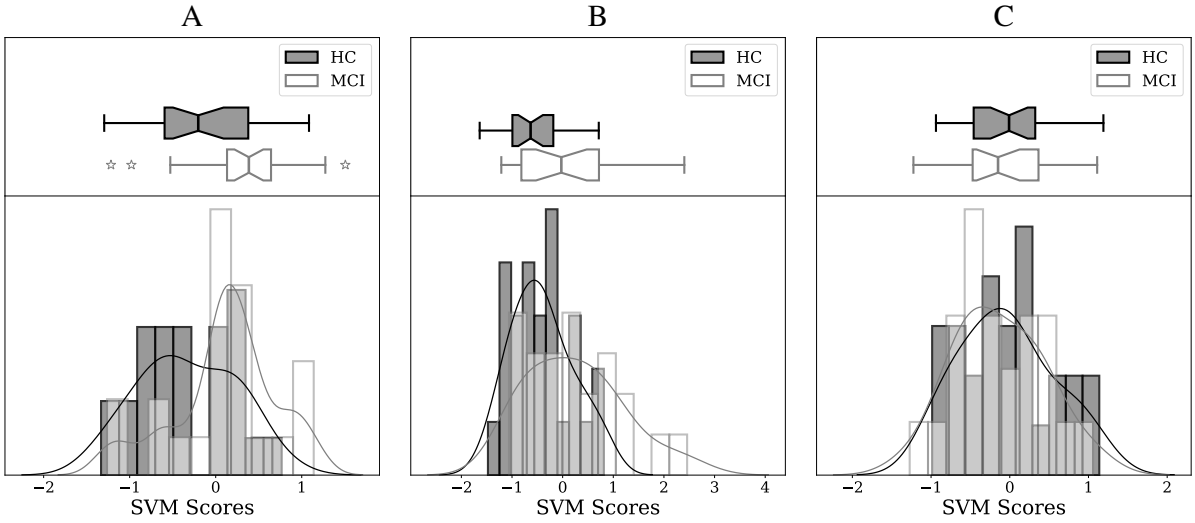

Figure 5.16: Scores for the assessment of Alzheimer's disease patients with MCI obtained for: A) Articulation. B) Prosody. C) W2V.

Figure 5.18 shows the ROC curves. Note that for this classification task the linguistic embeddings do not get satisfactory results. This may occur due to some unknown words by the algorithms related to characteristic lexicon from the region, or mispronunciations of the words. The GC and NGC subjects are at a mean age of close to 30 years, while the patients and HC are at a mean age 50. The MCI patients and HC subjects have a lower education level, and many come from rural regions, thus their lexicon is more characteristic of the region and tends to produce more mispronunciations of the words. Regarding BERT and BETO models, the results with BERT were slightly higher, which concludes that for our approach the translation to Spanish did not show a strong impact on the results.

The comparison with the regular LOSO strategy was performed, where the difference between the F-scores between both validation approaches was 3% on average, which confirms that the regular LOSO is more optimistic. More detail about the comparison between the two validation methods in Appendix A.

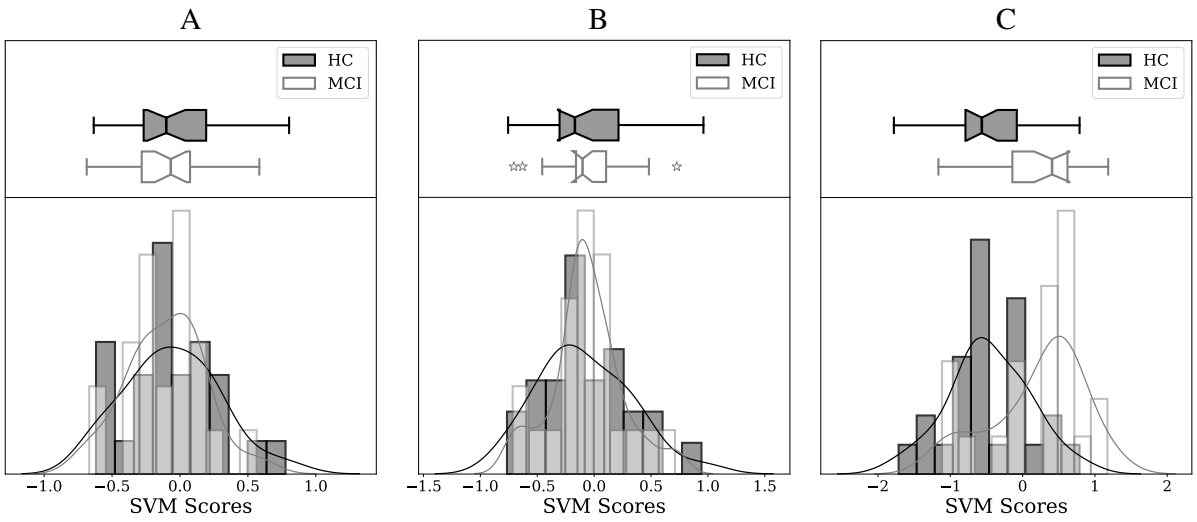

Figure 5.17: Scores for the assessment of Alzheimer's disease patients with MCI obtained for: A) BERT. B) BETO. C) Early fusion between articulation and prosody.

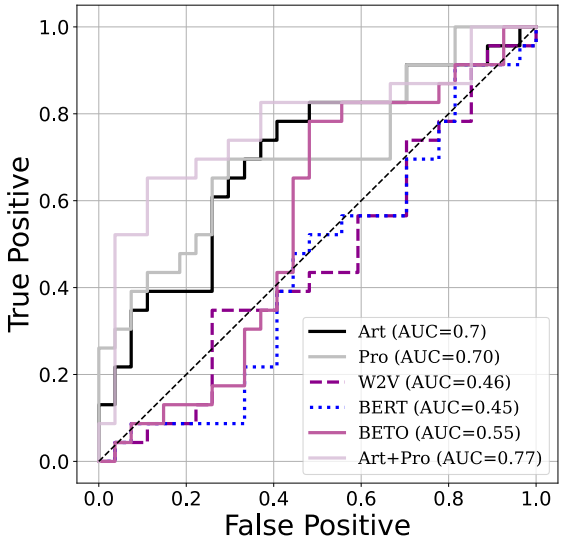

Figure 5.18: ROC Curve for the assessment of Alzheimer's disease patients with MCI obtained for different feature sets. Art: articulation. Pro: prosody.

5.3 Linguistic Analysis to Discriminate Parkinson's Disease

This thesis proposes the use of NLP methods to extract features from transcriptions to discriminate between HC subjects and PD patients. Although acoustic analysis has shown to be a suitable tool to study symptoms of PD patients, there are components related to language production that are not modeled with that approach. The aim is to to model language deficits exhibited by PD patients.

5.3.1 Methodology

The database for this approach was presented in Section 4.4. The transcriptions were obtained from spontaneous speech recordings, where the participants were asked to describe their daily routines. Figure 5.19 shows the main steps followed in this study to perform the text analysis. The data is split into train and test sets, in order to perform a LOSO strategy to validate this approach. Text processing for all NLP techniques is performed to remove noisy entities and to standardize and clean the text. The feature extraction step includes classical approaches such as BoW and TF-IDF, along with word-embeddings like W2V. Stemming process is not considered in the Lexicon normalization step for W2V. The classification is performed to discriminate between PD patients and HC subject.

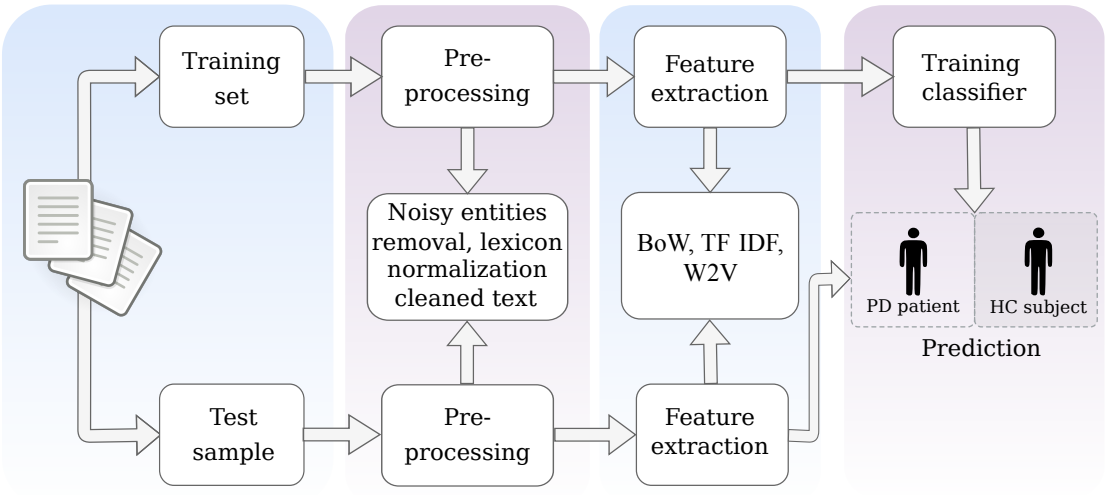

Figure 5.19: Scheme of the general methodology to discriminate Parkinson's disease using NLP

5.3.2 Optimization and Classification

The discrimination capability is performed using an RBF-SVM and Random Forest. The data is distributed following a LOSO strategy as it is shown in Figure 5.20. The criterion for the meta-parameters optimization is performed following a 10-fold cross validation strategy. The optimization is based on the accuracy obtained in the development data. Notice that this process might be slightly optimistic since we have a different set of meta-parameters for each test sample; however, the results show that the distribution of the meta-parameters is stable across the different test sets. The optimal parameters are found through a grid search, where $C \in \{10^{-4}, 10^{-3}, ..., 10^{4}\}$ and $\gamma \in \{10^{-4}, 10^{-3}, ..., 10^{4}\}$ for the RBF-SVM, and number of trees $N \in \{5, 10, 20, 30, 50, \ 100\}$ and maximum depth $D \in \{2, 5, 10, 20, 30, 50, 100\}$ for the RF. The optimal hyper-parameters are found based on the median of the values of the hyper-parameters obtained for each fold.

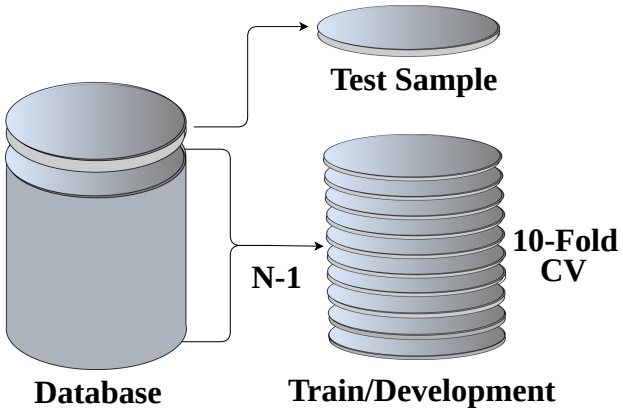

Figure 5.20: Database distribution for the assessment of Parkinson's disease using NLP. **CV:** cross-validation. **N:** number of samples.

5.3.3 Results and Discussion

The Word cloud representations between PD patients and HC subjects are shown in Figure 5.21. The text are preprocessed using noise removal and lexicon normalization, in order to build the word clouds. The patients frequently used words such as "casa" (house) or "ver televisión" (watch television). and fillers in Spanish like "pues", while words such as "trabajar" (work), "salir" (go out), "gustar" (like) or "bueno" (good) appear more frequently in HC subjects. In addition to the words related to the daily activities of the subjects, in PD patient appears more frequently a Colombian fuller word "pues" ($\approx$ well in English), denoting a lack of fluency in the speech of the patients. Note that the patients said the same words most of the time since there are fewer words

that look representative in the cloud in comparison to the cloud of the HC subjects, where there is more variety.

Figure 5.21: Word cloud representation for the assessment of Parkinson's disease using NLP: A) PD patient. B) HC subject

Classification results in Table 5.11 considers the three feature sets individually and their combination in an early fusion strategy. The performance of the classifiers is evaluated according to their Unweighted Average Recall (UAR), Sensitivity (Sens), Specificity (Spe), and the Area Under the receiver operating characteristic Curve (AUC). Highest results are obtained for W2V using RBF-SVM (UAR=72%) and for BoW (UAR=70%) using RF. On the one hand, the results related to BoW could indicate that the occurrence of the words may be relevant in order to discriminate the disease. On the other hand, W2V is a more robust method that allows extracting contextual information from the text. The early fusion strategy did not improve the performance, indicating that the considered features are not complementary and further research is required to find an optimal strategy to merge such information. However, this could be due to the high dimensionality and a large amount of sparse cells provide by BoW and TF-IDF.

Notice that in general specificity is lower than sensitivity, which indicates that PD patients were better discriminated in most of the cases. This difference between specificity and sensitivity suggests that AUC is a better statistic to compare the approaches. The highest AUC is obtained with the BoW features using RF (0.76).

The approaches that provide highest AUC values are obtained using RF, where the distribution of the scores is shown Figures 5.22. The dark gray bars correspond to the scores for HC subjects, the white bars are the scores computed for the PD patients, and the light gray bars correspond to the intersection between both sets, and reflect the classification errors. Note that the scores of the PD patients are less sparse than those obtained for the HC subjects, and the scores for the

Table 5.11: Classification results for the assessment of Parkinson's disease using NLP.

Features	RBF-SVM						RF					
	UAR	**Sens**	**Spe**	**AUC**	**C**	γ	**UAR**	**Sens**	**Spe**	**AUC**	**N**	**D**
BoW	62.0	70.0	54.0	0.60	10e+1	10e0	70.0	74.0	66.0	0.76	100	20
TF-IDF	58.0	58.0	56.0	0.60	10e+1	10e0	67.0	68.0	66.0	0.71	100	100
W2V	72.0	92.0	52.0	0.66	10e0	10e0	67.0	74.0	60.0	0.71	5	5
Fusion	60.0	62.0	58.0	0.62	10e+1	10e0	66.0	68.0	64.0	0.71	100	5

Notes: **UAR**: unweighted average recall. **Sens**: sensitivity. **Spe**: specificity. **AUC**: Area under the ROC curve. **C** and γ are the optimal meta-parameters obtained in development for RBF-SVM. **N** and **D** are the optimal meta-parameters obtained in development for RF. Unweighted average recall, Sensitivity and Specificity are given in [%].

TF-IDF features are more overlapped. W2V and fusion RF scores are more overlapped and sparse, especially for HC subjects, where a higher variance is observed, therefore this may provide a less stable solution.

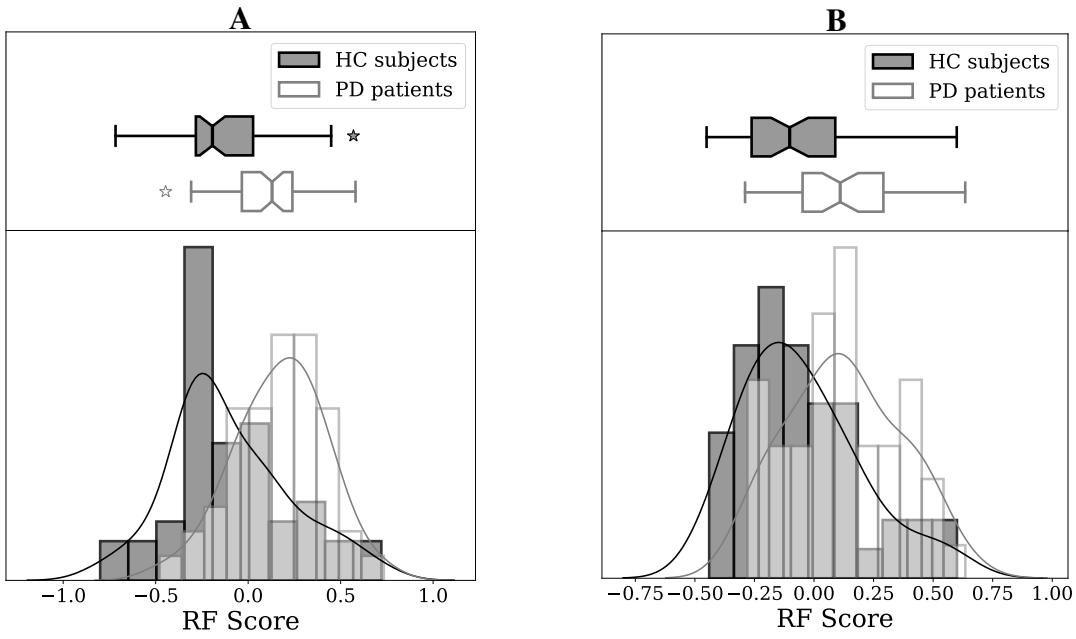

Figure 5.22: Scores for the assessment of Parkinson's disease using NLP obtained for the RF classifier for: A) BoW. B) TF-IDF.

The ROC curves obtained using the RF classifier are shown in Figure 5.24. This figure allows to show the results more compactly. ROC curve for BoW is the most stable, however, the AUC and the curves are close to each other.

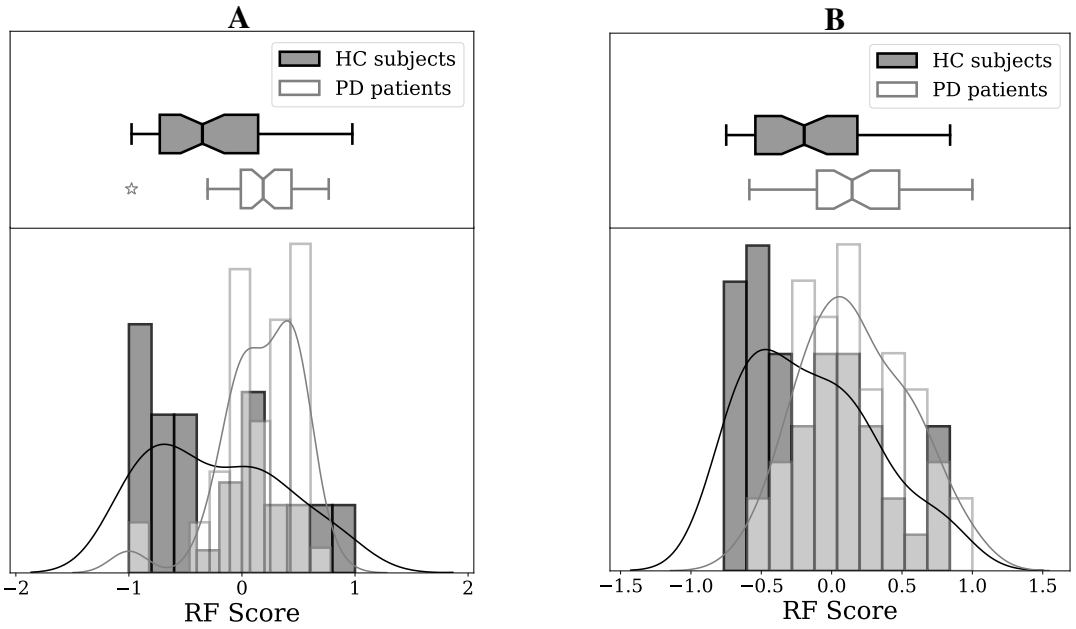

Figure 5.23: Scores for the assessment of Parkinson's disease using NLP obtained for the RF classifier for: A) W2V. B) Fusion.

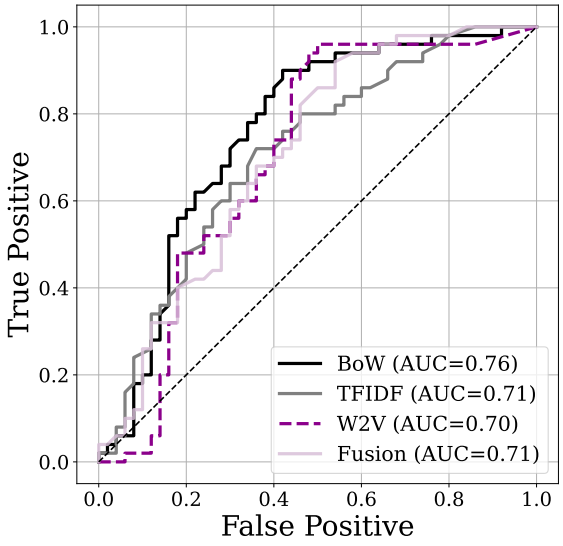

Figure 5.24: ROC Curve for the assessment of Parkinson's disease using NLP obtained for the RF classifier

5.4 Depression in Parkinson's Disease

Depression is one of the typical non-motor symptoms that most Parkinson's patients develop. Impairments in speech production together with depression produce negative effect in the communication capabilities and social interaction of patients. This thesis considers a combination of acoustic and linguistic methods to model depressive patterns in PD. The used datasets in this experiment are presented in Section 4.5 and Section 4.4. To the best of my knowledge, this is the first study focused on evaluating depression symptoms in PD patients combining acoustic analysis and NLP.

5.4.1 Methodology

Articulation and prosody dynamic features are considered for speech, while word embedding are extracted for linguistic analysis. All features are shown in Table 5.12, where for speech signals is taken chunks of 40 ms a shift of 10ms, only considering onset segments, and for linguistic the embedding for each word in the utterance.

Table 5.12: Dynamic features considered in this approach for the assessment of depression in Parkinson's disease.

Features type	Descriptors	Dimension
Articulation	The energy content in the transitions is modeled considering by the Bark scale, and 12 MFCCs along with their derivatives	Number of voice segments $\times$ 58
Prosody	Coefficients of 5-degree Legendre polynomials to model the pitch and the energy contour, separately, and the duration of each voiced segments	Number of voiced segments $\times$ 13
Word embeddings	BERT embeddings from the last layer	Number of words $\times$ 768

GMM-UBM to model acoustic and linguistic features is considered. Figure 5.25 shows the scheme of the GMM-UBM based approach implemented in this thesis. These models are well known for their effectiveness and scalability to model the spectral distribution of speech, especially for text-independent speaker recognition applications. The inputs of the GMM-UBM are the extracted features after applying PCA with 80% of cumulative variance.

GMMs represent the distribution of the feature for each ND-PD and D-PD patient. This PD patients are considered from depression in PD database presented in Section 4.5. The GMM model is derived from the UBM by adapting the parameters (mean vectors, covariance matrix,

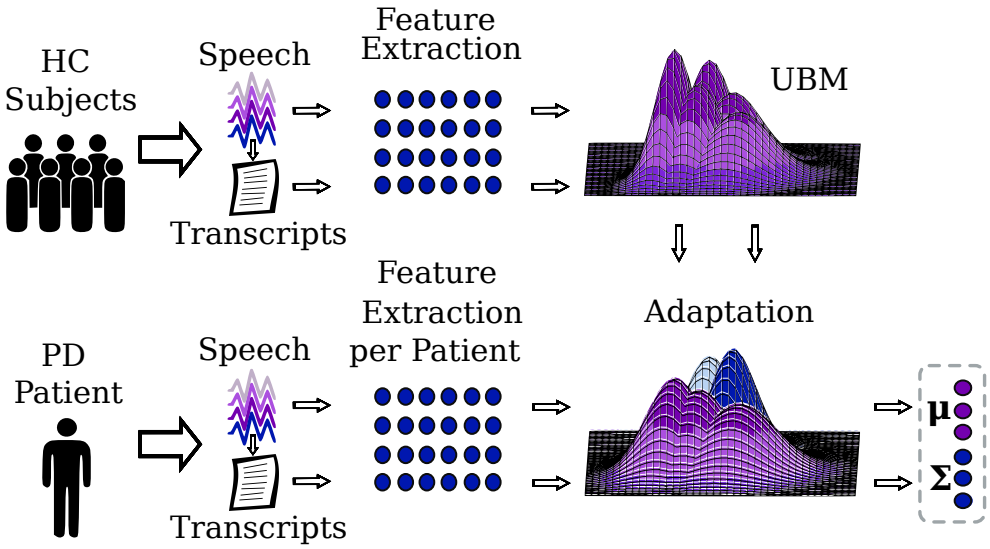

Figure 5.25: GMM-UBM based approach addressed in this thesis for the assessment of depression in Parkinson's disease

and mixture weights). The UBM is trained with two Gaussian components with the data from HC subjects of PC-GITA database presented in Section 4.4. The number of Gaussian components was selected based on the performance, and on the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). A user model for each patient is adapted to build a supervector by concatenating the mean vector and the diagonal of the covariance matrix, to form a static vector that represent each patient. Finally, each supervector from the GMM-UBM model is used as input features to discriminate between D-PD and ND-PD as is shown in Figure 5.26.

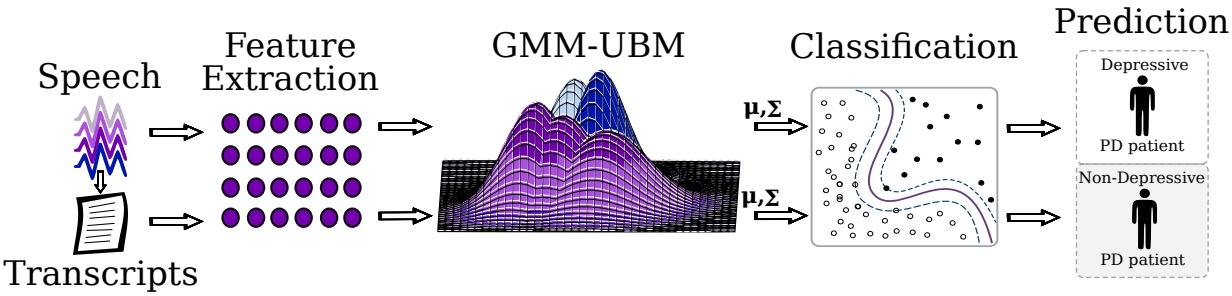

Figure 5.26: Scheme of the methodology addressed in this thesis for the assessment of depression in Parkinson's disease

5.4.2 Optimization and Classification

The extracted supervector is used as input in the classification. Similar to Section 5.2, the classification is performed using an RBF-SVM and following a LOSO strategy. The criterion for the meta-parameters optimization is performed following a 6-fold cross validation strategy. The optimal parameters $C \in \{10^{-4}, 10^{-3}, ..., 10^{4}\}$ and $\gamma \in \{10^{-4}, 10^{-3}, ..., 10^{4}\}$ of the RBF-SVM are found through a grid search. The classification is performed for each feature set separately and for the combination using an early fusion strategy to merge linguistic and acoustic features before performing the classification. Late fusion strategies were also considered, but the results were not satisfactory. The baseline considers four functionals (mean, standard deviation, kurtosis and skewness) to form a static vectors for each speaker. These functionals are computed over the described acoustic and linguistic features sets.

5.4.3 Results and Discussion

Linguistic differences between D-PD and ND-PD patients can be shown via word cloud representation (see Figure 5.27). The text are preprocessed using noise removal and lexicon normalization, in order to build the word clouds.There are sligthly differences between the word clouds of D-PD and ND-DP patients. Note that D-PD patients frequently used words that are directly linked to routine activities such as "breakfast" (desayunar), "lunch" (almorzar), and "dinner" (comer or cenar). Words as "salir" (go out), "bueno" (good) and "work" (trabajar) appear more frequently in ND-DP patients.

BERT embeddings are using to capture the contextual and linguistic information since word counting methods are mostly used for inference and may not provide enough information regarding this problem.

Four different feature sets are considered for acoustic and linguistic analyses: (1) baseline, (2) baseline after applying PCA, (3) the GMM-UBM supervectors, and (4) the GMM-UBM supervectors after applying PCA over the feature space. The results for each feature set separately are shown in Table 5.13. Highest results are obtained using articulation and BERT features. The GMM-UBM based approaches are the most accurate to model depression disturbances in PD, where articulation obtained an F-score=0.72, and with PCA prosody and BERT an F-score=0.70.

Table 5.14 shows the results for the combination of the different features sets using the early fusion strategy. GMM-UBM model after applying PCA shows the highest result while combining articulation and BERT features (F-score=0.77). This result is higher than the one obtained with the baselines, and with each individual feature set. The combinations of features that included

Figure 5.27: Word cloud representation for the assessment of depression in Parkinson's disease: A) D-PD patient. B) ND-PD patient

Table 5.13: Results for the assessment of depression in Parkinson's disease using each feature set separately

Experiment	Features	F-score	Sens	Spe	AUC	Number of features	Space Reduction
Baseline	Articulation	0.60	72.0	51.4	0.61	488	-
	Prosody	0.53	40.0	62.9	0.51	52	-
	BERT	0.65	60.0	68.6	0.70	3072	-
Baseline and PCA	Articulation	0.63	76.0	54.3	0.70	16	3.3
	Prosody	0.55	40.0	68.6	0.54	15	28.9
	BERT	0.54	52.0	54.3	0.55	37	1.2
GMM-UBM	Articulation	0.72	76.0	68.6	0.75	232	-
	Prosody	0.49	20.0	77.1	0.47	52	-
	BERT	0.64	48.0	77.1	0.63	3072	-
PCA and GMM-UBM	Articulation	0.65	72.0	60.0	0.72	56	24.1
	Prosody	0.70	72.0	68.6	0.69	16	30.8
	BERT	0.70	56.0	80.0	0.70	508	16.8

Notes: **AUC**: area under the curve. **Sens**:Sensitivity. **Spe**: Specificity. Space Reduction, Sensitivity and Specificity are given in [%]. Space reduction: the proportion of the modified features sets after applying PCA.

prosody do not show satisfactory results, thus future work will explore other features related to prosody to increase the robustness of the model.

The approaches that provide highest AUC values are obtained using PCA and GMM-UBM for articulation, BERT and its fusion, where the distribution of the scores is shown in Figures 5.28.A, 5.28.B, and 5.28.C, respectively. The dark gray bars correspond to the scores for ND-PD patients, the white bars are the scores computed for the D-PD patients, and the light gray bars correspond to

Table 5.14: Results for the assessment of depression in Parkinson's disease using early fusion of the different feature sets.

Experiment	Features	F-score	Sens	Spe	AUC	Number of features	Space Reduction
Baseline	Art-Pro	0.60	68.0	54.3	0.59	540	-
	Art-BERT	0.62	52.0	68.6	0.71	3560	-
	Pro-BERT	0.65	60.0	68.6	0.70	3124	-
	Art-Pro-BERT	0.62	56.0	65.7	0.69	3612	-
Baseline and PCA	Art-Pro	0.60	56.0	62.9	0.54	31	5.7
	Art-BERT	0.60	56.0	62.9	0.65	53	1.5
	Pro-BERT	0.53	40.0	62.9	0.49	52	1.7
	Art-Pro-BERT	0.55	52.0	57.1	0.56	68	1.9
GMM-UBM	Art-Pro	0.56	48.0	62.9	0.66	284	-
	Art-BERT	0.63	48.0	74.3	0.66	3304	-
	Pro-BERT	0.64	44.0	80.0	0.63	3124	-
	Art-Pro-BERT	0.61	44.0	74.3	0.66	3356	-
PCA and GMM-UBM	Art-Pro	0.67	60.0	71.4	0.70	72	24.4
	Art-BERT	0.77	76.0	77.1	0.78	564	17.1
	Pro-BERT	0.66	56.0	74.3	0.72	524	16.8
	Art-Pro-BERT	0.72	76.0	68.6	0.77	580	17.3

Notes: **AUC**: area under the curve. **Sens**:Sensitivity. **Spe**: Specificity. Art: Articulation. Pro: Prosody. Space Reduction, Sensitivity and Specificity are given in [%]. Space reduction: the proportion of the modified features sets after applying PCA.

the intersection between both sets, and reflect misclassifications. Note that the scores of the D-PD patients are less sparse than those obtained for the N-DP patients for articulation, contrary in BERT. The combination of articulation and BERT shows an improvement regarding the consistency in the discrimination of the reference and control class, therefore this may provide a more stable and complementary solution.

The ROC curves obtained using the GMM-UBM model, after applying PCA are shown in Figure 5.29. The highest AUCs are obtained with the early fusion strategy, for all cases. The approach seems to be promising, although further research and more data are required to improve these results.

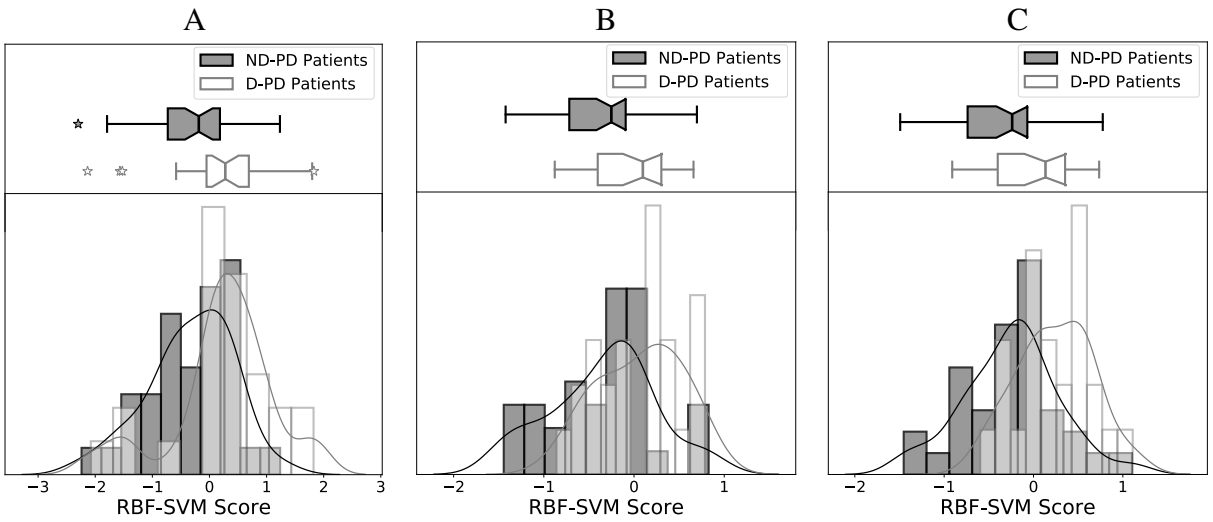

Figure 5.28: Scores for the assessment of depression in Parkinson's disease obtained with RBF-SVM classifier for PCA and GMM-UBM:
A) Articulation. B) BERT. C) Art-BERT

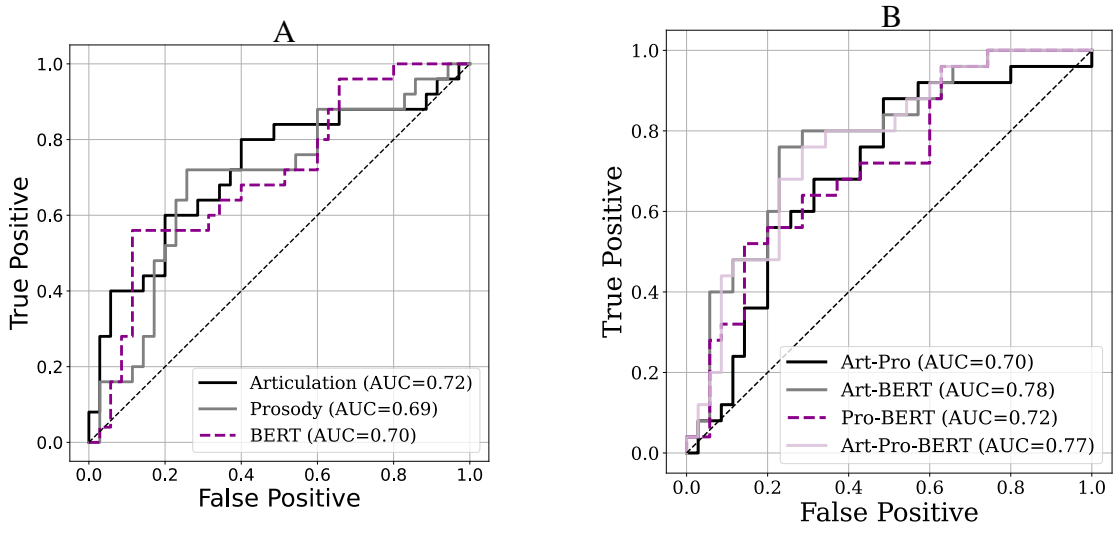

Figure 5.29: ROC curve graphics for the assessment of depression in Parkinson's disease for PCA and GMM-UBM: A) Each feature set separately. B) Early fusion strategy.

5.5 User State Modeling Based on the Arousal-Valence Plane for Customer Satisfaction and Health-Care

This thesis proposes a methodology focused on modeling the user's state using speech and natural language from spontaneous audio recordings and their transcripts for the evaluation of scenarios such as customer satisfaction and assessment of patients with neuro-degenerative diseases. Speech signals and their transcripts are used to train these systems based on the arousal-valence plane representation [6], which performs a quantitative analysis of emotions and can handle the fact of continuously emotional and mood changes that can be present in the human behavior. The trained systems are used in three scenarios related to customer satisfaction and health-care: (1) evaluation of customer satisfaction in call-centers, (2) assessment of depressive symptoms in PD patients, and (3) assessment and classification of AD.

Different models were trained to discriminate each quadrant in the arousal-valence plane and to obtain a set of posterior probabilities and embeddings to be used as input features. The challenge is to model the user state based on this information to be applied in different scenarios, such as customer satisfaction and health-care.

5.5.1 Methodology

The thesis considers to train different models based in several classification tasks to discriminate different quadrants in the arousal-valence plane:

- Bi-class classification of active vs. passive arousal (AA vs. PA): the Active Arousal (AA) class is the combination between the quadrants AP and AN, and the Passive Arousal (PA) class considers PP and PN quadrants as is shown in Figure 5.30.A.

- Bi-class classification of positive vs. negative arousal (PV vs. NV): in this case the Positive Valence (PV) class combines AP and PP quadrants, and Negative Valence (NV) class combines AN and PN quadrants as is shown in Figure 5.30.B.

- Multi-class classification of active positive vs. active negative vs. passive negative vs. passive positive (AP vs. AN vs. PN vs. PP): the last model is trained to classify the four different quadrants as is shown in Figure 5.30.C.

The aforementioned processes are performed in order to decide which is more effective for this problem, bi-classification or multi-class classification. The proposed models are implemented

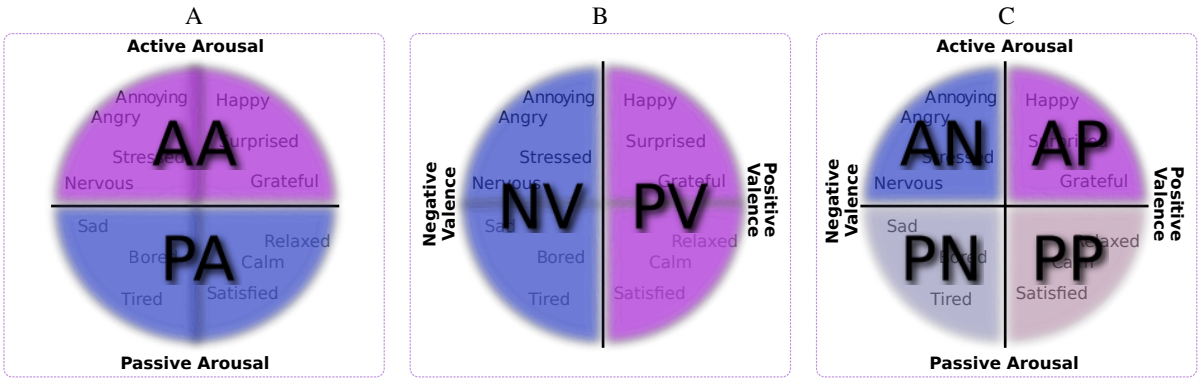

Figure 5.30: Classification tasks according to the arousal-valence plane: A) AA vs. PA, B) PV vs. NV, C) AP vs. AN vs. PN vs. PP

in both acoustic and linguistic analysis. Each analysis considers different architectures. From the speech signals, a 2D-Mel spectrogram is obtained to be used as input in the acoustic model, and the BERT embeddings are extracted from the transcripts, which are the input of the linguistic model. At the end, for each architecture is obtained a set of embeddings and posterior probabilities to be used as input features for different applications.

Acoustic Model

Figure 5.31 shows the proposed architecture for the acoustic analysis, which consists of three main parts: (1) a CNN with two layers, (2) a Bidirectional Gated Recurrent Unit (Bi-GRU) with two stacked layers, and (3) a dense layer at the end. The CNN and the Bi-GRU perform the operations in parallel, and considering as their input the 2D-Mel spectrogram. A summary of the proposed architectures for the CNN, GRU and dense layers are shown in Table 5.15.

The CNN considers two layers of 8 and 4 channels, both with a kernel size of (1,3). In each layer is performed a batch normalization, a max pooling of (1,2), a dropout, and as the activation an Exponential Linear Unit (ELU). The output of the CNN is finally flattened to obtain the vector C. Notice that the operations are only performed in the frequency dimension in order to keep the temporal information as much as possible. The Bi-GRU consists of 2 stacked layers of 128 hidden states, with batch normalization in the last layer. The stacked GRU, with the second GRU taking in outputs of the first GRU and computing the final results. The final hidden states H are then concatenated with the output of the CNN (C). The next step is the dense layer, that considers 2 linear layers of 1024, and 512 neurons. In the linear layers is applied a batch normalization, a dropout and as activation function a Gaussian Error Linear Unit (GELU). The final step is

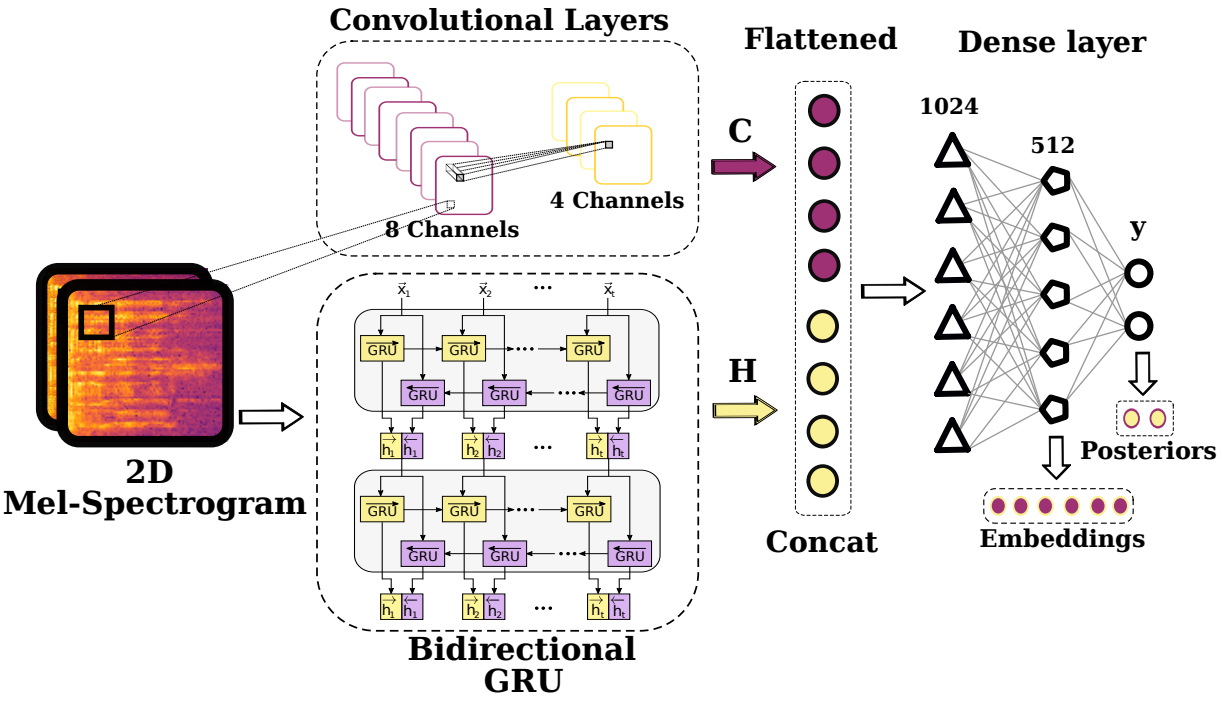

Figure 5.31: Acoustic architecture addressed in this study for modeling the user state based on the arousal-valence plane

Table 5.15: Dimensions of the proposed architecture for the acoustic model for modeling the user state based on the arousal-valence plane

Convolution			**Bi-GRU**		
Layers	**Input**	**Output**	**Layers**	**Input**	**Output**
Convolutional 8-ch k(1,3)	(2, 50, 64)	(8, 50, 62)	Stacked Bi-GRU 128h	(50, 128)	(50, 256)
Max pooling k(1,2)	(8, 50, 62)	(8, 50, 31)	Flatten	(50, 256)	12800
Convolutional 4-ch k(1,3)	(8, 50, 31)	(4, 50, 29)			
Max pooling k(1,2)	(4, 50, 29)	(4, 50, 14)			
Flatten	(4, 50, 14)	2800			
Dense Layer					
Layers		**Input**		**Output**	
Concatenate Covolutional+Bi-GRU		2800+12800		15600	
Linear		15600		1024	
Linear		1024		512	
Linear		512		2 or 4	

Notes: **ch**: channels. **k**: two dimensional kernel. **h**: number of hidden states. The input and output for convolutional and max pooling are defined as (ch, sequence length, frequency axis). The input in the Bi-GRU layer is defined as (sequence length, ch$\times$frequency axis) and the output as (sequence length, number of directions$\times$hidden states)

the classification layer, which is defined by the number of neurons y, 2 o 4 depending on the classification task, and then is followed by a Sigmoid or a Softmax activation function for bi-class or multi-class, respectively.

Multi-channel spectrograms and CNNs have been shown good performance in emotion recognition tasks. This work takes into account the CNN to emphasize on capturing the energy inside the different frequency bands of the spectrogram [96], [97]. The recurrent neural network models has been previously adopted for prosody-prediction in different speech systems [98], [99]. In our case, the Bi-GRUs are used in order to capture prosodic information given 500 ms sequences.

Linguistic Model

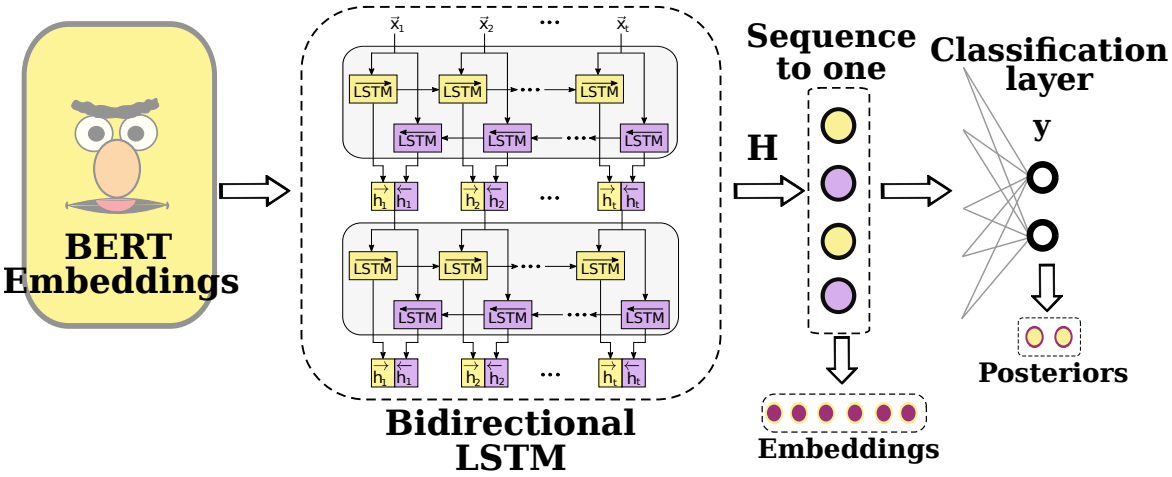

Figure 5.32: Linguistic architecture addressed in this study for modeling the user state based on the arousal-valence plane

Figure 5.32 shows the linguistic architecture that consists of 2 main steps: (1) a Bidirectional LSTM (Bi-LSTM), and (2) a Dense layer. The Bi-LSTM processes a sequence of 7 BERT word embeddings with an overlap of 1 word embedding. It consists of two bidirectional layers with 128 hidden states and a batch normalization layer. The hidden states consider the sequence to one as the final embedding. The proposed architecture is summarized in Table 5.16.

The classification layer is the same as in the acoustic scheme. Despite the fact that it is well known that BERT is capable of capturing contextual information, in this work a Bi-LSTM is considered in order to focus on a smaller sequences of 7 words based on the learned information.

Table 5.16: Dimensions of the proposed architecture for the linguistic model

Layers	Input	Output
Bi-LSTM 128h	(7, 768)	(2, 128)
Flatten	(2, 128)	256
Linear	256	2 or 4

Notes: **h**: number of hidden states. The input of the Bi-LSTM is defines as (sequence length, word embedding length) and the output as (number of directions, number of hidden states).

Application of the Model

The methodology addressed to apply the obtained models is shown in Figure 5.33. The pre-trained models were derived from training, separately acoustic and linguistic using the IEMOCAP dataset to discriminate different quadrants in the arousal-valence plane. In training, the learning rate was set at 10^{-4} for all models, and the ADAM optimizer and cross-entropy as loss function was used. There are three different pre-trained models given different classification tasks as was mentioned before: (1) AA vs. PA, (2) PV vs. NV, and (3) AP vs. AN vs. PN vs. PP.

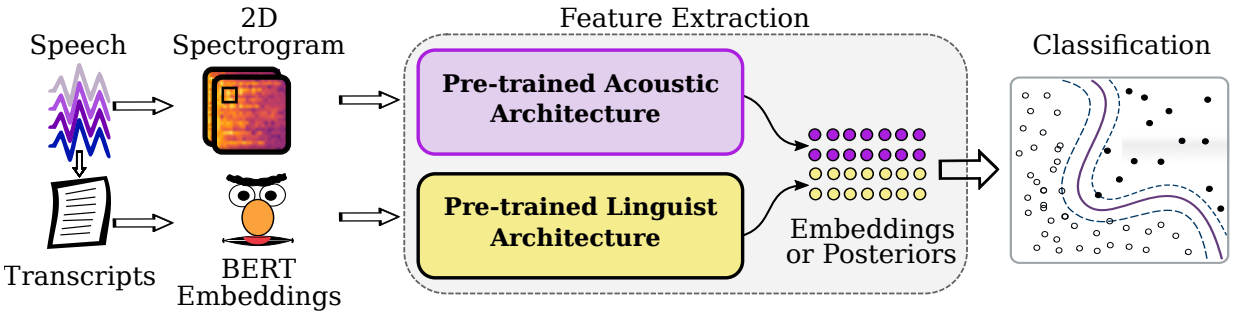

Figure 5.33: Methodology addressed in this study for modeling the user state based on the arousal-valence plane

The aim to train two different classification models is to evaluate the suitability of two different bi-class models vs. one multi-class model. After pre-training the models, it is assumed that information can be extracted from this plane to model emotional and mood changes. The remaining three datasets are used as test to validate this assumption. As the labels for the databases are not the same, the models are used to extract embeddings from the last layer and also the log-likelihood posterior probabilities as two different feature sets. Then, the classification for the remaining datasets are performed using an RBF-SVM.

5.5.2 Optimization and Classification

The pre-trained model was obtained using the sessions 1-4 from the IEMOCAP database and validate with session 5. The extracted posterior probabilities and embeddings for the three remaining dataset were classified with an RBF-SVM. The validation process used for the ADReSS challenge considers the pre-defined train and test set as it was mentioned in Section 4.3. However, a bootstraping strategy of 80% for training and 20% for development is addressed to optimize the meta-parameters of the classifier. The customer satisfaction dataset follows a bootstraping strategy of 70%-15%-15% as is shown in Figure 5.2.A. For the depression in PD dataset the validation process followed a modification of a nested leave one out cross-validation strategy, with an internal 6-fold cross-validation to optimize the hyper-paramaters of the SVM. This validation process is illustrated in Figure 5.10. It is the same as the one used in Section 5.4, since the same dataset is considered. The optimal parameters of the SVM in the validation processes were found by using a grid search where $C \in \{10^{-4}, 10^{-3}, ..., 10^{4}\}$ and $\gamma \in \{10^{-4}, 10^{-3}, ..., 10^{4}\}$. The optimal hyper-parameters were found based on the median of the values of the hyper-parameters obtained for each fold. The classification is performed for embeddings and posterior probabilities separately, and for the combination of different modals using an early fusion strategy. Early fusion consisted of merging linguistic and acoustic features before performing the classification and making the final decision. This is performed to analyze the suitability of the different features sets.

Different baseline models are considered in order to validate the suitability of this approach. For depression in PD and for call-centers dataset, the "IS16_ComParE-[100] feature set from openSMILE is considered in acoustics. For the Alzheimer's classification task the baseline was taken from the ADReSS challenge. The baseline for linguistic for the mentioned three cases was performed using the regular BERT embeddings. Four functionals (mean, standard deviation, kurtosis and skewness) were computed over the BERT embeddings to form a static vector for each speaker.

5.5.3 Results and Discussion

Four different experiments are performed related to this approach: (1) classification of the arousal-valence plane, (2) classification of customer satisfaction, (3) classification of depression in Parkinson, and (4) classification of Alzheimer's disease, The performance of the classifiers is evaluated according to their weighted F-score, Unweighted Average Recall (UAR), weighted Precision (Prec), weighted Recall (Rec), and Area Under the ROC curve (AUC).

Classification of the Arousal-Valence Plane

This experiment aims to test the proposed models. The classification is performed using the sessions 1-4 to train and to validate the model, and session 5 is used for testing. The classification performance was higher for the 2D-Mel spectrogram using the real and the imaginary part together, which improve respect to each part separately 3%. Table 5.17 shows the results of the acoustic and linguistic models to discriminate: (1) Arousal: AA vs. PA, (2) Valence: PV vs. NV, and (3) Quadrants: AP vs. AN vs. PN vs. PP. The linguistic model produces higher performance than the acoustic. The classification of the valence obtained the most accurate results for both models. Arousal proved to be a difficult task to classify considering that is a two-class problem and the obtained results.

Table 5.17: Test results for the model using session 5 from IEMOCAP dataset

Model		**UAR**	**F-score**	**Prec**	**Rec**
Acoustic Model	Arousal	69.1	0.67	0.68	0.69
	Valence	79.7	0.83	0.86	0.79
	Quadrants	59.7	0.58	0.61	0.60
Linguistic Model	Arousal	73.8	0.76	0.77	0.74
	Valence	80.8	0.84	0.89	0.81
	Quadrants	60.2	0.65	0.72	0.60

Notes: **UAR**: unweighted average recall. **Prec**: precision. **Rec**: recall. Unweighted average recall is given in [%].

Figures 5.34, and 5.35 show two Mel spectrogram examples for active and passive arousal . The audio recording from a female speaker consisted on the phrase in Spanish “Vamos a conversar a la sala” (in English “Let’s talk in the dining room”). The absolute subtraction between the real and imaginary parts are considered in order to visualize the differences between each spectrogram. Note that the active arousal presents slight differences in the high and low frequencies, while the passive arousal exhibits more differences in the low frequencies.

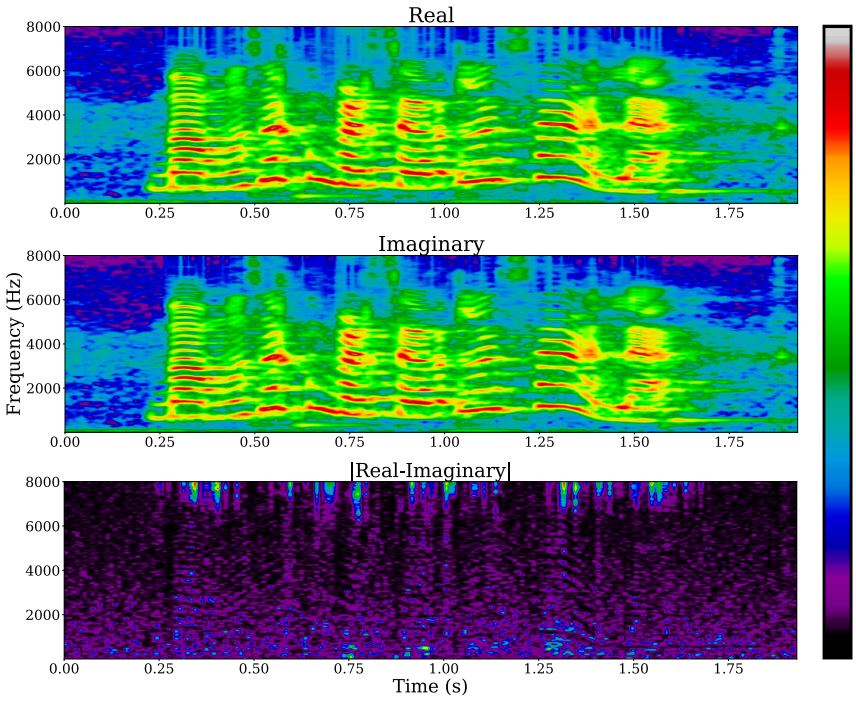

Figure 5.34: Mel spectrogram of the active arousal in the arousal-valence plane.

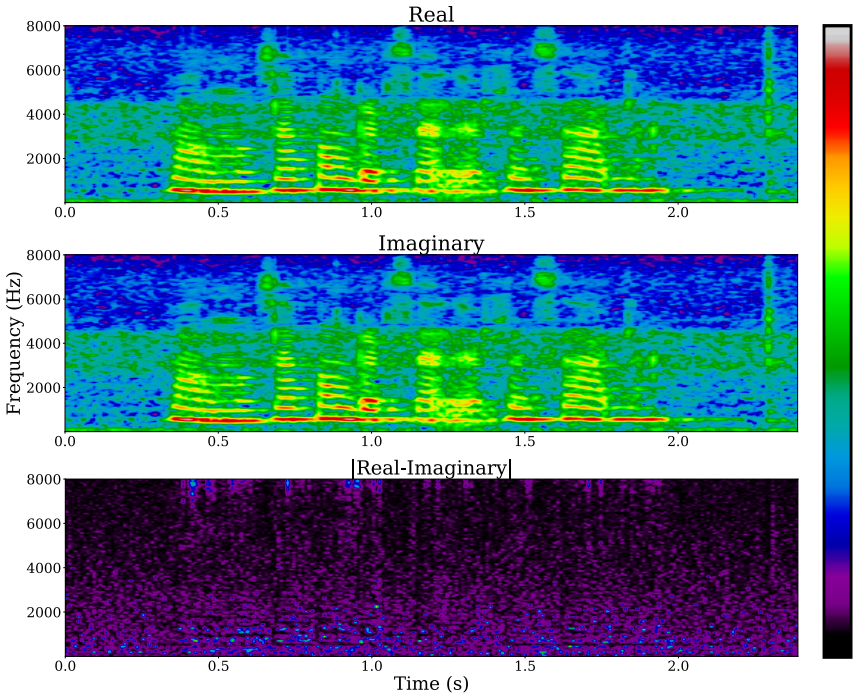

Figure 5.35: Mel spectrogram of the passive arousal in the arousal-valence plane.

Classification of Customer Satisfaction

The banking call-center dataset in Section 4.1 is considered for this experiment. The input features consist of the embeddings and the log-likelihood posterior probabilities from the linguistic and acoustic models. The classification is performed using an RBF-SVM, following the same validation process described in Section 5.1. Table 5.18 shows the results of the baseline models for acoustics and linguistics to compare the suitability of the proposed approach. Acoustics consider as baseline openSMILE, which obtained an F-score of 0.73. BERT is used as baseline for linguistics with an F-score of 0.51.

Table 5.18: Baseline results for classification of customer satisfaction in the banking call-center dataset

Features	Baseline	UAR	F-score	Prec	Rec
Acoustic	openSMILE	72.7	0.73	0.72	0.74
Linguistic	BERT	51.5	0.51	0.51	0.52

Notes: **UAR**: unweighted average recall. **Prec**: precision. **Rec**: recall. Unweighted average recall is given in [%].

Table 5.19: Results for classification of customer satisfaction in the banking call-center dataset using the log-likelihood posterior probabilities

Features	Experiment	UAR	F-score	Prec	Rec	AUC
Acoustic Model	Arousal	65.9	0.66	0.66	0.66	0.70
	Valence	64.8	0.65	0.65	0.65	0.69
	Arousal+Valence	64.5	0.64	0.64	0.65	0.72
	Quadrants	69.0	0.69	0.69	0.69	0.77
Linguistic Model	Arousal	75.8	0.76	0.76	0.76	0.81
	Valence	74.9	0.75	0.75	0.75	0.82
	Arousal+Valence	76.6	0.77	0.77	0.77	0.84
	Quadrants	78.9	0.79	0.79	0.79	0.87
Early Fusion Acoustic + Linguistic	Arousal	72.7	0.73	0.73	0.73	0.77
	Valence	63.1	0.63	0.63	0.63	0.66
	Arousal+Valence	69.9	0.70	0.70	0.70	0.72
	Quadrants	78.6	0.79	0.79	0.79	0.85

Notes: **UAR**: unweighted average recall. **Prec**: precision. **Rec**: recall. Unweighted average recall is given in [%].

Table 5.19 shows the classification performance using the log-likelihood posterior probabilities as input. Arousal+Valence are defined as the early fusion of those posteriors. Highest results are obtained using the posteriors from the quadrant's model. Linguistics obtained the most accurate

results (F-score=0.79), while the early fusion strategy does no improve the performance of each feature set separately.

Figure 5.36 shows the confusion matrices for the best classification results using acoustics, linguistic, and early fusion respectively. The upper left boxes represent the proportion of true negatives (specificity), the upper right the proportion of false negatives, the lower left the proportion of false positives, and the lower right the proportion of true positives (sensitivity). The darker the box, the higher the proportion. Note that the sensitivity and specificity are more balanced for linguistics (see Figure 5.36.B).

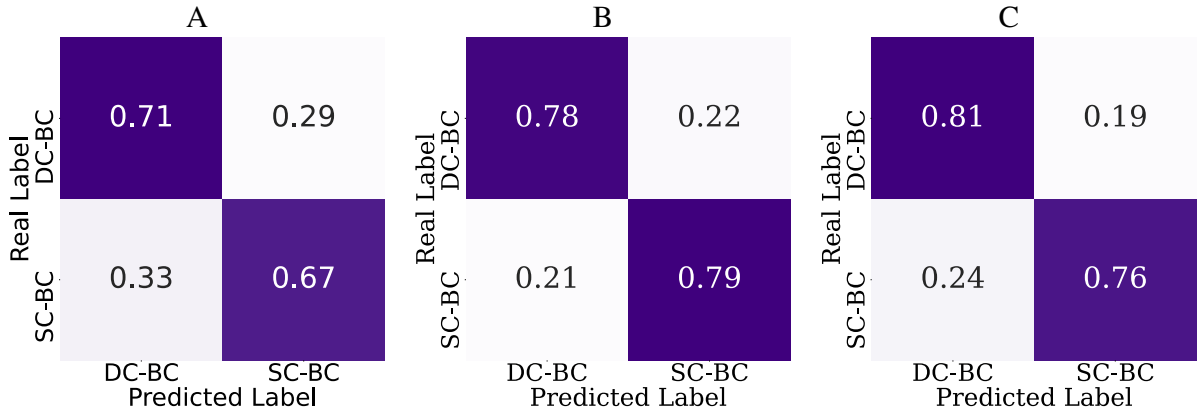

Figure 5.36: Confusion matrices of the highest results for classification of customer satisfaction in the banking call-center dataset using the log-likelihood posterior probabilities: A)Acoustic model-Quadrants. B) Linguistic model-Quadrants. C) Acoustic+Linguistic model-Quadrants.

Figures 5.37 and 5.38 show the bar plot of the log-likelihood posterior probabilities obtained from the acoustic and linguistic model respectively. The bar plots allow showing the behavior of the posteriors more compactly. The bar on the left side of each plot is the posteriors of the SC-BC subjects and on the right side the posteriors of the DC-BC subjects. The difference between SC-BC and DC-BC can be easily observed in the two-class problem of AA vs. PA, and PV vs. NV (see Figures 5.37.A and 5.38.A). Note that for the SC-BC, the arousal tends to be higher for the active and lower for the passive. The valence for the SC-BC subjects tends to be higher for positive and lower for negative, which agrees with what has expected if the customer is satisfied with the service.

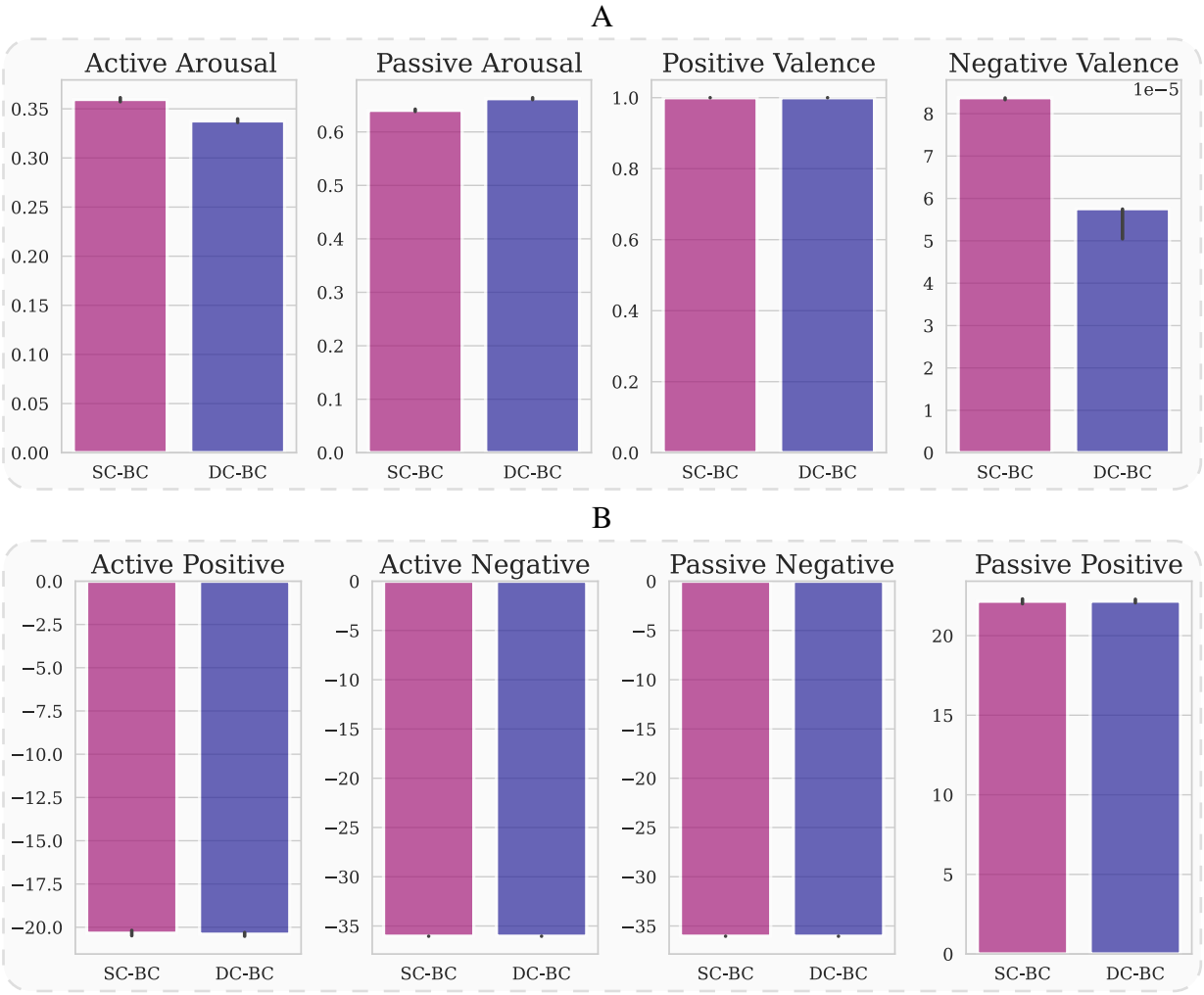

Figure 5.37: Bar plots for customer satisfaction in the banking call-center dataset using the log-likelihood posterior probabilities from the acoustic model: A) Arousal and Valence. B) Quadrants.

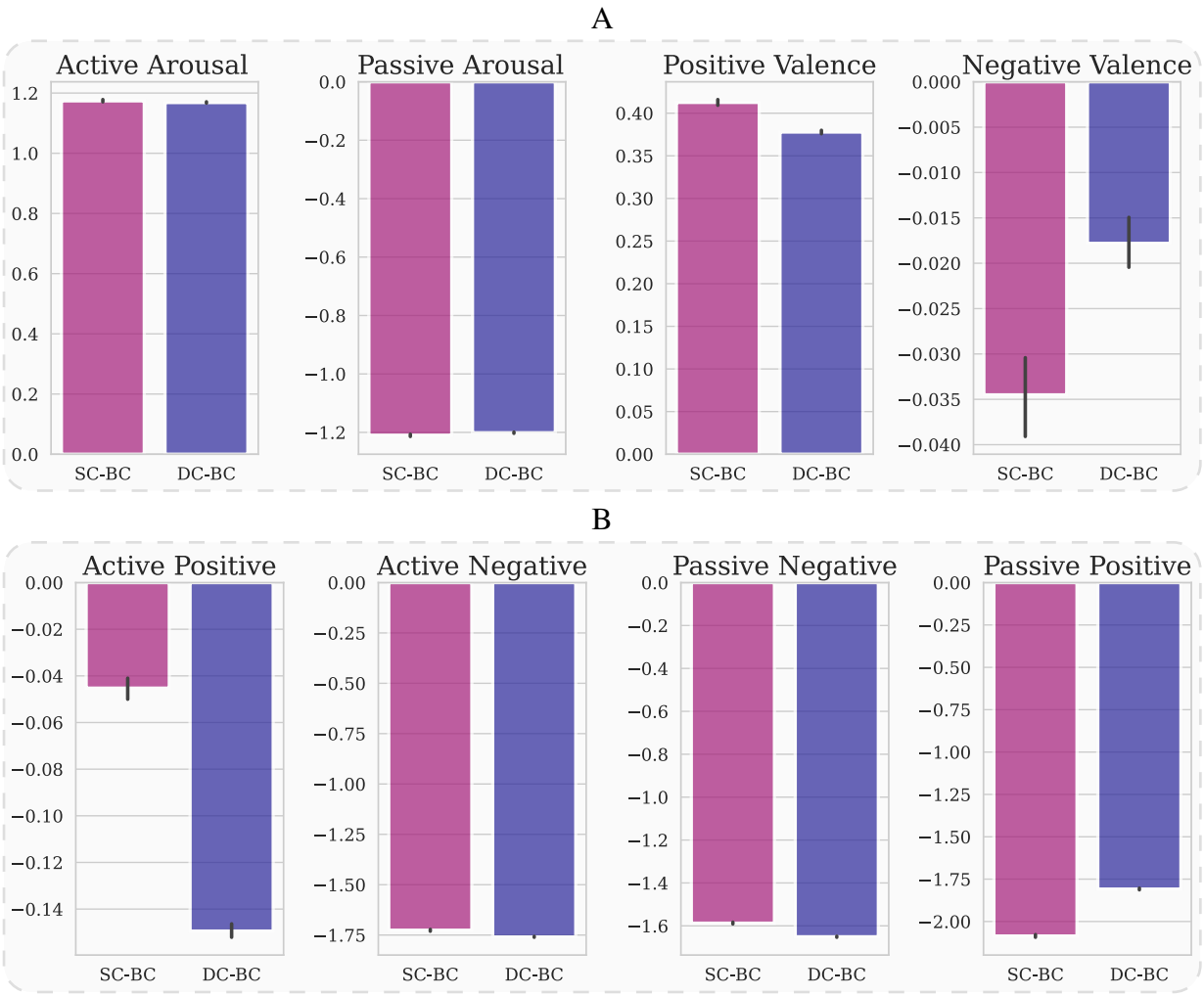

Figure 5.38: Bar plots for customer satisfaction in the banking call-center dataset using the log-likelihood posterior probabilities from the linguistic model: A) Arousal and Valence. B) Quadrants.

Table 5.20 shows the classification performance using the embeddings as input. The highest results are obtained using the early fusion of the arousal and valence and with the quadrants. The information embedded around the entire arousal-valence plane is most accurate for this task in order to capture information related to customer satisfaction.

Table 5.20: Results for classification of customer satisfaction in the banking call-center dataset using the embeddings

Features	Experiment	UAR	F-score	Prec	Rec	AUC
Acoustic Model	Arousal	79.2	0.79	0.79	0.79	0.85
	Valence	74.1	0.74	0.74	0.74	0.84
	Arousal+Valence	78.3	0.78	0.78	0.78	0.85
	Quadrants	74.6	0.75	0.75	0.75	0.84
Linguistic Model	Arousal	84.5	0.85	0.85	0.85	0.93
	Valence	84.5	0.85	0.85	0.85	0.93
	Arousal+Valence	87.3	0.87	0.87	0.87	0.94
	Quadrants	88.5	0.89	0.88	0.88	0.94
Early Fusion Acoustic + Linguistic	Arousal	85.6	0.86	0.86	0.86	0.93
	Valence	83.9	0.84	0.84	0.84	0.92
	Arousal+Valence	85.6	0.86	0.86	0.86	0.93
	Quadrants	84.8	0.85	0.85	0.85	0.93

Notes: **UAR**: unweighted average recall. **Prec**: precision. **Rec**: recall. Unweighted average recall is given in [%].

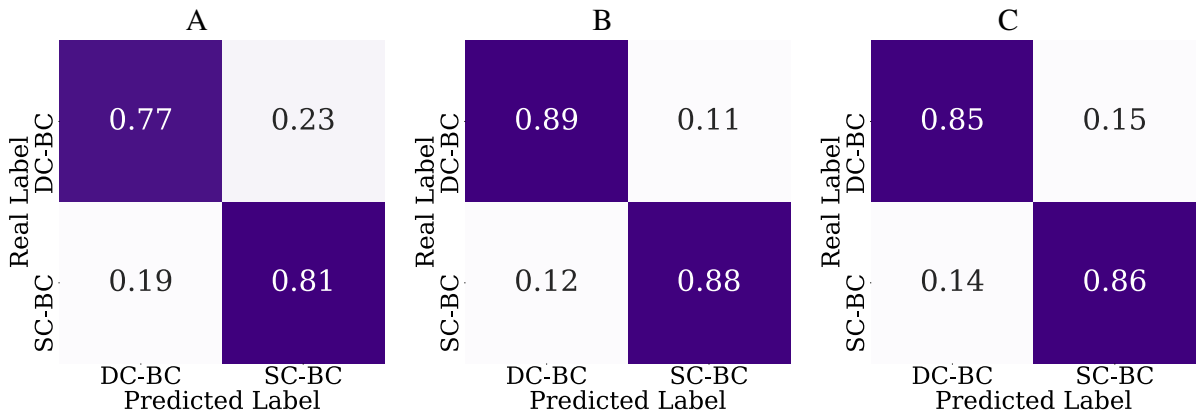

Figure 5.39: Confusion matrices of the highest results for classification of customer satisfaction in the banking call-center dataset using the embeddings: A)Acoustic model-Arousal. B) Linguistic model-Quadrants. C) Acoustic+Linguistic model-Arousal.

Figure 5.39 shows the confusion matrices for the best classification results with the embeddings

using acoustics, linguistic, and early fusion respectively. The sensitivity and specificity are more balanced for linguistics (see Figure 5.39.B), which concludes that the linguistic model using the quadrant information is the most suitable for this task. Note that the performance with respect to the baseline improves with the proposed approach by 6% absolute for acoustics and by 38% absolute for linguistics.

Classification of Depression in Parkinson

This experiment considers the depression in PD dataset (Section 4.5). The same procedure is performed, using the embeddings and log-likelihood posterior probabilities from the linguistic and acoustic models. The classification is performed using a RBF-SVM, following the same validation process described in Sections 5.2 and5.4. Table 5.21 shows the baseline models for acoustics and linguistics to compare the suitability of the proposed approach to discriminate depression in PD. Acoustics consider as baseline openSMILE, which obtains an F-score of 0.55. BERT is used as a baseline for linguistics with an F-score of 0.65.

Table 5.21: Baseline results for classification of D-PD patients

Features	Baseline	UAR	F-score	Prec	Rec
Acoustic	openSMILE	55.0	0.55	0.53	0.53
Linguistic	BERT	65.0	0.65	0.64	0.64.3

Notes: **UAR**: unweighted average recall. **Prec**: precision. **Rec**: recall. Unweighted average recall is given in [%].

The classification performance using the log-likelihood posterior probabilities is shown in Table 5.22. The most accurate results are obtained using the early fusion of arousal and valence for the three models (F-score=0.82). The highest performance is obtained with acoustics, while the early fusion strategy improves the performance for valence classification.

Figure 5.40 shows the confusion matrices for the best classification results using acoustics, linguistic, and early fusion respectively. Note that the acoustics and the early fusion strategy between acoustics and linguistics obtained the same results, i.e., early fusion does not improve the performance.

Figures 5.41 and 5.42 show the bar plot of the log-likelihood posterior probabilities obtained from the acoustic and linguistic model respectively. The bar on the left side of each plot is the posteriors of the D-PD patients and on the right side the posteriors of the ND-PD patients. Note that for the D-PD patients, the arousal tends to be lower for the active and higher for the passive, which may indicate that depression in PD for this task is related to the passive arousal, i.e., emotions in the low quadrants.

Table 5.22: Results for classification of D-PD patients using the log-likelihood posterior probabilities

Features	Experiment	UAR	F-score	Sens	Spe	AUC
Acoustic Model	Arousal	68.3	0.69	0.69	0.68	0.72
	Valence	63.3	0.63	0.63	0.63	0.61
	Arousal+Valence	81.7	0.82	0.82	0.82	0.86
	Quadrants	51.7	0.52	0.51	0.52	0.48
Linguistic Model	Arousal	61.7	0.62	0.64	0.62	0.56
	Valence	66.7	0.67	0.70	0.67	0.68
	Arousal+Valence	70.0	0.71	0.72	0.70	0.71
	Quadrants	51.7	0.52	0.51	0.52	0.48
Early Fusion Acoustic + Linguistic	Arousal	68.3	0.69	0.69	0.68	0.67
	Valence	78.3	0.78	0.80	0.78	0.81
	Arousal+Valence	81.7	0.82	0.82	0.82	0.86
	Quadrants	46.7	0.47	0.47	0.47	0.40

Notes: **UAR**: unweighted average recall. **Prec**: precision. **Rec**: recall. Unweighted average recall is given in [%].

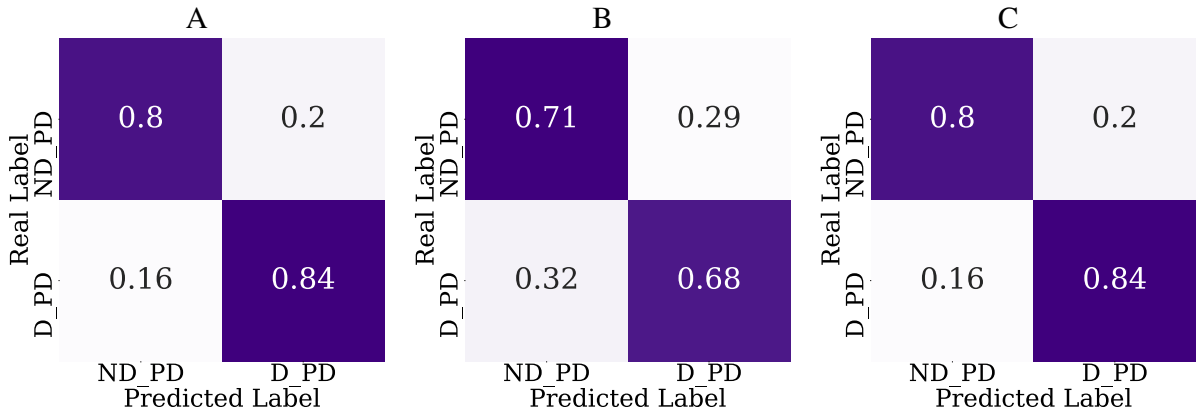

Figure 5.40: Confusion matrices of the highest results for classification of D-PD patients using the posterior probability: A) Acoustic model-Arousal+Valence. B) Linguistic model-Arousal+Valence. C) Acoustic+Linguistic model-Arousal+Valence.

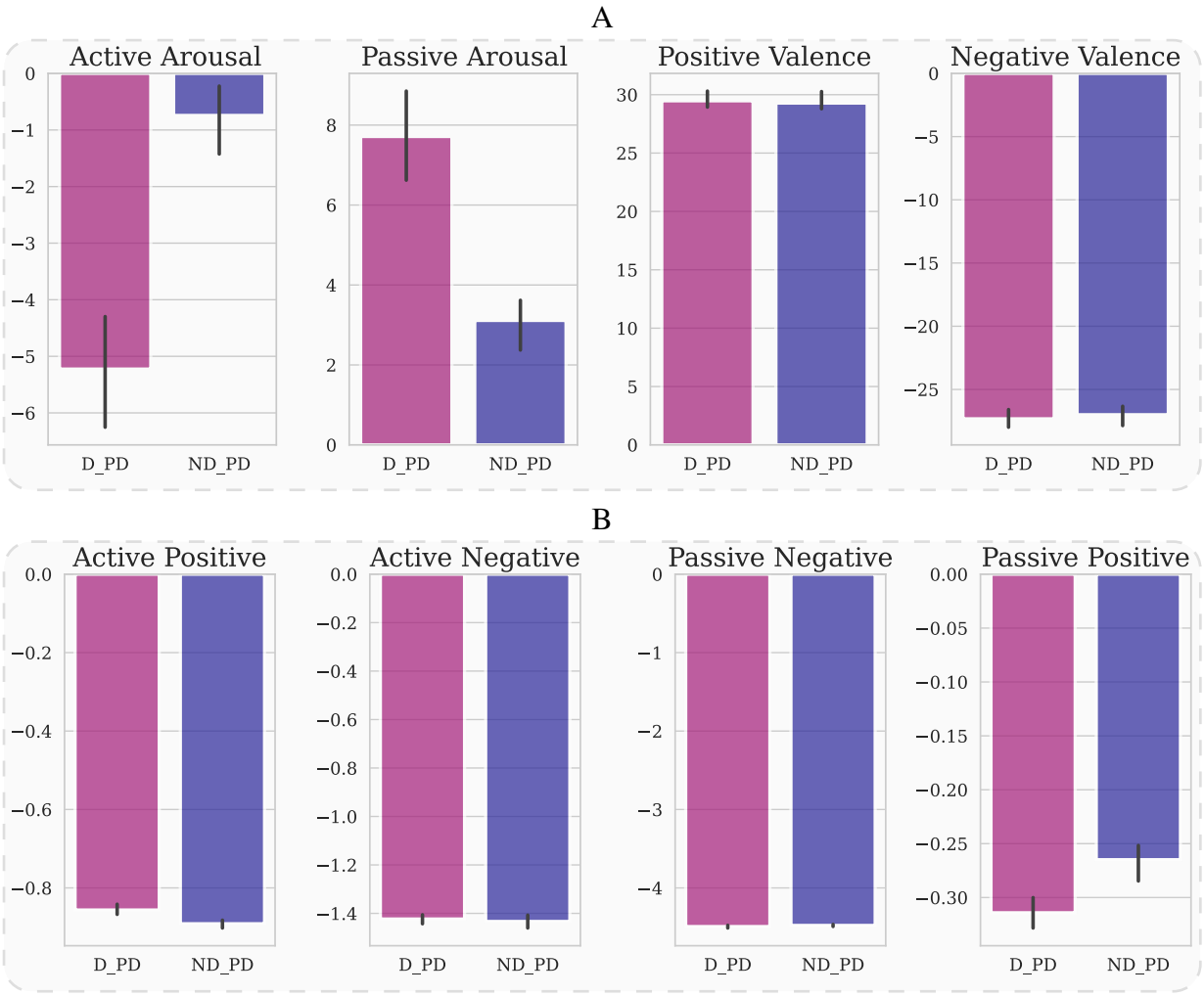

Figure 5.41: Bar plots for D-PD patients using the log-likelihood posterior probabilities from the acoustic model: A) Arousal and Valence. B) Quadrants.

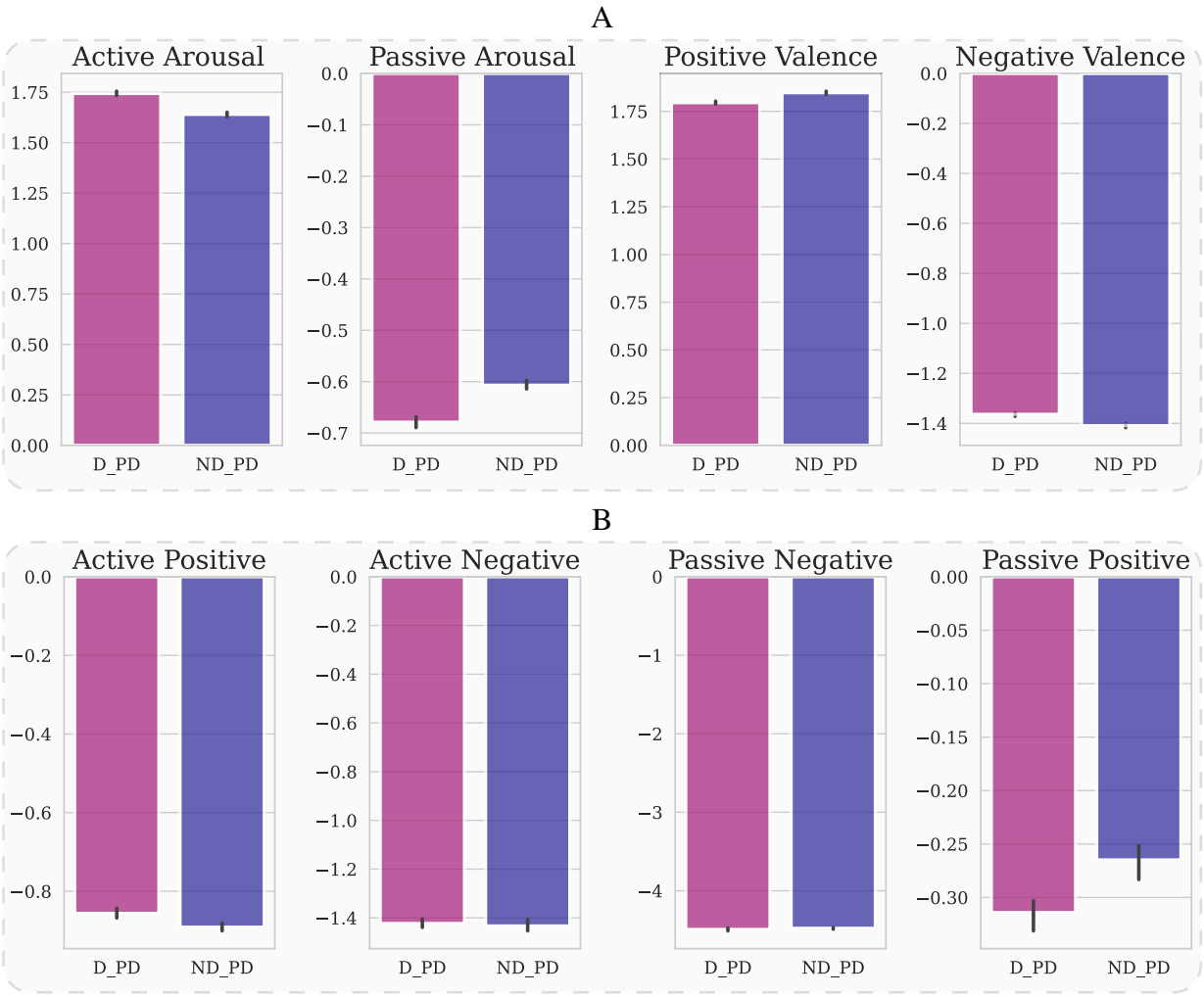

Figure 5.42: Bar plots for D-PD patients using the log-likelihood posterior probabilities from the linguistic model: A) Arousal and Valence. B) Quadrants.

Table 5.23 shows the classification performance using the embeddings as input. The highest results are obtained using the early fusion of the quadrants. The information embedded around the entire arousal-valence plane is most accurate for this task to capture information related to depression in PD. However, the posteriors obtained higher performance for this task in comparison with the embeddings.

Table 5.23: Results for classification of D-PD patients using the embeddings

Features	Experiment	UAR	F-score	Sens	Spe	AUC
Acoustic Model	Arousal	60.0	0.60	0.61	0.60	0.60
	Valence	65.0	0.65	0.65	0.65	0.63
	Arousal+Valence	68.3	0.68	0.68	0.68	0.66
	Quadrants	60.0	0.59	0.59	0.60	0.68
Linguistic Model	Arousal	60.0	0.60	0.60	0.60	0.60
	Valence	56.7	0.54	0.54	0.57	0.57
	Arousal+Valence	60.0	0.59	0.59	0.60	0.56
	Quadrants	66.7	0.67	0.66	0.67	0.63
Early Fusion Acoustic + Linguistic	Arousal	56.7	0.56	0.56	0.57	0.59
	Valence	61.7	0.62	0.62	0.62	0.61
	Arousal+Valence	58.3	0.58	0.59	0.58	0.61
	Quadrants	70.0	0.70	0.70	0.70	0.73

Notes: **UAR**: unweighted average recall. **Prec**: precision. **Rec**: recall. Unweighted average recall is given in [%].

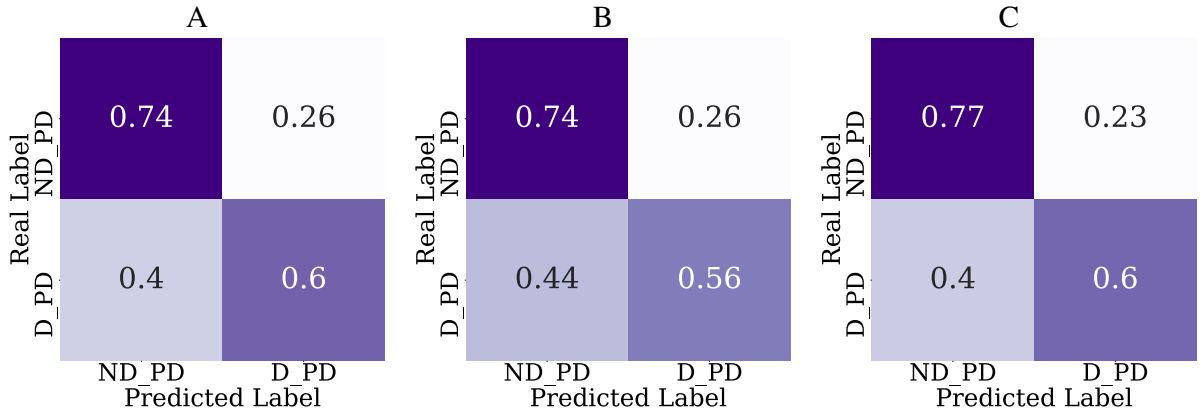

Figure 5.43: Confusion matrices of the highest results for classification of D-PD patients using the embeddings: A)Acoustic model-Arousal+Valence. B) Linguistic model-Quadrants. C) Acoustic+Linguistic model-Quadrants.

Figure 5.43 shows the confusion matrices for the best classification results with the embeddings

using acoustics, linguistic, and early fusion respectively. The sensitivity and specificity are more balanced for acoustics (see Figure 5.43.A). In general, the acoustic model is the most suitable for this task. Note that the performance with respect to the baseline improves with the proposed approach in 27% for acoustics and in 5% for linguistics.

Classification of Alzheimer's Disease

The ADReSS dataset in Section 4.3 is considered for this experiment. This dataset does not contain directly linked emotional or affective labels, but it is used to analyze AD using the arousal-valence plane. The input features consist of the embeddings and log-likelihood posterior probabilities from the linguistic and acoustic models. The classification is performed using an RBF-SVM, using the test set provided by the Interspeech ADReSS challenge 2020 [91]. Table 5.24 shows the baseline models for acoustics and linguistics to compare the suitability of the proposed approach. Acoustics consider as baseline ADReSS with an F-score of 0.62. BERT is used as a baseline for linguistics with an F-score of 0.75.

Table 5.24: Baseline results for classification of AD patients from the ADReSS dataset

Features	Baseline	UAR	F-score	Prec	Rec
Acoustic	ADReSS	62.0	0.62	0.64	0.63
Linguistic	ADReSS-BERT	75.0	0.75	0.77	0.75

Notes: **UAR**: unweighted average recall. **Prec**: precision. **Rec**: recall. Unweighted average recall is given in [%].

Table 5.25 shows the classification performance using the log-likelihood posterior probabilities as input. The highest results are obtained using the arousal for acoustics (F-score=0.75) and the valence for linguistics (F-score=0.80). Linguistics obtained the most accurate results, while the early fusion strategy does not improve the performance of each feature set separately.

Figure 5.44 shows the confusion matrices for the best classification using acoustics, linguistic, and early fusion respectively. Note that the sensitivity and specificity are more balanced for acoustic (see Figure 5.44.A), while linguistics tend to discriminate better AD patients (see Figure 5.44.B).

Figures 5.45 and 5.46 show the bar plot of the log-likelihood posterior probabilities obtained from the acoustic and linguistic model respectively. The bar on the left side of each plot is the posteriors of the AD patients and on the right side the posteriors of the HC subjects. Note that for AD, the arousal tends to be slightly lower for the active and higher for the passive. The valence for the AD patients tends to be higher for positive and lower for negative. The quadrants show

Table 5.25: Results for classification of AD patients from the ADReSS dataset using the log-likelihood posterior probabilities

Features	Experiment	UAR	F-score	Prec	Rec	AUC
Acoustic Model	Arousal	75.0	0.75	0.75	0.75	0.80
	Valence	47.9	0.32	0.24	0.48	0.45
	Arousal+Valence	68.8	0.69	0.69	0.69	0.72
	Quadrants	57.7	0.58	0.59	0.58	0.63
Linguistic Model	Arousal	48.0	0.48	0.48	0.48	0.43
	Valence	79.2	0.80	0.82	0.79	0.85
	Arousal+Valence	52.1	0.47	0.53	0.52	0.63
	Quadrants	54.2	0.51	0.56	0.55	0.72
Early Fusion Acoustic + Linguistic	Arousal	58.3	0.56	0.61	0.58	0.59
	Valence	68.8	0.69	0.69	0.69	0.74
	Arousal+Valence	66.7	0.66	0.67	0.67	0.73
	Quadrants	60.4	0.60	0.61	0.60	0.60

Notes: **UAR**: unweighted average recall. **Prec**: precision. **Rec**: recall. Unweighted average recall is given in [%].

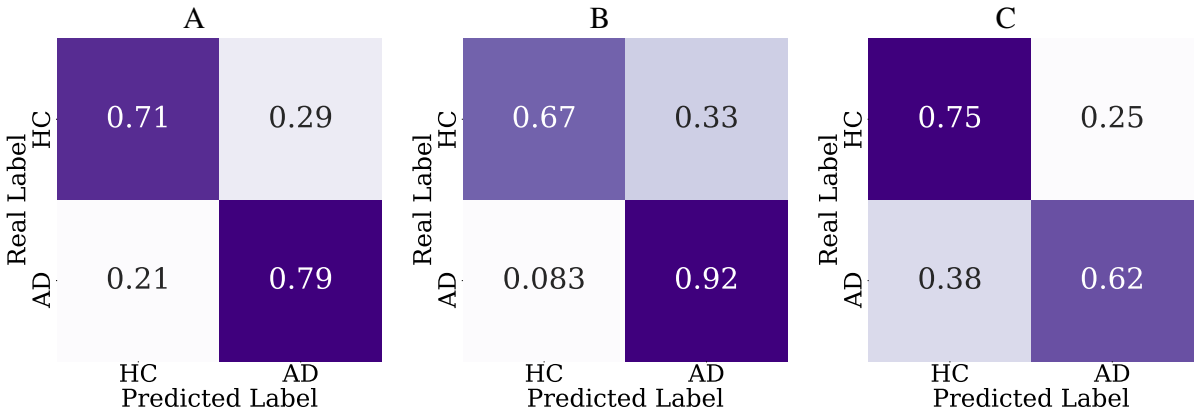

Figure 5.44: Confusion matrices of the highest results of AD patients from the ADReSS dataset using the log-likelihood posterior probabilities: A)Acoustic model-Arousal. B) Linguistic model-Valence. C) Acoustic+Linguistic model-Valence.

that AD patients tend to be lower for active and passive positive, and slightly higher for passive negative. This may be because the labels are not directly linked to affective or emotion. However, there is a visual difference between AD patients and HC according to the bar plot.

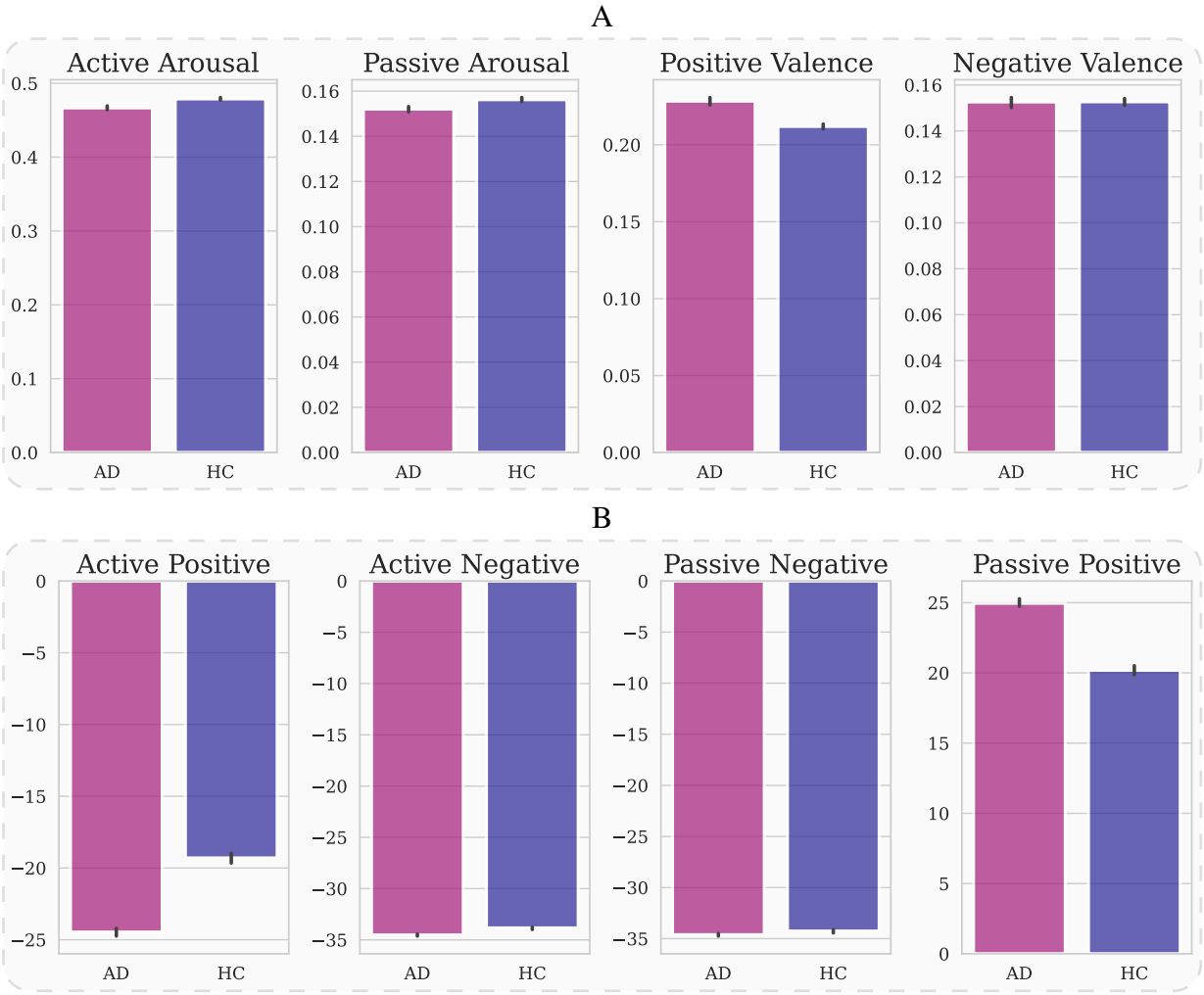

Figure 5.45: Bar plots for AD patients from the ADReSS dataset using the log-likelihood posterior probabilities from the acoustic model: A) Arousal and Valence. B) Quadrants.

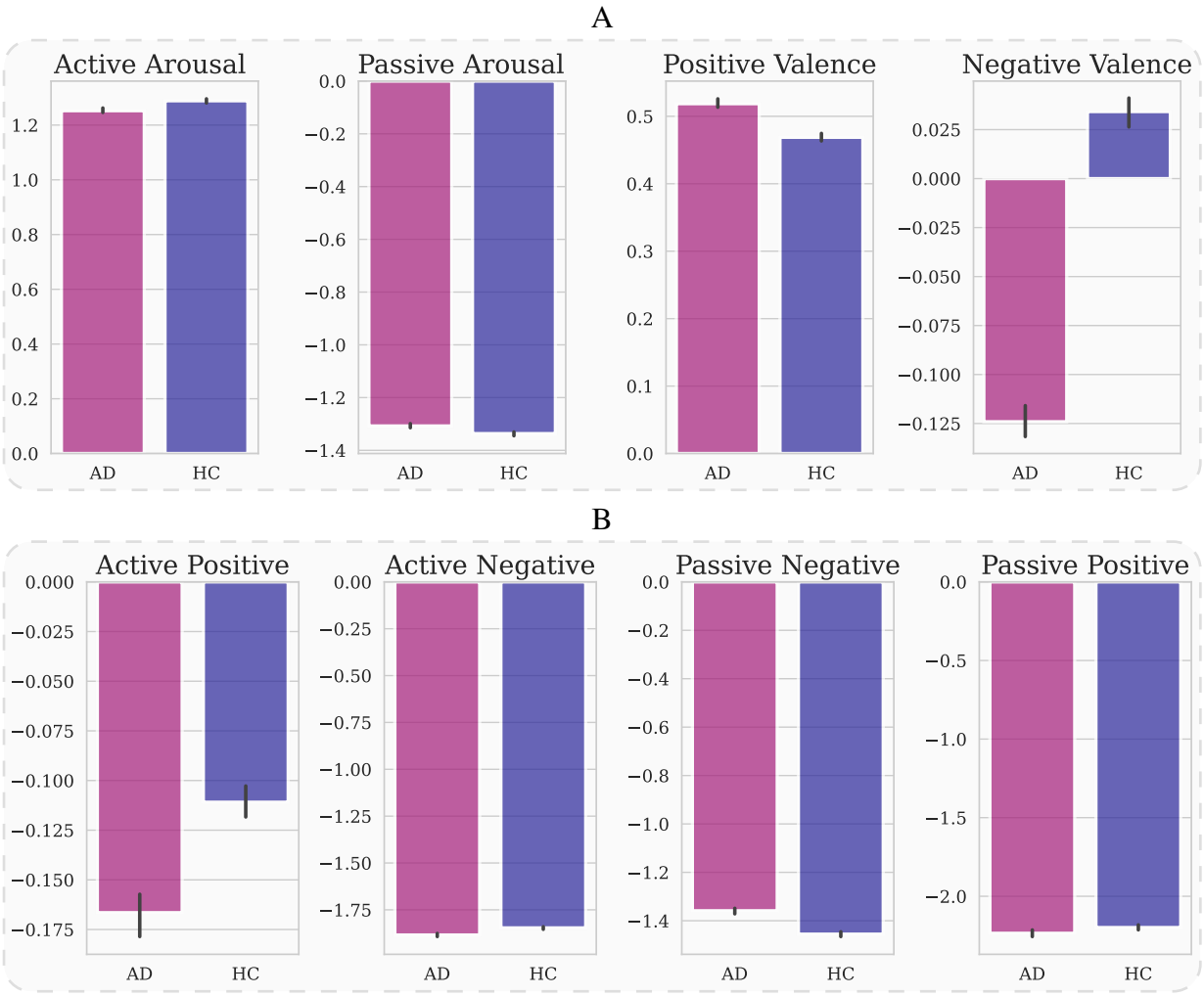

Figure 5.46: Bar plots for AD patients from the ADReSS dataset using the log-likelihood posterior probabilities from the linguistic model: A) Arousal and Valence. B) Quadrants.

Table 5.26 shows the classification performance using the embeddings as input. The highest results are obtained using the early fusion of the arousal and valence for linguistics (F-score=0.69). However, The posteriors obtained higher performance for this task in comparison with the embeddings.

Table 5.26: Results for classification of AD patients from the ADReSS dataset using the embeddings

Features	Experiment	UAR	F-score	Prec	Rec	AUC
Acoustic Model	Arousal	50.0	0.37	0.50	0.50	0.55
	Valence	64.6	0.63	0.68	0.65	0.68
	Arousal+Valence	52.1	0.38	0.76	0.52	0.45
	Quadrants	56.3	0.52	0.60	0.56	0.53
Linguistic Model	Arousal	56.3	0.50	0.63	0.56	0.68
	Valence	66.7	0.67	0.67	0.67	0.71
	Arousal+Valence	68.8	0.69	0.70	0.69	0.68
	Quadrants	56.3	0.50	0.63	0.56	0.66
Early Fusion Acoustic + Linguistic	Arousal	50.0	0.37	0.50	0.50	0.55
	Valence	54.2	0.48	0.58	0.54	0.51
	Arousal+Valence	52.1	0.38	0.76	0.52	0.51
	Quadrants	56.3	0.53	0.59	0.56	0.55

Notes: **UAR**: unweighted average recall. **Prec**: precision. **Rec**: recall. Unweighted average recall is given in [%].

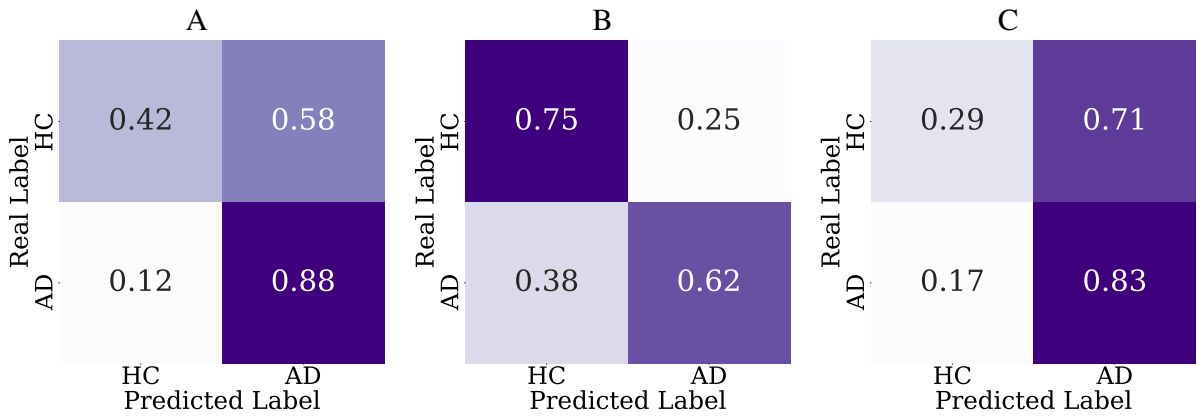

Figure 5.47: Confusion matrices of the highest results for classification of AD patients from the ADReSS dataset using the embeddings: A)Acoustic model-Valence. B) Linguistic model-Arousal+Valence. C) Acoustic+Linguistic model-Quadrants.

Figure 5.47 shows the confusion matrices for the best classification with the embeddings using

acoustics, linguistic, and early fusion respectively. Note that the performance with respect to the baseline improves with the proposed approach in 13% for acoustics and in 5% for linguistics, which concludes that there may be information embedded in the arousal-valence plane to classify AD.

Chapter 6

Summary and Outlook

This thesis proposed methodologies for acoustic and linguistic analyses in different scenarios such as customer satisfaction, cognitive disorders in AD, and depression in PD. Five different experiments were considered: (1) evaluation of customer satisfaction, (2) assessment of Genetic Alzheimer's disease, (3) linguistic analysis to discriminate Parkinson's disease, (4) depression in Parkinson's disease, and (5) User state modeling based on the arousal-valence plane.

6.1 Evaluation of Customer Satisfaction

Call-centers collect many speech recordings from different industry sectors such as banks, insurances, telecommunications, among others. The evaluation of customer satisfaction is performed considering two different datasets related to insurance and banking call-centers (see Section 5.1. The aim was to detect whether a customer is satisfied or dissatisfied with the service. This experiment considers two different datasets from banking and insurance call-centers. The feature set for the acoustic analysis includes methods based on F_0, MFFCs, BBEs, energy, voiced rates, among others. The linguistics features include word-embeddings as W2V and BERT. The results showed that the performed acoustic analysis is more suitable for the banking dataset, and the linguistic analysis for the insurance dataset. The classification of the customer satisfaction in the banking dataset exhibited higher performance using prosody (F-score=0.72), while the early fusion did not improve the results. However, linguistic features provided better results to discriminate customer satisfaction in the insurance dataset (F-score=0.71) as same as the early fusion that outperformed the results with an F-score up to 0.73. According to these results despite the fact that it is the same classification problem, the information derived from different sources may influence the obtained performance for each analysis, since the information is derived from the interaction between client

and advisor for the insurance dataset, and from customer opinions recorded in a voicemail for the banking call-center dataset. Further work will explore additional features as same as methods to segment, analyze, and extract information of each interaction client-advisor.

6.2 Assessment of Genetic Alzheimer's Disease

Acoustic and linguistic analyses were considered to discriminate genetic carriers of the *Paisa mutation* as well as EOA (see Section 5.2). This experiment consisted of two different tasks: (1) GC vs. NGC, and (2) MCI vs. HC. Other classification tasks were not included to avoid the effect of aging between the groups. The classification between MCI patients with early-onset Alzheimer and HC subjects was performed. The acoustic analysis mainly consisted of articulation and prosody, while linguistic analysis was based on word-embeddings methods. Early fusion of articulation and prosody features exhibited the highest performance for MCI patients vs. HC subjects with F-scores of up to 0.74. The linguistic-based analysis did not show satisfactory results in this experiment, which may occur due to the proportion of unknown words that could affect the performance of models. The effectiveness of the word-embeddings methods is directly linked to the number of known words by the predefined vocabulary in the corpus on which they were trained. This may occur due to some unknown words by the algorithms related to characteristic lexicon from the region, or mispronunciations of the words. The unknown words on average are 21.4% of the total of words per utterance using the word-embedding methods. Regarding BERT and BETO models, the results with BERT were slightly higher, which concludes that for our approach the translation to Spanish did not show a strong impact on the results. The same features were used to classify GC vs. NGC. Good results were obtained considering the difficult task that involves the classification of two healthy groups without any AD symptom. Prosody and BERT exhibited the highest performance with F-scores of 0.67 and 0.68, respectively. The influence of depression in the GC subjects was discarded by performing a Man-Whitney U-test ($p = 0.89$) between GC and NGC regarding the geriatric depression scale of Yesavage. According to Table 4.1 and to the results, there is no bias at cognitive (MMSE, MoCA) and depression level between the groups, even though the machine learning algorithm was able to find significant differences above chance between GC vs. NGC. Therefore we need to investigate other possible causes that influenced the results. Further work will explore other features related to linguistics to improve the results, as same as acoustics to analyze the impact of the mutation, and the influence of the demographic aspects for each group. To the best of my knowledge, this is the first study focused on automatically evaluating genetic AD using acoustics and linguistics.

6.3 Linguistic Analysis in Parkinson's Disease

The linguistic analysis in PD was performed in order to analyze the suitability of NLP methods to discriminate PD patients vs. HC subjectssee Section 5.3). The proposed approach is a step forward in the assessment of language impairments that affect communication capabilities in PD. Thus, it allows studying of different communication deficits that cannot be observed in motor activities. The feature sets included classical methods such as BoW and TF-IDF along with other techniques based on word-embeddings like W2V. The classification performance was relatively accurate to classify PD patients and HC subjects with F-scores of up to 0.72. The results suggested that there is information in a language that may reflect disturbances in the communication capabilities of PD patients, as it also observed in previous studies [49]. Further, this information can be used to discriminate between PD and HC subjects and to evaluate the neurological state of the patients. The main limitation of this experiment is related to the task performed by the participants, where they were asked to describe their daily routines. This task cannot reflect properly deficits in the communication of PD patients since it may introduce an explicit bias in the recordings. PD patients tend to do mainly passive activities such as having meals, thinking, and take their medication, while HC subjects showed more variety in their daily activities. Future work will address other linguistic features and the analysis of the disease severity using linguistics. The results reported here were the first step towards the automatic evaluation of language impairments in PD patients in this thesis.

6.4 Depression Assessment in Parkinson's Disease

This thesis proposed an automatic detection of the depression in PD patients, based on acoustic, linguistic information, and an approach based on user modeling (see Section 5.4). It aims to discriminate between depressive and non-depressive PD patients. Depression was labeled according to the depression item in the first part of the MDS-UPDRS evaluation. The acoustic analysis was based on articulation and prosody features, while BERT embeddings were used to perform the linguistic analysis. The GMM-UBM supervector paradigm was addressed to model the acoustic and linguistic features. The early fusion strategy of acoustic and linguistics was the most accurate in classification with F-scores of up to 0.77. The proposed approach using GMM-UBM and the combination of articulation and BERT embeddings increased the performance of 15% in comparison with the baseline model. The results reported here suggest that there is information in speech and language that can be directed linked to the depression state of PD patients. Prosody methods were less accurate, consequently, furthermore prosody related

features will be explored. The main limitation of this experiment is the amount of data that needs to be increased in the future. GITA research group is currently collecting more data for further research not only including tasks of spontaneous conversations but also picture descriptions. Further experiment will consider to explore other fusion techniques, data modeling methods, and other set of features. To the best of my knowledge, this is the first study focused on evaluating depression symptoms in PD patients combining acoustic and linguistic analyses.

6.5 User State Modeling Based on the Arousal-Valence Plane for Customer Satisfaction and Health-Care

In Section 5.5 a novel approach to evaluate scenarios such as customer satisfaction and the assessment of patients with neuro-degenerative diseases is proposed, using deep learning techniques and the arousal-plane information. The proposed methodology focused on analyze emotional and mood changes in acoustics and natural language from spontaneous speech recordings and their transcriptions. This approach considers the arousal-valence plane representation to perform a quantitative analysis of emotions, which distributes the human emotions into a 2-dimensional space to capture information about excitation and polarity of the emotions [6]. Acoustic and linguistic analyses are used to train multimodal models based on the arousal-valence plane representation [6]. Different models were trained to discriminate each quadrant in the arousal-valence plane and to obtain a set of posterior probabilities and embeddings in order to used them as features. The trained models were used in three scenarios related to customer satisfaction and health-care: (1) evaluation of customer satisfaction in call-centers, (2) assessment of depressive symptoms in PD patients, and (3) assessment and classification of AD. The acoustic analysis consisted on Mel spectrograms combined with a CNN-RNN based approach to capture the energy inside the different frequency bands of the spectrogram and prosody information. The linguistic model was based on the combination of BERT embeddings and a Bi-LSTM approach to focus on a smaller sequences based on learned information from BERT. Note that the proposed approach outperformed all baselines in the three classification tasks. On the one hand, the performance improved for acoustics in about 6% for SC vs. DC, 27% for D-PD vs. ND-PD, and 13% for AD vs. HC. On the other hand, the linguistic model outperformed the baseline in about 38% for SC vs. DC, 5% for D-PD vs. ND-PD, and 5% for AD vs. HC. In general, the results with the proposed model obtained highest performance in comparison to the baselines, which concludes that there may be information embedded in the arousal-valence plane to discriminate customer satisfaction, depression in PD, and AD. Customer satisfaction was better discriminate using all of

the information embedded in the entire arousal-valence plane using the quadrants, as same as for depression in PD with the difference that the early fusion between the arousal and valence models obtained the highest results. The discrimination of AD produce higher results using the arousal information for acoustics and the valence information for linguistic. The log-likelihood posterior probabilities showed good performance to classify problems related to neuro-degenerative disease, while the embeddings were suitable to discriminate customer service. Further work will explore more robust approaches and also information from other datasets will be included to improve the performance of the system. Additionally, I am working on training linguistic models with the Spanish translated corpus since this is the original language of most of the tested datasets.

List of Figures

List of Tables

Bibliography

[1] M. L. Knapp, *Essentials of nonverbal communication*. Holt, Rinehart and Winston New York, 1980.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, *et al.*, "Emotion recognition in human-computer interaction", *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[3] M. L. Finucane, A. Alhakami, P. Slovic, and other, "The affect heuristic in judgments of risks and benefits", *Journal of behavioral decision making*, vol. 13, no. 1, pp. 1–17, 2000.

[4] R. E. Wragg and D. V. Jeste, "Overview of depression and psychosis in alzheimer's disease.", *The American journal of psychiatry*, vol. 146, no. 5, pp. 577–587, 1989.

[5] A. Schrag, P. Barone, R. Brown, *et al.*, "Depression rating scales in parkinson's disease: Critique and recommendations", *Movement disorders*, vol. 22, no. 8, pp. 1077–1092, 2007.

[6] J. Russell, "A circumplex model of affect.", *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[7] N. O. Ndubisi, "Service quality: Understanding customer perception and reaction, and its impact on business", *Gadjah Mada International Journal of Business*, vol. 5, no. 2, pp. 207–219, 2003.

[8] N. Hill, G. Roche, and R. Allen, *Customer satisfaction: the customer experience through the customer's eyes*. The Leadership Factor, 2007.

[9] J. Cho, R. Pappagari, P. Kulkarni, *et al.*, "Deep neural networks for emotion recognition combining audio and transcripts", *Proc. Interspeech 2018*, pp. 247–251,

[10] C. Segura, D. Balcells, M. Umbert, *et al.*, "Automatic speech feature learning for continuous prediction of customer satisfaction in contact center phone calls", in *International Conference on Advances in Speech and Language Technologies for Iberian Languages*, Springer, 2016, pp. 255–265.

[11] Y. Park and S. C. Gates, "Towards real-time measurement of customer satisfaction using automatically generated call transcripts", in *Proceedings of the 18th ACM conference on Information and knowledge management*, ACM, 2009, pp. 1387–1396.

[12] F. Boller, T. Mizutani, U. Roessmann, and P. Gambetti, "Parkinson disease, dementia, and alzheimer disease: Clinicopathological correlations", *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 7, no. 4, pp. 329–335, 1980.

[13] A. P. Association *et al.*, *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[14] M. J. Prince, *World Alzheimer Report 2015: the global impact of dementia: an analysis of prevalence, incidence, cost and trends*. Alzheimer's Disease International, 2015.

[15] M. de Salud y Protección Social de Colombia-Grupo Gestión Integrada para la Salud Mental, *Boletin de salud mental demencia,* https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/PP/ENT/Boletin-demencia-salud-mental.pdf, 2017.

[16] M. Lalli, H. Cox, M. Arcila, *et al.*, "Origin of the psen1 e280a mutation causing early-onset alzheimer's disease", *Alzheimer's & Dementia*, vol. 10, S277–S283, 2014.

[17] N. Acosta-Baena, D. Sepulveda-Falla, C. M. Lopera-Gómez, *et al.*, "Pre-dementia clinical stages in presenilin 1 e280a familial early-onset alzheimer's disease: A retrospective cohort study", *The Lancet Neurology*, vol. 10, no. 3, pp. 213–220, 2011.

[18] M. F. Folstein, L. N. Robins, and J. E. Helzer, "The mini-mental state examination", *Archives of general psychiatry*, vol. 40, no. 7, pp. 812–812, 1983.

[19] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, *et al.*, "The montreal cognitive assessment, moca: A brief screening tool for mild cognitive impairment", *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005.

[20] J. T. Olin, L. S. Schneider, I. R. Katz, *et al.*, "Provisional diagnostic criteria for depression of alzheimer disease", *The American journal of geriatric psychiatry*, vol. 10, no. 2, pp. 125–128, 2002.

[21] C. G. Lyketsos and H. B. Lee, "Diagnosis and treatment of depression in alzheimer's disease", *Dementia and geriatric cognitive disorders*, vol. 17, no. 1-2, pp. 55–64, 2004.

[22] M. S. Goodkind, A. Gyurak, M. McCarthy, *et al.*, "Emotion regulation deficits in frontotemporal lobar degeneration and alzheimer's disease.", *Psychology and aging*, vol. 25, no. 1, p. 30, 2010.

[23] J. D. Henry, P. G. Rendell, A. Scicluna, M. Jackson, *et al.*, "Emotion experience, expression, and regulation in alzheimer's disease.", *Psychology and aging*, vol. 24, no. 1, p. 252, 2009.

[24] K. Lopez-de Ipiña, J. B. Alonso, J. Solé-Casals, *et al.*, "On automatic diagnosis of alzheimer's disease based on spontaneous speech analysis and emotional temperature", *Cognitive Computation*, vol. 7, no. 1, pp. 44–55, 2015.

[25] A. König, A. Satt, A. Sorin, *et al.*, "Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease", *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 1, no. 1, pp. 112–124, 2015.

[26] S. Wankerl, E. Nöth, and S. Evert, "An n-gram based approach to the automatic diagnosis of alzheimer's disease from spoken language.", in *Proc. Interspeech 2017*, pp. 3162–3166.

[27] P. Klumpp, J. Fritsch, and E. Noeth, "ANN-based Alzheimer's disease classification from bag of words", in *Speech Communication; 13th ITG-Symposium*, VDE, 2018, pp. 1–4.

[28] O. Hornykiewicz, "Biochemical aspects of Parkinson's disease", *Neurology*, vol. 51, no. 2 Suppl 2, S2–S9, 1998.

[29] J. R. Duffy, *Motor speech disorders: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2013.

[30] O. O. Ojo, N. U. Okubadejo, F. I. Ojini, and M. A. Danesi, "Frequency of cognitive impairment and depression in parkinson's disease: A preliminary case-control study", *Nigerian medical journal: journal of the Nigeria Medical Association*, vol. 53, no. 2, p. 65, 2012.

[31] S. E. Starkstein, H. S. Mayberg, R. Leiguarda, *et al.*, "A prospective longitudinal study of depression, cognitive decline, and physical impairments in patients with parkinson's disease.", *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 55, no. 5, pp. 377–382, 1992.

[32] C. G. Goetz *et al.*, "Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): Scale presentation and clinimetric testing results", *Movement Disorders*, vol. 23, no. 15, pp. 2129–2170, 2008.

[33] M. Seki, K. Takahashi, D. Uematsu, *et al.*, "Clinical features and varieties of non-motor fluctuations in parkinson's disease: A japanese multicenter study", *Parkinsonism & related disorders*, vol. 19, no. 1, pp. 104–108, 2013.

[34] K. M. Smith and D. N. Caplan, "Communication impairment in parkinson's disease: Impact of motor and cognitive symptoms on speech and language", *Brain and language*, vol. 185, pp. 38–46, 2018.

[35] L. L. Murray and L. P. Lenz, "Productive syntax abilities in huntington's and parkinson's diseases", *Brain and Cognition*, vol. 46, no. 1-2, pp. 213–219, 2001.

[36] S. Vanhoutte, M. De Letter, P. Corthals, *et al.*, "Quantitative analysis of language production in parkinson's disease using a cued sentence generation task", *Clinical linguistics & phonetics*, vol. 26, no. 10, pp. 863–881, 2012.

[37] D. Weintraub and M. B. Stern, "Psychiatric complications in parkinson disease", *The American journal of geriatric psychiatry*, vol. 13, no. 10, pp. 844–851, 2005.

[38] H. Soltau, G. Saon, and B. Kingsbury, "The ibm attila speech recognition toolkit", in *2010 IEEE Spoken Language Technology Workshop*, IEEE, 2010, pp. 97–102.

[39] I. Mierswa, M. Wurst, R. Klinkenberg, *et al.*, "Yale: Rapid prototyping for complex data mining tasks", in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2006, pp. 935–940.

[40] N. Kamaruddin, A. W. A. Rahman, and A. N. R. Shah, "Measuring customer satisfaction through speech using valence-arousal approach", in *2016 6th International Conference on Information and Communication Technology for The Muslim World (ICT4M)*, IEEE, 2016, pp. 298–303.

[41] N. Kamaruddin, A. Wahab, and C. Quek, "Cultural dependency analysis for understanding speech emotion", *Expert Systems with Applications*, vol. 39, no. 5, pp. 5115–5133, 2012.

[42] F. R Lucini, L. M. Tonetto, F. S. Fogliatto, *et al.*, "Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews", *Journal of Air Transport Management*, vol. 83, p. 101 760, 2020.

[43] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor", in *Proceedings of the 18th ACM international conference on Multimedia*, ACM, 2010, pp. 1459–1462.

[44] Y. Park, "Automatic call section segmentation for contact-center calls", in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, ACM, 2007, pp. 117–126.

[45] P. Lieberman, E. Kako, J. Friedman, *et al.*, "Speech production, syntax comprehension, and cognitive deficits in parkinson's disease", *Brain and language*, vol. 43, no. 2, pp. 169–189, 1992.

[46] J. L. Cummings, A. Darkins, M. Mendez, *et al.*, "Alzheimer's disease and parkinson's disease: Comparison of speech and language alterations", *Neurology*, vol. 38, no. 5, pp. 680–680, 1988.

[47] I. Rektorova, J. Mekyska, E. Janousova, *et al.*, "Speech prosody impairment predicts cognitive decline in parkinson's disease", *Parkinsonism & related disorders*, vol. 29, pp. 90–95, 2016.

[48] E. Mioshi, K. Dawson, J. Mitchell, *et al.*, "The addenbrooke's cognitive examination revised (ace-r): A brief cognitive test battery for dementia screening", *International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences*, vol. 21, no. 11, pp. 1078–1085, 2006.

[49] A. M. García, F. Carrillo, J. R. Orozco-Arroyave, N. Trujillo, *et al.*, "How language flows when movements don't: An automated analysis of spontaneous discourse in parkinson's disease", *Brain and language*, vol. 162, pp. 19–28, 2016.

[50] K. M. Smith, J. R. Williamson, and T. F. Quatieri, "Vocal markers of motor, cognitive, and depressive symptoms in parkinson's disease", in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2017, pp. 71–78.

[51] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech", *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[52] K. C. Fraser, F. Rudzicz, and G. Hirst, "Detecting late-life depression in alzheimer's disease through analysis of speech and language", in *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, 2016, pp. 1–11.

[53] J. T. Becker, F. Boiler, O. L. Lopez, *et al.*, "The natural history of alzheimer's disease: Description of study cohort and accuracy of diagnosis", *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[54] R. S Bucks and S. A. Radford, "Emotion processing in alzheimer's disease", *Aging & mental health*, vol. 8, no. 3, pp. 222–232, 2004.

[55] D. Bowers, L. Blonder, and K. M. Heilman, *Florida affect battery*. Center for Neuropsychological Studies, Department of Neurology Florida, USA, 1998.

[56] H. Goodglass and E. Kaplan, *Boston Naming Test: scoring booklet*. Lea & Febiger, 1983.

[57] J. Fritsch, S. Wankerl, and E. Nöth, "Automatic diagnosis of alzheimer's disease using neural network language models", in *ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 5841–5845.

[58] G. Gosztolya, V. Vincze, L. Tóth, *et al.*, "Identifying mild cognitive impairment and mild alzheimer's disease based on spontaneous speech using asr and linguistic features", *Computer Speech & Language*, vol. 53, pp. 181–197, 2019.

[59] P. Boersma and D. Weenink, "Praat v. 4.1. 7 [computer software]", *Amsterdam, the Netherlands: Institute of Phonetic Sciences*, 2003.

[60] J. Zsibrita, V. Vincze, and R. Farkas, "Magyarlanc: A toolkit for morphological and dependency parsing of hungarian", 2013.

[61] H. Hotelling, "Analysis of a complex of statistical variables into principal components.", *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[62] V. Vapnik, I. Guyon, and T. Hastie, "Support vector machines", *Mach. Learn*, vol. 20, no. 3, pp. 273–297, 1995.

[63] A. J. Smola and B. Schölkopf, *Learning with kernels*. Citeseer, 1998, vol. 4.

[64] L. Breiman, "Random forests", *UC Berkeley TR567*, 1999.

[65] D. A Reynolds, T. F Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models", *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[66] T. Bayes, "Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s", *Philosophical transactions of the Royal Society of London*, no. 53, pp. 370–418, 1763.

[67] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities", *Ieee Access*, vol. 5, pp. 8869–8879, 2017.

[68] Z. Huang, M. Dong, Q. Mao, and other, "Speech emotion recognition using cnn", in *Proceedings of the 22nd ACM international conference on Multimedia*, ACM, 2014, pp. 801–804.

[69] A. Nassif, I. Shahin, I. Attili, *et al.*, "Speech recognition using deep neural networks: A systematic review", *IEEE Access*, vol. 7, pp. 19 143–19 165, 2019.

[70] T. Mikolov, K. Chen, G. Corrado, *et al.*, "Word2vec", *URL https://code. google. com/p/-word2vec*, 2013.

[71] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation", in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[72] J. Devlin, M. Chang, K. Lee, and other, "Bert: Pre-training of deep bidirectional transformers for language understanding", *arXiv preprint arXiv:1810.04805*, 2018.

[73] D. M. Powers, "Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation", 2011.

[74] C. Darwin and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[75] S. Basu, J. Chakraborty, A. Bag, *et al.*, "A review on emotion recognition using speech", in *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, IEEE, 2017, pp. 109–114.

[76] D. Hussein, "A survey on sentiment analysis challenges", *Journal of King Saud University-Engineering Sciences*, vol. 30, no. 4, pp. 330–338, 2018.

[77] H. Schröter, T. Rosenkranz, A. N. Escalante-B, and A. Maier, *Clc: Complex linear coding for the dns 2020 challenge*, 2020. arXiv: 2006.13077 [eess.AS].

[78] S. G Koolagudi and K. S. Rao, "Emotion recognition from speech: A review", *International journal of speech technology*, vol. 15, no. 2, pp. 99–117, 2012.

[79] B. Stasiak and K. Rychlicki-Kicior, "Fundamental frequency extraction in speech emotion recognition", in *International Conference on Multimedia Communications, Services and Security*, Springer, 2012, pp. 292–303.

[80] K. R. Scherer, "Vocal affect expression: A review and a model for future research", *Psychological bulletin*, vol. 99, no. 2, p. 143, 1986.

[81] I. Fonagy and K. Magdics, "Emotional patterns in intonation and music", *STUF-Language Typology and Universals*, vol. 16, no. 1-4, pp. 293–326, 1963.

[82] G. Fairbanks and W. Pronovost, "An experimental study of the pitch characteristics of the voice during the expression of emotion", *Communications Monographs*, vol. 6, no. 1, pp. 87–104, 1939.

[83] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion", *The Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, 1993.

[84] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2095–2103, 2007.

[85] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)", *The Journal of the Acoustical Society of America*, vol. 33, no. 2, pp. 248–248, 1961.

[86] K. S. Rao and S. G. Koolagudi, *Robust emotion recognition using spectral and prosodic features*. Springer Science & Business Media, 2013.

[87] C. E. Williams and K. N. Stevens, "Vocal correlates of emotional states", *Speech Evaluation in Psychiatry*, pp. 221–240, 1981.

[88] P. Johnson-Laird and K. Oatley, "The language of emotions: An analysis of a semantic field", *Cognition and emotion*, vol. 3, no. 2, pp. 81–123, 1989.

[89] S. Bird and E. Loper, "Nltk", in *Proceedings of the COLING/ACL on Interactive presentation sessions*, 2004, pp. 69–72.

[90] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need", in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[91] S. Luz, F. Haider, and S. o. de la Fuente, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge", *arXiv preprint arXiv:2004.06833*, 2020.

[92] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, *et al.*, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease", in *LREC*, 2014, pp. 342–347.

[93] C. Busso, M. Bulut, C. C. Lee, *et al.*, "Iemocap: Interactive emotional dyadic motion capture database", *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[94] J. R. Orozco-Arroyave, *Analysis of speech of people with Parkinson's disease*. Logos Verlag Berlin GmbH, 2016, vol. 41.

[95] J. R. Orozco-Arroyave, J. C. Vásquez-Correa, and E. Nöth, "Current methods and new trends in signal processing and pattern recognition for the automatic assessment of motor impairments: The case of parkinson's disease", *Chapter 8, In: Neurological Disorders and Imaging Physics. Applications in dyslexia, epilepsy and Parkinson's. Eds.*, vol. 5, 2020.

[96] A. M. Badshah, J. Ahmad, N. Rahim, *et al.*, "Speech emotion recognition from spectrograms with deep convolutional neural network", in *2017 international conference on platform technology and service (PlatCon)*, IEEE, 2017, pp. 1–5.

[97] M. Chen, X. He, J. Yang, *et al.*, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition", *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[98] R. Fernandez, A. Rendel, B. Ramabhadran, *et al.*, "Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks.", in *Proc. Interspeech 2014*, pp. 2268–2272.

[99] R. Fernandez, A. Rendel, and B. a. Ramabhadran, "Using deep bidirectional recurrent neural networks for prosodic-target prediction in a unit-selection text-to-speech system", in *Proc. Interspeech 2016*.

[100] B. Schuller, S. Steidl, A. Batliner, *et al.*, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language", in *Proc. Interspeech 2016*.

Appendices

Appendix A

Regular vs. Proposed Leave-One-Speaker-Out Strategy to Classify Genetic Alzheimer's Disease

Table A.1 shows the results of the classification of genetic AD (Section 5.2) using the regular LOSO as same as the proposed validation strategy. Note that for this experiment only prosody obtained higher results using the proposed approach in both classification tasks. The average of the F-scores for each strategy was computed, where the regular approach was 3% higher. It may confirms that for this experiment the results performed with the proposed strategy are less optimistic.

Table A.1: Comparison between the regular and the proposed LOSO strategy for the classification of genetic AD.

Features	Experiment	F1-Reg LOSO	F1-Prop LOSO
GC vs. NGC			
Acoustic	Articulation	0.50	0.47
	Prosody	0.65	0.67
Linguistic	W2V	0.58	0.53
	BERT	0.69	0.68
	BETO	0.67	0.65
MCI vs. HC			
Acoustic	Articulation	0.69	0.66
	Prosody	0.65	0.66
Linguistic	W2V	0.53	0.48
	BERT	0.56	0.50
	BETO	0.53	0.50
Average		0.61	0.58

Appendix B

Conferences & Publications

The following publications were derived from the development of this thesis, and the joint work with the GITA research group[1] and the Pattern Recognition Lab (LME)[2]:

B.1 Journals

- **Pérez-Toro, P. A.**, Vásquez-Correa, J. C., Arias-Vergara, T., Nöth, E., & Orozco-Arroyave, J. R.(2020). "Nonlinear dynamics and Poincaré sections to model gait impairments in different stages of Parkinson's disease"(2020). *Nonlinear Dynamics volume*,100,(pp. 3253–3276).

- Orozco-Arroyave, J. R., Vásquez-Correa, J. C., Klumpp, P., **Pérez-Toro, P. A.**, Escobar-Grisales, D., Roth, N., et al. "Apkinson: the smartphone application for telemonitoring Parkinson's patients through speech, gait and hands movement" (2020). *Neurodegenerative Disease Management*, 10(3), (pp. 137-157)

- **Pérez-Toro, P. A.**, Vásquez-Correa, J. C., Bocklet, T., Nöth, E., & Orozco-Arroyave, J. R. "User State Modeling Based on the Arousal-Valence Plane: Applications in Customer Satisfaction and Health-Care". *IEEE Transactions on Affective Computing* , under rewiew.

[1] https://gita.udea.edu.co/
[2] https://lme.tf.fau.de/

B.2 Book Chapters

- **Pérez-Toro, P.A.**, Vásquez-Correa, J.C., Strauss, M., Orozco-Arroyave, J.R., & Nöth, E. (2019, September). "Natural Language Analysis to Detect Parkinson's Disease". *In International Conference on Text, Speech, and Dialogue* (pp. 82-90). Springer, Cham.

B.3 Conferences

- Vásquez-Correa, J.C., Arias-Vergara, T., Klumpp, P., Strauss, M., Küderle, A., Roth, N., Bayerl, S., Garcıa-Ospina, N., **Pérez-Toro, P.A.**, Parra-Gallego, L.F, Rios-Urrego, C.R., Escobar-Grisales, Orozco-Arroyave, J.R., D., Eskofier, B., & Nöth, E. (2019). "Apkinson: a Mobile Solution for Multimodal Assessment of Patients with Parkinson's Disease". *Proc. Interspeech 2019*, (pp. 964-965).

- Klumpp, P., Arias-Vergara, T., Vásquez-Correa, J. C., **Pérez-Toro, P. A.**, Hönig, F., Nöth, E., & Orozco-Arroyave, J. R. (2020). Surgical mask detection with deep recurrent phonetic models. Proc. Interspeech 2020, (pp. 2057-2061).

- **Pérez-Toro, P. A.**,Vásquez-Correa, J. C., Arias-Vergara, T., Klumpp, P., et al. "Acoustic and Linguistic Analyses to Assess Early-Onset and Genetic Alzheimer's Disease". *2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, in press.

- Vásquez-Correa, J. C., Arias-Vergara, T., Klumpp, P., **Pérez-Toro, P. A.**, Orozco-Arroyave, J. R., & Nöth, E. "End-2-End Modeling of Speech and Gait from Patients with Parkinson's disease: Comparison between High Quality vs. Smartphone Data". *2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, in press.

Appendix C

Academic events

- Talk (2020). "Depression Assessment in Parkinson's Disease Using Acoustic Analysis and Natural Language Processing". *Apkinson-workshop. Voice and speech analysis. Open Lecture and Workshop*. IEEE Poland Section. AGH University of Science and Technology. Krakow, Poland.
- Workshop 2020, "Speech processing and understanding". Czech Technical University in Prague and Friedrich-Alexander-Universität Erlangen-Nürnberg. Prague, Czech Republic.
- Workshop 2019, "Speech processing and understanding". Friedrich-Alexander-Universität Erlangen-Nürnberg and Czech Technical University in Prague. Erlangen, Germany.
- Speech and Movement Analysis using your SMArt phone for neurological diseases (SMA^2). Financed by Bundesministerium für Bildung und Forschung (BMBF). 2018 – 2019. Role: Co-researcher.
- Exchange student at Friedrich-Alexander-Universität Erlangen-Nürnberg (15.10.2019-30.03.2021)

Appendix D

Awards and honors

- Research stay at Pattern Recognition Lab-LME (15.10.2019-31.04.2020). This grant was funded by Bayerische Hochschulzentrum für Lateinamerika-BAYLAT.
- Scholarship "Estudiante instructor" (01.02.2019-31.12.2020). This grant is currently funded by the University of Antioquia.
- Scholarship for exchange students from Universities with a partnership agreement with Friedrich-Alexander-Universität Erlangen-Nürnberg. This grant is currently funded by Deutscher Akademischer Austauschdienst-DAAD.