TÉCNICO LISBOA

Speech Biomarker Analysis: Application in Psychiatry

João Pedro Santos Monteiro

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisors: Prof. João Miguel Raposo Sanches
Prof. Miguel de Sequeiros Constante

Examination Committee

Chairperson: Prof. Rita Homem de Gouveia Costanzo Nunes
Supervisor: Prof. João Miguel Raposo Sanches
Member of the Committee: Prof. Alberto Abad Gareta

October 2024

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Preface

The work presented in this thesis was carried out at the Instituto de Sistemas e Robótica, Instituto Superior Técnico (Lisboa, Portugal), starting in September 2023 and ending in October 2024. It was carried out under the supervision of Professor **Prof. João Sanches** at Instituto Superior Técnico and **Prof. Miguel Constante** from the Department of Psychiatry, Hospital Beatriz Ângelo.
During this period, I had the opportunity to work on a project that aimed to develop a tool for affective disorder assessment using speech processing techniques with weekly meetings with the supervisors.

Acknowledgments

I would like to thank my parents for their encouragement and love at the most stressful time in my life. I would also like to thank Leonor Rodrigues for keeping me sane and supporting me through the final stages of this Thesis.

I would also like to acknowledge my dissertation supervisor Prof. João Sanches for his guidance, support and sharing of knowledge that has made this project possible.

To each and every one of you – Thank you.

Abstract

Affective disorders, including Major Depressive Disorder (MDD) and Bipolar Disorder (BD), impact over 320 million people globally. Current diagnosis and monitoring rely on evaluating symptoms through questionnaires and interviews and with the evolution of telemedicine technologies there is the possibility of conducting these interviews through virtual videoconferencing.

This thesis explores the potential of extracting speech biomarkers, in real time, from audio recordings of patients during telemedicine consultations. The main objective is to develop a system that can assist in the assessment and monitoring of affective disorders.

Through an extensive literature review, the most relevant speech features for affective disorders are identified and the most effective methods for their extraction are discussed. The selected group of features includes Jitter, Shimmer, Pitch, Harmonics-to-Noise Ratio, Formants, and other more specific prosodic measurements.

A Python GUI Application is demonstrated, which not only integrates the over time display of the selected speech features values but also integrates an analysis module of valuable depression biomarkers collected from the patients' video stream.

In order to validate the system, a group of benchmark tests were conducted using audio data from healthy individuals. The results show that the obtained data is consistent with the expected values for healthy individuals. Another study conducted using speech data from the DAIC-WOZ dataset also permitted to infer that the system is capable of detecting similar biomarker variation as the one characterized in the state-of-the-art when comparing the speech of individuals with and without depression.

Keywords

Telemedicine; Affective Disorders; Speech Processing; Machine Learning; Real-time Analysis.

Resumo

As perturbações afectivas, como a perturbação depressiva major (PDM) e a perturbação afetiva bipolar (PAB), afetam mais de 320 milhões de pessoas no mundo. O diagnóstico e monitorização destas doenças baseiam-se na avaliação de sintomas através de questionários e entrevistas. Com a evolução das tecnologias de telemedicina, estas entrevistas podem agora ser realizadas através de videoconferência.

Esta tese explora o potencial de extrair biomarcadores da fala, em tempo real, a partir de gravações de áudio de pacientes durante consultas de telemedicina, desenvolvendo um sistema que auxilia na avaliação e monitorização de perturbações afectivas. Através de uma revisão bibliográfica, são identificadas e discutidas as características da voz mais relevantes para este tipo de perturbações e os métodos eficazes para a sua extração. Foram selecionados: Jitter, Shimmer, Frequência Fundamental da Voz, Índice sinal-ruído, Formantes e outras medidas prosódicas tais como velocidade da fala e número de pausas.

É apresentada uma aplicação GUI em Python que permite a visualização temporal das características da fala selecionadas, além de um módulo de extração de biomarcadores do estado mental do paciente, obtidos através da análise do sinal do seu vídeo.

Para validar o sistema, foi conduzido um grupo de testes de referência utilizando dados de áudio de indivíduos saudáveis, demonstrando consistência com os valores esperados. Outro estudo, usando dados do dataset DAIC-WOZ, confirmou a capacidade do sistema em detetar variações de biomarcadores entre indivíduos com e sem depressão.

Palavras Chave

Telemedicina; Perturbações Afetivas; Processamento de Fala; Aprendizagem Automática; Análise em tempo real.

Contents

1 Introduction **1**

1.1 Topic Overview . . . 2

1.2 Motivation . . . 3

1.3 Objectives . . . 4

1.4 Innovations . . . 4

1.5 Methodology . . . 5

1.5.1 Search Method . . . 5

1.5.2 Eligibility criteria . . . 5

1.6 Report Layout . . . 5

2 Background **7**

2.1 Clinical Background . . . 8

2.1.1 Affective Disorders . . . 8

2.1.2 Physiological Process of Speech Production . . . 11

2.1.3 Speech Biomarkers . . . 12

2.1.4 Telemedicine . . . 12

2.2 Technical Background . . . 15

2.2.1 Brief History of Speech Processing . . . 15

2.2.2 Speech Processing Tasks . . . 16

2.2.3 Speech Features . . . 18

2.2.4 Overview of Speech-Based ML Technologies . . . 23

3 State-Of-The-Art **27**

3.1 Literature Review . . . 28

3.2 Feature Selection . . . 30

3.2.1 Feature Value Thresholds . . . 31

3.2.2 Feature Behavior in Affective Disorders . . . 32

3.3 Conclusions . . . 33

4 Methods **35**
4.1 Audio Processing Pipeline . . . 36
4.1.1 Audio Capture . . . 36
4.1.2 Feature Extraction . . . 37
4.1.3 Tool Versatility . . . 40
4.2 Python GUI . . . 40
4.2.1 Pre-Existent Graphical User Interface . . . 40
4.2.2 Multimodal Graphical User Interface . . . 42
5 Results **45**
5.1 Efficacy Tests . . . 46
5.1.1 Syllable Counting . . . 46
5.1.2 Feature Benchmarking . . . 47
5.2 Feature Behaviour with Real-World Data . . . 49
5.3 Conclusions . . . 51
6 Conclusion **53**
6.1 Future Work . . . 54
6.2 Final Remarks . . . 55
Bibliography **55**
A Support Algorithms **65**
A.1 Fourier Transform (FT) . . . 65
A.1.1 Discrete Fourier Transform (DFT) . . . 65
A.1.2 Fast Fourier Transform (FFT) . . . 66
A.2 Long Short-Term Memory (LSTM) . . . 66
A.3 Welford Algorithm . . . 67
B Supplementary Material **69**
B.1 Pitch Benchmarking Figures . . . 70
B.2 Jitter Benchmarking Figures . . . 71
B.3 Shimmer Benchmarking Figures . . . 72
B.4 Formant Benchmarking Figures . . . 73
B.5 Example Report . . . 74
B.6 Software Guide . . . 75
B.6.1 Miminum System Requirements . . . 75
B.6.2 Installation . . . 75
B.6.3 Starting Conditions . . . 75

List of Figures

2.1 Speech production system sketch [1, p.8] . . . 11
2.2 Pipeline from a classical SER application [2, p. 2] . . . 17
2.3 Pipeline from a real-time diarization application [3] . . . 17
2.4 Graphical representation of shimmer and jitter [4, p. 176] . . . 20
2.5 Spectrum of a speech segment annotated with its formants [5] . . . 20
2.6 Spectrogram of a speech signal with Fundamental Frequency (F0) annotated . . . 22
2.7 Graph of mel-scale values vs Hertz scale values . . . 22
2.8 Graphical illustration of different Machine Learning (ML) models [6] . . . 24
2.9 Transformer of two stacked encoders and decoders by Jay Alammar [7] . . . 25

4.1 Representation of the Audio Processing Pipeline. . . . 36
4.2 Abstract scheme of the Response Time calculation with a binary representation of the Voice Activity Detection (VAD) output. . . . 38
4.3 Pre-existent Graphical User Interface with real time video processing. . . . 41
4.4 Example of a hover event over the Pitch feature. . . . 42
4.5 Multimodal Graphical User Interface with real time video and audio processing. . . . 43
4.6 Graph windows for the Video Signal Features (left) and Audio Signal Features (right). . . . 43

5.1 Graphical representation of the overall Pitch values obtained in the benchmark separated by sex . . . 48
5.2 Graphical representation of the overall **Jitter** values obtained in the benchmark separated by sex . . . 48
5.3 Initial window of the developed tool . . . 51
5.4 Final version of the developed tool's Graphical User Interface (GUI) in a split screen configuration with Zoom . . . 52

B.1 Graphs of the Pitch values obtained in the benchmark . . . 70
B.2 Graphs of the Jitter values obtained in the benchmark . . . 71

B.3 Graphs of the Shimmer values obtained in the benchmark 72
B.4 Graphs of the Formant values obtained by the system and the database documentation . 73

List of Tables

3.1 Research articles found on speech processing for affective disorders, with the respective year of publication, features extracted, evaluation method and computational approaches. 28
3.1 Research articles found on speech processing for affective disorders, with the respective year of publication, features extracted, evaluation method and computational approaches (Continued). 29
3.2 Normative feature measurements presented in Isabel Guimarães' book [8], organized by study author, age range and gender, vowel phonation analyzed and features extracted. . 32

5.1 Syllable counting comparison for English and Portuguese sentences. 46
5.2 Average feature values extracted from speech data of individuals with and without Major Depressive Disorder (MDD) (*$p < .05$). 50

B.1 Example of an audio data report generated by the audio section of the multimodal tool. . 74

Acronyms

AI	Artificial Intelligence
ANNs	Artificial Neural Networks
APQ	Amplitude Perturbation Quotient
ASR	Automatic Speech Recognition
BASIS-24	24-item Behavior and Symptom Identification Scale
BDI-II	Beck Depression Inventory-Second Edition
BD	Bipolar Disorder
CHUC	Centro Hospitalar e Universitário de Coimbra
CBT	Cognitive Behavioral Therapy
CNNs	Convolutional Neural Networks
CNTS	Portuguese National Centre of Telehealth
CPP	Cepstral Peak Prominence
CPPS	Cepstral Peak Prominence Smoothed
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DL	Deep Learning
DNNs	Deep Neural Networks
DSM-5	Diagnostic and Statistical Manual of Mental Disorders, 5th Edition
EAR	Eye Aspect Ratio
F0	Fundamental Frequency
FFT	Fast Fourier Transform
FT	Fourier Transform
GMMs	Gaussian Mixture Models

GNE	Glottal-to-Noise Excitation Ratio
GUI	Graphical User Interface
HAMD	Hamilton Depression Rating Scale
HAMD-17	Hamilton Depression Rating Scale 17-item version
HF	High Frequency
HMMs	Hidden Markov Models
HNR	Harmonics-to-Noise Ratio
HPP	The Health Policy Partnership
HRV	Heart Rate Variability
Hz	Hertz
ICD-10	International Statistical Classification of Diseases and Related Health Problems, 10th Revision
IDFT	Inverse Discrete Fourier Transform
IPT	Interpersonal Psychotherapy
LF	Low Frequency
LSTM	Long Short-Term Memory
LPC	Linear Predictive Coding
MAE	Mean Absolute Error
MDD	Major Depressive Disorder
MFCC	Mel-Frequency Cepstral Coefficient
MHA	Multi-Head Attention
ML	Machine Learning
NLP	Natural Language Processing
PALOP	Portuguese-speaking African countries
PENTS	National Strategic Telehealth Plan
PHQ-8	Eight-Item Patient Health Questionnaire
PMR	Psychomotor Retardation
PPQ	Period Perturbation Quotient
PSD	Power Spectral Density
QIDS	Quick Inventory of Depressive Symptomatology

RMSE Root Mean Squared Error

RMSSD Root Mean Square of Successive RR interval Differences

RNN Recurrent Neural Network

SD Speaker Diarization

SDNN Standard Deviation of NN Intervals

SER Speech Emotion Recognition

SF-12 Short Form Health Survey

SI Suicidal Ideation

SNRI Serotonin-Norepinephrine Reuptake Inhibitor

SR Speech Recognition

SSRI Selective Serotonin Reuptake Inhibitor

SVM Support Vector Machine

TEOC Teager energy operator coefficients

TCA Tricyclic Antidepressant

TTS Text-to-Speech

VAD Voice Activity Detection

1

Introduction

Contents

1.1 Topic Overview 2
1.2 Motivation 3
1.3 Objectives 4
1.4 Innovations 4
1.5 Methodology 5
1.6 Report Layout 5

As Aristotle once said, “Man is by nature a social animal”, which means that we, as a species, had to perfect the way that information is exchanged among community members in order to promote increasingly complex societal organizations that led to the civilization that we know today. Communication among human beings can be done through verbal and non-verbal methods of transmitting information. Verbal communication represents a way of conveying messages in linguistic form, resorting to the use of spoken, written communication and sign language [9]. Since we are able to innately detect how our interlocutor is feeling through characteristics of their speech, humans can also recur to spoken language as a way to express sentiments, emotions and humor.

1.1 Topic Overview

Psychiatry can be defined as a discipline within the medical field whose primary concern is the understanding and management of mental, emotional, and behavioural phenomena, which are defined as “mental disorders” [10]. One of the groups of disorders treated by this branch of medicine is the affective disorders which are characterized by disturbances in a person’s mood or affect.

Among the multiple types of affective disorders, the most common ones are Major Depressive Disorder (MDD) and Bipolar Disorder (BD). In 2019 the Institute of Health Metrics and Evaluation accounted for approximately 280 million people (around 5% of adults) in the world suffer from depression and around 40 million people experience BD [11]. Both these disorders are characterized by leading to periods of profound unhappiness and a diminished enthusiasm or enjoyment for once gratifying or pleasurable activities [12, p. 123, 155]. There may also be a decrease in vital energy, cognitive, sleep and appetite changes and ideas of guilt or death. During these periods, speech can be characterized by a decrease in volume and output, as well as an increase in response latency time.

Based on data from the Global Burden of Disease 2019 study, mental disorders were among the leading causes of health loss worldwide over the past 30 years in younger age groups (14–49) [13, 14]. One of the most tragic consequences of prolonged untreated depression is Suicidal Ideation (SI) and attempt which lacks specificity as a predictor making it difficult to infer what MDD characteristics lead to suicidal thoughts and ideation [15].

In the case of depression prevalence in Portugal, it is estimated that more than 12% of people over 15 suffer from this specific affective disorder, making Portugal the European country with the highest share of people suffering from it [16, p. 6–7]. It should also be noted that there is a considerable difference in the prevalence of current depressive disorder between different genders since its prevalence is twice as high in women than in men [17].

During a psychiatric consultation, doctors look for specific psychopathological biomarkers which can be obtained through the observation of the patient’s behaviour and speech. One of the most common

symptoms of MDD is Psychomotor Retardation (PMR) which is characterized by a slowing-down of thought and a reduction of physical movements in an individual [18]. Some abnormalities can be detected in a patient's speech, mainly increased pauses, decreased volume, reduced tone and inflection and, delayed response [18].

Speech processing is a way to analyze speech through the use of emergent audio processing technologies. By acquiring, processing and manipulating audio signals one can detect characteristics in the speaker's speech that offer valuable insights into various aspects of human communication and cognition. This enables the extraction of meaningful features such as pitch, tone, rhythm, and linguistic patterns that, nowadays, are applied to different audio technologies such as Automatic Speech Recognition (ASR), Voice Activity Detection (VAD), Speech Emotion Recognition (SER) and Text-to-Speech (TTS) [1, p. 2–3]. These different applications can be found in diverse fields, ranging from linguistics and psychology to human-computer interaction and healthcare.

1.2 Motivation

As technology evolves, the way we, as a species, interact with the world and with each other undergoes profound transformations. These changes can directly affect how certain industries work, improving the way they are structured and organized and how products are manufactured and/or services are provided.

With the increasing use of telemedicine over the past decade and its rapid growth during the SARS-CoV-2 pandemic crisis, non-urgent consultations were recommended to be conducted remotely as a consequence of isolation and quarantine measures [19]. This method of providing care thus became a viable solution.

As presented in the 2023 The Health Policy Partnership (HPP) report [16], even though the mental health services in Portugal have been undergoing significant reform, there is still a lack of resources and professionals to provide the necessary care to the population in need. One of the main challenges is the fact that changes in our health system are slow and the implementation of new technologies is not always a priority. In light of this, it is apparent that presenting a tool that, through the use of telemedicine, can help clinicians in their evaluation of psychiatric patients, is a step in the right direction.

The idea behind the creation of a speech processing tool that could be used to extract speech features in real-time, which is the initial statement of this project, was devised as an add-on tool to an existent telemedicine platform developed by Diogo Ramalho et al [20, 21]. This platform's functionality is to analyze the patient's video and extract, in real-time, information that can help the doctor make an informed prognosis about the patient's state of health. It currently has the capacity to collect various video-based biomarkers such as blink frequency, heart rate and cardiac variability which are obtained by a combination of variations in skin color, head movement and change in fiducial markers on the subject's face [20].

One of the limitations of working with state-of-the-art technology is that these innovations even when applied with the most possible simplicity tend to only be applied to (for-profit) private institutions, so working with a hospital that is integrated into the Portuguese health system is important as it can a good step in applying new technology to services that an everyday citizen could have access. This ensures that they receive the best state-of-the-art care, regardless of their financial situation.

Having this in mind, one can understand that it is important to improve on existent biomedical technologies as a way to provide the best quality of care that those tools can offer, which is why I have faith that the implementation of speech analysis, into an existent platform, as a supporting tool for mental health assessment in psychiatry should help clinicians on their disease evaluation process.

1.3 Objectives

The main aim of this project is to develop a tool that be used as a support mechanism for the assessment of the mental state of patients suffering from affective disorders through the analysis of their voice features and characteristics, in real-time, during a telemedicine consultation.

In order to achieve the aforementioned goal, it is necessary to first identify which are the audio features that should be extracted and processed from the recorded audio signal and how their variation over time can help extrapolate the patient's mental state. Then it is required to comprehend how those features can be coupled and organized together and how can they be displayed in the telemedicine platform so that medical professionals can observe them during the consultation if and when they wish to do so.

1.4 Innovations

In this work, a novel yet straightforward approach to response time calculation based on real-time monitoring of microphone activity between two speakers is introduced. This method is based on the assumption that the time between the end of the first speaker's speech and the beginning of the second speaker's speech is a good indicator of the response time of the second speaker. With this simple application of VAD technologies, one can efficiently calculate response times with minimal computational overhead while maintaining measurement efficiency.

1.5 Methodology

The methodology employed in this report aimed to gather comprehensive and up-to-date information on the topic of speech processing and its application to the medical field by utilizing reputable academic databases and platforms. The primary sources of information included Google Scholar[1], PubMed[2], and ResearchGate[3], which are vast repositories for academic literature and peer-reviewed articles.

1.5.1 Search Method

A systematic search strategy was devised to identify pertinent articles, studies and reviews that could help characterize the state-of-the-art technology presently used in speech analysis. It was mostly done through the use of specific keywords and controlled vocabulary relevant to the topic at hand.

1.5.2 Eligibility criteria

In order to guarantee the significance and reliability of the selected literature, criteria for inclusion and exclusion were established. Only peer-reviewed articles and studies written in English were selected, and an effort was made to favor papers with a more recent publication date.

The exclusion criteria took into account the relevance of each study for this work and the accessibility of the full-text article.

1.6 Report Layout

This Thesis is composed of 6 Chapters and is organized as follows: **Introduction** contemplates the description of the motivation, objectives and methodology used in this report's research and presents a brief overview of the themes at hand. **Background** introduces the reader to the speech-producing mechanism and, the different types of speech processing technologies and which features can be extracted from the audio signal, presenting a both clinical and technical overview of this Thesis's main themes. **State-Of-The-Art** expands on the previous chapter, but focuses on presenting some promising works that have been done in the field of speech processing and its application to the field of psychiatry and presents the importance of selecting a suitable group of features for the development of speech processing tools. **Methods**, presents structure and contents of the tool that was developed. The **Results**' chapter demonstrates, through benchmark and real-world data tests, the efficacy and validity of the developed tool. Finally, **Conclusion** summarizes the main findings of this report and presents some plausible avenues for future work.

[1] https://scholar.google.com/
[2] https://pubmed.ncbi.nlm.nih.gov/
[3] https://www.researchgate.net/

2

Background

Contents

2.1 Clinical Background 8
2.2 Technical Background 15

This chapter presents the background information necessary to understand the context of this thesis. It is divided into two main sections: the first section is associated with the medical background of affective disorders and telemedicine, and the second section is associated with technologies surrounding speech processing.

2.1 Clinical Background

This section starts with the characterization of affective disorders and then introduces the human speech production mechanism followed by a brief presentation of the history of this field and ending with a description of the different speech tasks and features that will be referenced in the following chapters.

2.1.1 Affective Disorders

Affective disorders are a group of psychiatric conditions characterized by disturbances in mood and emotions. Even though mood changes are transversal to most psychiatric disorders such as schizophrenia, anxiety, and personality disorders, only in affective disorders they are the main underlying feature [22]. Two major types of affective disorders can be identified: MDDs and BDs.

The Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5) characterizes MDD as being a condition that leads to discrete episodes with a duration of at least two weeks, in which the individual experiences changes in mood, cognitive function, and physical well-being [12].

BDs are defined by the DSM-5 as conditions in which individuals experience depressive episodes with similar characteristics to those of MDD, but also have episodes of mania or hypomania [12]. A maniac episode represents a period of mood elevation, irritability, and increased energy with a duration of at least one week and where the individual suffers impaired social or occupational functioning [12]. A hypomanic episode is similar to a maniac episode but with a shorter duration and less severe symptoms [12].

Aetiology

The aetiology of affective disorders is complex and multifactorial, being influenced by genetic, environmental, and neurobiological factors and their interactions.

In the case of MDD, it is known that childhood trauma and stress can increase the risk of developing the disorder later in life [23, p. 205]. The genetic component of MDD is also well established, with studies estimating the heritability of MDD at 37% which is a lot lower than the predicted 85% for BD [24].

As for BD, the genetic component is even more pronounced, as confirmed by the above-mentioned heritability estimates. The environmental factors that can influence the development of BD are also var-

ied, with studies showing that childhood trauma and substance abuse can increase the risk of developing the disorder [22, 23].

Those predisposed for MDD and BD tend to have alterations in the levels of neurotransmitters such as serotonin, norepinephrine, and dopamine. The monoamine hypothesis of depression states that the reduction of these neurotransmitters in the brain is the main cause of the disorder [23, p. 209]. It is assumed that those who are predisposed to depression can have it triggered by current life events, which can also happen in the case of BD patients, in which episodes of both depression and mania can be triggered [23].

Diagnosis

In a clinical setting, the diagnosis of affective disorders is based on the patient's medical history, physical examination, and the use of standardized questionnaires and scales. The DSM-5 and the International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10) present the most commonly used diagnostic criteria for affective disorders [12, 25].

The DSM-5 imposes the presence of at least five of the following symptoms, during at least two weeks, for a diagnosis of MDD:

- Depressed mood most of the day, nearly every day;
- Diminished interest or pleasure in all, or almost all, activities;
- Significant weight loss, weight gain, or changes in appetite;
- Insomnia or hypersomnia;
- Psychomotor agitation or retardation;
- Fatigue or loss of energy;
- Feelings of worthlessness or excessive or inappropriate guilt;
- Reduced ability to think or concentrate, or indecisiveness;
- Recurrent thoughts of death and/or suicidal ideation.

The DSM-5 also presents the diagnostic criteria for BD, stating that the presence of a manic episode is necessary for its diagnosis and that three or more of the following symptoms must be present during the period of mood disturbance:

- Inflated self-esteem or grandiosity;
- Decreased need for sleep;

- More talkative than usual or pressure to keep talking;
- Decrease in attention or easily distracted;
- Increase in goal-directed activity or psychomotor agitation;
- Excessive involvement in high-risk activities;

It should be noted that for a conclusive diagnosis, the symptoms should not be attributable to the physiological impacts of any substance/drug, or another medical condition and their impact causes considerable distress or impairment in an individual's normal functioning [12, 22].

Treatment

In the area of psychiatry, the treatment of affective disorders can be done through a combination of pharmacological and psychotherapeutic interventions. Each treatment plan should be tailored to the individual's needs as well as the severity of the disorder.

Psychotherapy treatment and mitigation of the symptoms of affective disorders can be achieved through the use of varied types of therapy. The most commonly used are Cognitive Behavioral Therapy (CBT), Interpersonal Psychotherapy (IPT), and psychodynamic therapy [26].

The pharmacological treatment of affective disorders is based on the use of antidepressants, mood stabilizers, and antipsychotics. The choice of medication is dependent on the type of mood disorder and the severity of the symptoms [22].

In the case of MDD, common treatment options include the use of Tricyclic Antidepressants (TCAs), which inhibit the reuptake of serotonin and norepinephrine, Selective Serotonin Reuptake Inhibitors (SSRIs), which block the reuptake of serotonin and Serotonin-Norepinephrine Reuptake Inhibitors (SNRIs) which inhibit the uptake of both serotonin and norepinephrine. None of these medications are free of side effects, and TCAs are known to also inhibit the reuptake of other neurotransmitters, which can lead to a variety of side effects such as weight gain, sedation and cardiac effect [22, 27]. SSRIs are the most commonly prescribed depression medication and can lead to sexual dysfunction, gastrointestinal symptoms as well as anxiety and insomnia. SNRIs have similar adverse effects to SSRIs but are also known to cause excessive sweating (diaphoresis) and tachycardia [22, 27].

Pharmacological treatments for BD are more complex than those of MDD since they are dependent on the phase of the disorder. For over 70 years, lithium has been the gold standard for the treatment of BD due to its mood stabilizer properties [28]. However, its use has to be monitored closely since it has a narrow therapeutic window and can lead to severe side effects such as tremors, weight gain, and hypothyroidism [22, 28]. Regarding bipolar depression, the use of SNRIs and SSRIs is common, but due to the risk of inducing mania, they are often used in conjunction with mood stabilizers [22]. Some studies

have presented plausible evidence that the use of antipsychotics can be beneficial in the treatment of BD, especially in the manic phase of the disorder [29].

2.1.2 Physiological Process of Speech Production

The Springer Handbook of Speech Processing [1] defines speech as a "wave of air that originates from complex actions of the human body, supported by three functional units: generation of air pressure, regulation of vibration, and control of resonators."

Speech Apparatus

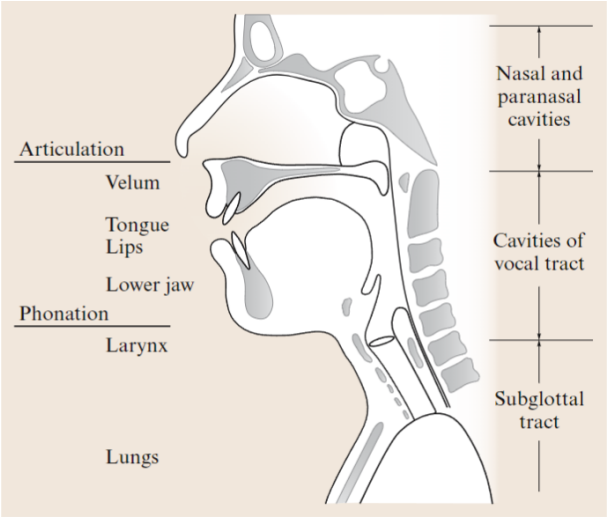


Figure 2.1: Speech production system sketch [1, p.8]

The speech apparatus, as seen in Figure 2.1, is divided into two groups of organs with different functions: the vocal tract which encapsulates organs of phonation responsible for voice production such as the lungs and larynx, and the articulators which are the settings of speech organs and comprehend the lower jaw, tongue, lips, and velum [1, p. 7–8].

The vocal folds, which are housed by the larynx, oscillate during natural speech in order to generate sound. Changes in the cycle-to-cycle variation of the waveform of the produced sound can be measured through the detection of frequency and amplitude perturbations. These perturbations are known as Jitter and Shimmer, respectively, and are used to measure the quality of the speaker's voice [1, p. 11–12] [4].

Such as every other periodic signal, the speech signal can be decomposed into a sum of sinusoidal components. The Fundamental Frequency (F0) of the speech signal is the lowest frequency component of the signal and is also known as the pitch of the voice. The pitch of the voice is determined by the rate of vocal fold vibration and has different ranges depending on the sex of the speaker, ranging from about

80 to 400 Hertz (Hz) in males, and about 120 to 800 Hz in females [1, p. 12].

2.1.3 Speech Biomarkers

A biomarker can be defined as an externally observed medical sign that can be measured with precision and consistency. These differ from medical symptoms which are subjective and can vary from patient to patient since they depend on self-perceived health indications [30].

In the area of psychiatry, the identification of biomarkers is crucial for the development of diagnostic tools and the monitoring of treatment efficacy. During a psychiatric interview, clinicians rely on past examination experience to identify any changes in how a patient acts and behaves concerning external stimuli. This method of diagnosis tends to be subjective and is prone to bias and errors, which is why the use of biomarkers as an objective support tool can be beneficial.

In the preceding work of Diogo et al. [21], to which this project will add, the author identifies some non-speech biomarkers that can be used in the assessment of affective disorders. These include the detection of head movements, blinks, and heart rate variability.

In a 2021 article, Fagherazzi et al. [31], defined voice or speech biomarkers as "a signature, a feature, or a combination of features from the audio signal of the voice that is associated with a clinical outcome". This definition is in line with a speech processing approach to the identification of biomarkers, which is based on the extraction of features from the speech signal and will be further explained in Section 2.2.3.

During a psychiatric evaluation or interview, concerning the speech of the patients, clinicians focus on changes in the content of the speech, the tone of voice, rate of speech, pauses and response time. Some of these features can be directly translated to audio signal features through the use of speech processing techniques, and others such as the content of the speech can be analyzed with the help of machine learning algorithms [31].

2.1.4 Telemedicine

With the advent of the internet and the development of new communication technologies, the field of medicine as a whole has been revolutionized. This led to the development of the field of telehealth, which can be defined as the use of new and emerging technologies to provide better healthcare services to patients in a remote setting [32, p. 11]. In this format, the way a patient receives care is changed, as they can now not only have direct access to their healthcare provider through video calls and messaging services but also have the possibility of monitoring their health status through the use of wearable devices and mobile applications.

Telemedicine is a subset of telehealth focused on the provision of remote clinical services, in which there is interaction between the patient and the healthcare provider [32, p. 11]. This field has been

slowly growing in the past years, as healthcare systems become more digitalized and the need for remote care increases. With the COVID-19 pandemic, and the requirement for social distancing, the field of telemedicine saw a significant increase in demand [19].

Telemedicine in Portugal

Portugal's history with telemedicine dates back to 1998 when the first telemedicine project was launched in the country by the pediatric cardiology service at Centro Hospitalar e Universitário de Coimbra (CHUC) and nowadays works both at the national level and with the Portuguese-speaking African countries (PALOP) [33, 34].

The Portuguese government has been investing in the development of telemedicine services, with the creation of the Portuguese National Centre of Telehealth (CNTS) in 2016 and later, in 2019, the creation of the National Strategic Telehealth Plan (PENTS) whose aim is to promote and develop telemedicine services in the country [35].

In a 2022 study by Caceiro et al. [33], the author presents the results from surveys conducted with 1016 Portuguese physicians, in which around 55% of participants stated that telehealth is a part of their regular practice. The percentage of medical professionals who participated in at least one teleconsultation in the previous six months was higher, at around 72%, with the most common method of communication being the telephone, followed by messaging services and video consultations [33, p. 54–56]. Those who participated in video consultations preferred the Zoom platform for teleconsultations, and those who communicated through messages mainly used email [33, p. 58]. As for the perception of telemedicine, most doctors believe that it is a useful tool for the provision of healthcare services, but that it does not always provide the same level of care as face-to-face consultations [33, p. 61–63]. Finally, regarding limitations to the use of telemedicine, the most common reasons cited were the lack of time to do so or the inadaptation of the clinical software, so the majority believes that there should exist some kind of support team specialized in telemedicine to help with the organization and implementation of these services [33, p. 63].

In the Hospital Beatriz Ângelo of which the psychiatric department is the receptor of the application development presented in this project, the telemedicine services are still in the early stages of development, and even the use of video consultations is not yet a common practice. What is believed is that if a well-structured telemedicine service is implemented, it can help both the patients and the healthcare providers understand the benefits of the use of telemedicine services, increasing its application in this sector.

Telepsychiatry

Telepsychiatry is part of the telemedicine field and is defined by the use of different communication technologies to connect patients with mental health professionals. It has been proven that not only can telepsychiatry be as effective as face-to-face consultations [36], but it can also increase access to mental health services for those who live in remote areas or have mobility issues [37, 38], and provides a cost-effective way of improving how a healthcare system treats those who suffer from affective disorders [39].

As is the case with telemedicine, telepsychiatry can be done through different types of communication technologies. These can be synchronous and associated with real-time communication, such as video consultations, asynchronous where the patient and the healthcare provider communicate through messaging services or email or dependent on telemonitoring services with continuous biomarker information collection through wearable devices or phone apps [40, p. 53]. Luis Gutiérrez-Rojas et al. [40] were also able to identify that the use of telepsychiatry through videoconferencing when applied to patients with affective disorders seems to lead to improvements in depressive symptoms and quality of life, with the added benefit of alleviating the symptoms of those who suffer from anxiety disorders such as agoraphobia.

Current Telepsychiatry Tools and Platforms

With the ubiquity of the use of smartphones and the internet, in the last couple of years, many projects have been developed to provide telemedicine services to patients and healthcare providers alike. One of those projects is **MindLogger**[1] which was responsible for the development of mobile and web applications in which patients can be asked to fill out questionnaires and surveys, monitor their mood and sleep patterns, and even participate in behavioural tests [41]. Another project is the **Moodtracker**[2], which lets users track varied measurements from mood and sleep patterns to water intake.

In a 2017 study [42], in which the authors presented a preliminary design of an application of sentiment analysis and affective computing for depression monitoring, it was demonstrated that it is possible to monitor depression through audio and video analysis. This system is expected to process the biomarker data collected through the use of wearable devices, mobile applications, and social media data, to provide a more comprehensive view of the patient's mental health status. As the aforementioned projects, this application was developed around an asynchronous communication system, in which patient data is collected over time and then presented for analysis to the healthcare provider through a Graphical User Interface (GUI) [42].

When comparing the project concepts presented in the previous paragraph with the work developed in this thesis, it is possible to see that there are similarities in the way the data is collected and presented

[1] https://mindlogger.org/
[2] https://www.moodtracker.com/

to the healthcare provider. However, the main difference lies in the fact that the application developed in this project is focused on the real-time analysis of the speech signal, which can be used to monitor the patient's mental health status during a teleconsultation, with no need for the patient to interact with the application since the data collection and analysis are done all through the clinician's side.

2.2 Technical Background

This section presents the background of the underlying concepts around speech processing, with a brief presentation of the history of this field, followed by a description of the different speech tasks and features that will be referenced in the following chapters and ending with an overview of the state-of-the-art machine learning applications for speech analysis.

2.2.1 Brief History of Speech Processing

As previously referenced in Section 1.1 speech processing can be defined as a group of emerging tools and techniques that, through the acquisition, processing and manipulation of audio signals support the detection of different characteristics in a speaker's speech that offer valuable insights into various aspects of human communication and cognition.

The history of speech processing can be dated back to the 13th century when there was undocumented evidence of the utilization of mechanical models to simulate the human vocal apparatus. Pioneering endeavors emerged in the 18th century, exemplified by Kratzenstein's resonant cavities and von Kempelen's speech synthesizer, which marked the commencement of speech processing. In the latter half of the 19th century, Alexander Graham Bell, renowned for his invention of the telephone, also engineered a device capable of articulating speech [1, p. 1–2].

Subsequently, in the 20th century, the focus shifted from mechanical to electrical devices and, with Homer Dudley's work in the 1930s, the concept of using carriers for speech signals was introduced. Dudley's electrical speech synthesizer, the Voder and later the Vocoder, demonstrated significant progress in speech synthesis and compression. Despite applications in military and secure voice systems during World War II, analog implementations of this technology faced challenges in commercial telephony due to quality issues. Digital hardware availability in the 1970s paved the way for significant advancements in speech coding, synthesis, and recognition [1, p. 1–2].

In the present time, speech processing continues to thrive as a field of innovation and integration, increasingly dependent on cutting-edge technologies such as Artificial Intelligence (AI). AI, particularly through machine learning and deep learning algorithms, has become instrumental in refining speech recognition and synthesis systems, which will be vital technologies in the future of human-computer interaction [1, p. 1–2] [43].

2.2.2 Speech Processing Tasks

Among the varied applications of speech processing, some are more relevant to this project than others. In the following subsections, the most relevant applications will be presented and explained.

Automatic Speech Recognition (ASR)

ASR is a technology that enables the recognition and translation of spoken language into text or commands [43, p. 40–41]. This technology is used in a wide range of applications, from voice search on mobile devices to voice commands in smart home devices. ASR systems are composed of three main components: audio feature extraction, an acoustic model and, a language model. While the acoustic model is responsible for converting the acoustic signal into a sequence of phonemes, the language model is used to determine the probability of a sequence of words occurring in a given language and then their outcomes are merged in order to produce the transcription of spoken words [43, p. 40–41].

Its applications in the medical field are also varied, ranging from helping people with speech impairments to aiding doctors with clinical note-taking.

Voice Activity Detection (VAD)

VAD is a process in which the presence or absence of human speech is detected in an audio signal. Nowadays, its use is extremely common in the development of ASR systems such as home devices, as it is used to determine the start and end of speech segments [43, p. 65]. VAD is also important in the distinction of which parts of the audio signal correspond to background or ambient noise and which parts correspond to speech, making sure that the systems this technology is integrated in only process the speech segments of the audio signal. Systems of this type tend to be used in conjunction with other more complex speech processing tasks, as the separation of speech from noise and artifacts tends to be an extremely important process in the pre-processing section of speech analysis applications.

Speech Emotion Recognition (SER)

SER is a domain of research associated with the task of Speech Recognition (SR) that aims to characterize or classify the emotional state of a speaker through the analysis of the speech signal [2]. Such like ASR, SER systems are composed of multiple steps which can be divided into signal pre-processing, feature extraction, feature selection, and classification. The signal pre-processing step is responsible for removing noise and silence sections from the audio signal, the feature extraction step assures that the most relevant information is extracted from the audio signal, the feature selection step is responsible for selecting the most relevant features, recurring to feature optimization techniques which will be dependent on the classification algorithm used in the classification step [2].

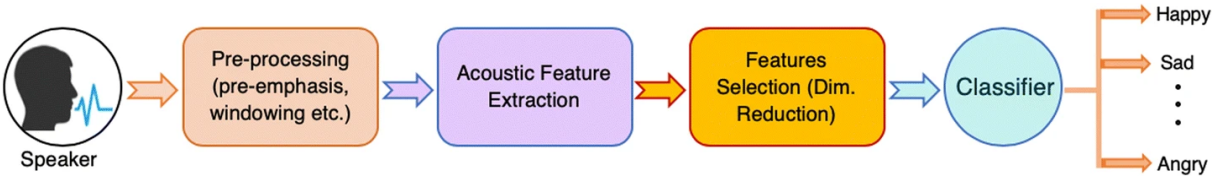


Figure 2.2: Pipeline from a classical SER application [2, p. 2]

Speaker Diarization (SD)

SD is a speech processing task used in situations where the audio data one wants to analyze or process has multi-speaker information [43, p. 56]. Using this technique it is possible to identify when each speaker is speaking, which is necessary in situations where there is a need to selectively analyze the speech of a specific speaker. SD systems are composed of three main steps: speaker segmentation, speaker embedding and, embedding clustering. The segmentation step is responsible for dividing the multi-speaker audio signal into smaller segments based on periods of silence between speakers (which might be dependent on VAD systems), the speaker embedding extraction step is responsible for obtaining a low-dimensional representation of each speaker's voice in a compact and meaningful way and, the clustering step uses clustering algorithms to group embeddings based on the proximity of its audio features [43, p. 57].

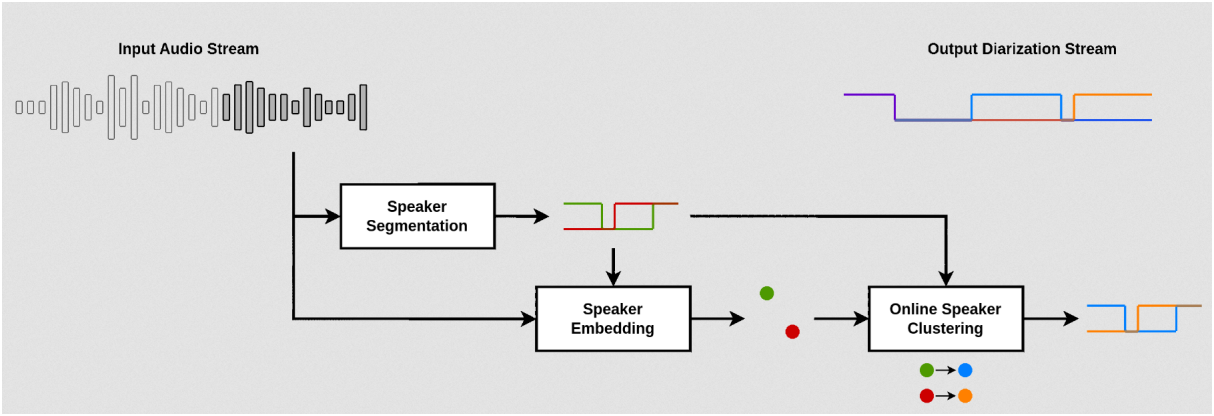


Figure 2.3: Pipeline from a real-time diarization application [3]

2.2.3 Speech Features

Speech features are representations of speech signals that are used in speech processing tasks. These features can be divided into two main categories: linguistic and acoustic features [44, p. 2–3].

Linguistic Features

This group of features encapsulates what is being said through the analysis of how the speaker uses certain words and grammar in order to formulate sentences. They have an increased importance when applied to speech processing tasks in the medical field since they can be used to detect abnormalities in the speech of a patient, such as the use of inappropriate words or the lack of coherence in the sentences they formulate [44, p. 3].

Most of the linguistic features are extracted from the text transcription of the speech signal, which is obtained through ASR applications, leading to complex and computationally expensive systems that are subject and context-dependent. This limits the identification of clearly defined and generalizable linguistic features.

One linguistic feature that can be characterized through the analysis of the syntax of the sentences uttered by the speaker is **valence**, which can be defined as the objects or concepts that attract or repel individuals based on their "(un)pleasantness, goal obstructiveness/conduciveness, low or high power, self-(in)congruence, and moral badness/goodness" – as defined by the authors of article Levels of Valence [45]. Having this in mind, one can use it to represent positive or negative emotional states of the speaker.

Acoustic Features

These types of features tend to relate to the physical properties of the speech signal and are the most commonly used in speech processing tasks and can be divided into two different categories: time-domain and frequency-domain features.

Time-domain Features

This group of features is obtained by deriving the amplitude of the audio signal over time. They are simple to compute and often used in real-time speech processing applications.

Energy is the temporal variation of the intensity of the speech signal and represents the loudness of the voice of the speaker [4, 43]. It can be computed using the following formula:

$$E_i = \sum_{k=i-N+1}^{i} [x(i)w(i-k)]^2, \tag{2.1}$$

where E_i is the energy value at sample i, $w()$ is the window function and N is the length of the window frame [4, p. 175].

Zero-crossing rate is the number of times the speech signal crosses the zero-axis within a defined time frame and, can be computed by counting the number of polarity changes in the signal during that window [4, 43]. It characterizes the smoothness of the signal and helps distinguish between voiced and unvoiced speech since the latter tends to have a significantly higher zero-crossing rate [5]. One implementation of the zero-crossing rate is given by the following formula:

$$ZCR_k = \sum_{h=kM}^{kM+N} |sign(x_h) - sign(x_{h-1})|, \tag{2.2}$$

where ZCR_k is the zero-crossing rate of the k-th window, M is the step between windows and, N is the window length [5].

One can also calculate the autocorrelation of the signal at $lag - k$ and position n in the sample x, using the formula:

$$r_k = \frac{1}{N-1} \sum_{k=1}^{N-1} x_n x_{n-k} \tag{2.3}$$

As presented in Section 2.1.2 the **Pitch** or **Fundamental Frequency (F0)** is a speech feature that relates to the perceived tonality of the voice of the speaker [43].

From the speech signal, one can also extract the **Jitter** and **Shimmer** features. **Jitter** is a perturbation in the audio signal that is characterized by cycle-to-cycle pitch variation [4,5] and can be computed using the following formula:

$$J = \frac{1}{N-1} \sum_{k=1}^{N-1} |T_k - T_{k+1}|, \tag{2.4}$$

where T_k is the time length of each **F0** period and N the number of periods extracted [4, 5].

Shimmer represents the amplitude variation of the signal and can help identify energy perturbations in the speech wave [4, 5]. It can be obtained using the following formula:

$$S = \frac{1}{N-1} \sum_{k=1}^{N-1} \left| 20 \log \left(\frac{A_{k+1}}{A_k} \right) \right|, \tag{2.5}$$

where A_k is the amplitude of the k-th period and N the number of periods extracted [4, 5].

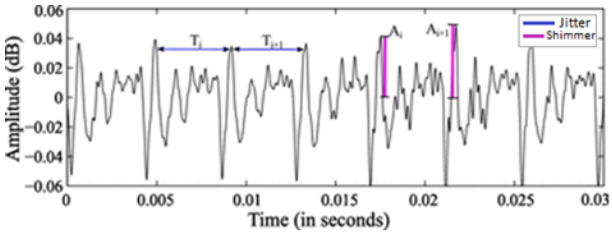


Figure 2.4: Graphical representation of shimmer and jitter [4, p. 176]

The above-presented formulas for jitter and shimmer calculations are one of the many ways to compute these features, and the choice of the formula is dependent on the application and the desired outcome of the analysis. The demonstrated formulas can be characterized as local absolute values of the perturbations and are the easiest to comprehend given the fact that the obtained output is presented in commonly used audio analysis units.

Formants correspond to the different frequencies of resonance of the vocal tract with a high degree of energy. They tend to be associated with the vowels of the voiced signal and can be used to characterize the speaker's emotions since vocal tract characteristics change with different emotion expression [1, p. 18] [4]. **Formants** are identified through the detection of energy peaks in the signal, and numbered in order of increasing frequency [5].

As it can be seen in Figure 2.5 as the frequency increases the harder it is to identify **Formants**, which is why the first three **Formants** are the most commonly used in speech processing tasks [1, 5].

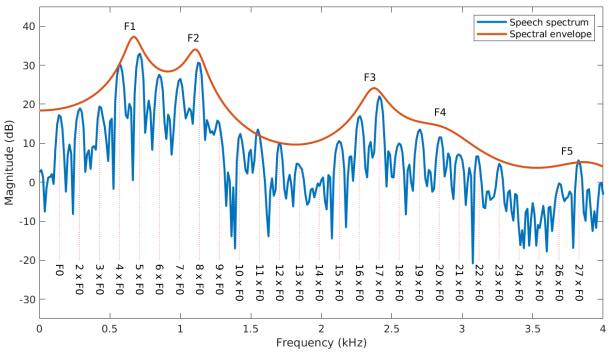


Figure 2.5: Spectrum of a speech segment annotated with its formants [5]

Harmonics-to-Noise Ratio (HNR) which, as the name indicates, quantifies the ratio between the periodic (harmonic part) and aperiodic (noise) component of the audio signal [46]. In some contexts can be preferred over **Jitter**, as it can be measured without the need for a pitch tracker [46]. The **HNR** (in dB) can be computed using the following formulas:

$$r_x(\tau) = \int x(t)x(t+\tau)dt, \tag{2.6}$$

$$HNR = 10.\log_{10}\frac{r'_x(\tau_{max})}{1 - r'_x(\tau_{max})}, \tag{2.7}$$

where $r_x(\tau)$ is the autocorrelation as a function of the lag τ, and $r'_x(\tau_{max})$ is the local maximum value of the normalized autocorrelation [47].

The following group of features cannot be computed directly from the audio signal and requires the extraction of specific information from the voiced segments of the signal in order to be obtained. They are associated with timing measurements related to suprasegmental characteristics of the speech signal which are defined as **prosody** [48].

Speech rate is the number of syllables spoken per time interval and can be used to characterize the speed of the speaker's speech [49, p. 7]. It can be computed using the following formula:

$$\text{Speech Rate} = \frac{\text{Number of Syllables}}{\text{Total Time}} \tag{2.8}$$

Articulation rate is the number of syllables spoken per phonation time interval which characterizes the duration of voiced speech segments [49, p. 7]. It can be computed using the following formula:

$$\text{Articulation Rate} = \frac{\text{Number of Syllables}}{\text{Phonation Time}} \tag{2.9}$$

Average syllable duration is the average time duration of each syllable vocalized by the speaker and can be used to characterize the rhythm of the speaker's speech [49, p. 7]. To obtain this feature one can use the following formula:

$$\text{Average Syllable Duration} = \frac{\text{Phonation Time}}{\text{Number of Syllables}} \tag{2.10}$$

Such as **Phonation Time** one can also compute **Pause Time** and **Pause Number**, but in order to do so, there is a need for segmentation of the audio signal into voiced and unvoiced segments and then calculation the duration of the pauses between the voiced segments. The most common way to segment the audio signal is through the use of VAD or ASR systems as presented in the previous section.

Frequency-domain Features

This group of features is obtained by deriving the frequency spectrum of the audio signal. Spectrograms constitute the graphical representation of the spectrum of a signal over time and are more complex to compute than time-domain features since they tend to recur to time-frequency transform operations such as Fourier Transform (FT)[3] [5, 43].

[3] For the mathematical formula of the Fourier Transform, refer to Appendix A, Section A.1

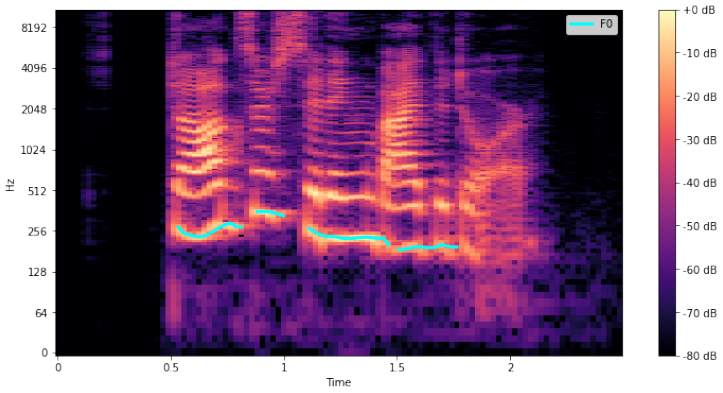


Figure 2.6: Spectrogram of a speech signal with F0 annotated

Mel-frequency spectrogram is a representation of the short-term power spectrum of a sound signal. It is obtained by applying a Fast Fourier Transform (FFT)[4] to short overlapping frames of the audio signal and then convert the power values at different frequencies to their corresponding mel-frequency bands [43]. These mel-frequency bands are part of the mel-scale which maps the frequency of a signal to the pitch that is perceived through the human auditory system [5, 43], as it can be observed in Figure 2.7.

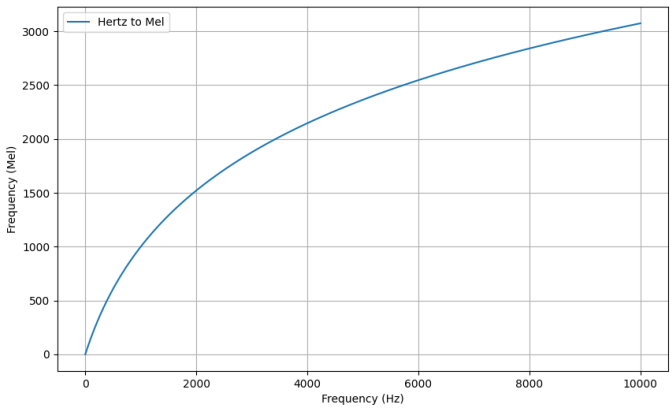


Figure 2.7: Graph of mel-scale values vs Hertz scale values

The formula used to convert the frequency of a signal to its corresponding mel-frequency is given by:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.11}$$

[4]For the mathematical formula of the Fast Fourier Transform, refer to Appendix A, Section A.1.2

Mel-Frequency Cepstral Coefficients (MFCCs) are a set of coefficients that result from the application of a Discrete Cosine Transform (DCT) over the log-mel spectrum of the audio signal [5]. These types of coefficients are one of the most commonly used features in speech processing tasks since they are able to capture the characteristics of the human voice, by ensuring the independence of the extracted features, and are robust to noise and speech variations [43].

2.2.4 Overview of Speech-Based ML Technologies

Although slightly out of the scope of the principal objectives envisioned for this project, the following section delves into notable applications within the field. The purpose of this section is to provide a brief overview of the current state-of-the-art applications of machine learning in the field of speech processing, with a particular focus on the detection of affective disorders.

One of the most common applications of speech processing in psychiatry is its use in the detection of anxiety, stress and different depression levels through audio feature analysis in **Speech Recognition (SR)** tasks. These works recur to different machine learning architectures to build classifiers with the purpose of predicting the presence of these conditions based on the audio data collected from the patients.

The more traditional speech learning algorithms encapsulate **Gaussian Mixture Models (GMMs)**, **Support Vector Machines (SVMs)** and **Hidden Markov Models (HMMs)** [43, p. 7–8]. **GMMs** combine multiple Gaussian distributions with varied weights over the probability distribution of a speech feature vector, **SVMs** are mainly used for classification tasks as they are able to separate data into different groups by finding the optimal hyperplane that separates different data classes and **HMMs** are used to model the probability distribution of a sequence of speech features based on the arrangement of hidden states accompanied by the respective observation [43, p. 7–8] [5].

In the last decade, this type of algorithms has been replaced by more modern machine learning architectures such as **Deep Neural Networks (DNNs)**, **Convolutional Neural Networks (CNNs)** and **Recurrent Neural Networks (RNNs)**. These deep learning methods represent an evolution over the legacy **Artificial Neural Networks (ANNs)** where a group of connected nodes is organized in a structure resembling the human brain. The main difference between **ANNs** and **DNNs** is that the latter have multiple hidden layers between the input and output layers, which allows them to learn more complex representations of the training data [6].

RNNs are a **Deep Learning (DL)** architecture that is used to process varying sequential patterns [43, p. 9]. They differ from other types of **ANNs** such as **Feed-Forward Networks** as they have a feedback loop that allows them to process sequential data by taking into account the previous inputs [6]. A standard **RNN** has structure as observed in Figure 2.8 and is represented by the following equations:

$$h_t = H(W_{hh}h_{t-1} + W_{xh}x_t + b_h), \tag{2.12}$$

$$y_t = W_{hy}h_t + b_y, \tag{2.13}$$

where W_{hh}, W_{hx} and W_{hy} are the weight matrices, b_h and b_y are the bias vectors, h_t is the hidden state at time t, x_t is the input at time t and y_t is the output at time t. $H()$ is the activation function, which is usually a non-linear function such as the hyperbolic tangent or the sigmoid function [43, p. 9] [6]. Their most common application is in speech recognition tasks with a focus on the prediction of phonetic segments from audio signals.

As a consequence of having vanishing gradients, **RNNs** are not able to learn long-term dependencies, which means that they are not able to learn from inputs that are far apart in the sequence. This problem was solved through the creation of **Long Short-Term Memory (LSTMs)**[5] which made it possible for the network to keep (or disregard) information over prolonged durations [43, p. 10]. Recently LSTM networks have been utilized in speech post-filtering in order to improve the quality of synthesized speech in applications that depend on it, such as TTS [43, p. 11].

CNNs are a type of **DL** architecture that is used to process data that assumes a grid-like topology, and were first created for image processing tasks [6]. They are composed of one or more sets of convolutional and pooling layers arranged alternately. Convolutional layers apply varied filters over local parts of the input, computing feature maps, which then are replicated over the entire data. The convolutional operation is represented by the equation $(h_k)_{ij} = (W_k \otimes q) + b_k$, where $(hk)_{ij}$ is the $(i,j)^{th}$ element for the kth output feature map, q represents the input feature maps, and W_k and b_k denote the k^{th} filter and bias, respectively [6]. The $\otimes$ operator represents the 2D convolution operation. Pooling layers are used to reduce the dimensionality of the data by applying a function over local parts of the input while maintaining the most important features [43, p. 11] [6].

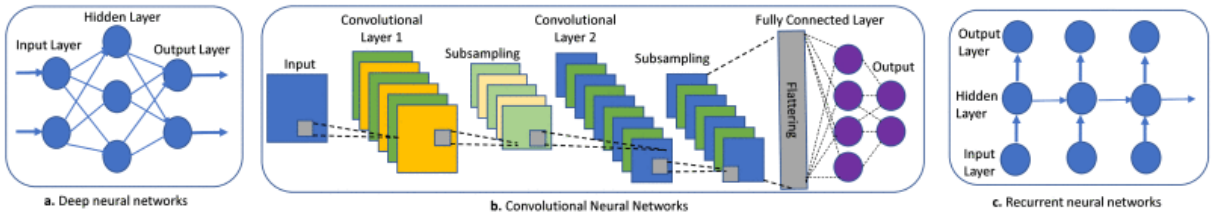


Figure 2.8: Graphical illustration of different Machine Learning (ML) models [6]

[5] For the mathematical formula of the Long Short-Term Memory RNN, refer to Appendix A, Section A.2

In the last couple of years, a technology that has completely revolutionized the field of Natural Language Processing (NLP) and speech processing is the **Transformer** architecture. It is characterized by presenting a specialized learning mechanism, named the attention mechanism, which coupled with parallel sequence analysis instead of sequential sequence analysis improves efficiency when compared with RNN architectures [50]. In the 2017 research paper "Attention Is All You Need", in which the transformer architecture was introduced, the attention function is defined as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors and the output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key [51].

The transformer model is composed of an encoder and a decoder, each comprised of a stack of blocks that consist of two sub-layers: a **Multi-Head Attention (MHA)** mechanism and a position-wise **fully connected feedforward network** [51].

In the encoder block the MHA mechanism allows it to simultaneously consider different parts of the input sequence, capturing intricate patterns and dependencies efficiently and in the decoder blocks MHA together with the encoder-decoder attention layer allows the decoder to focus on the relevant parts of the input sequence [51] [43, p. 16-18].

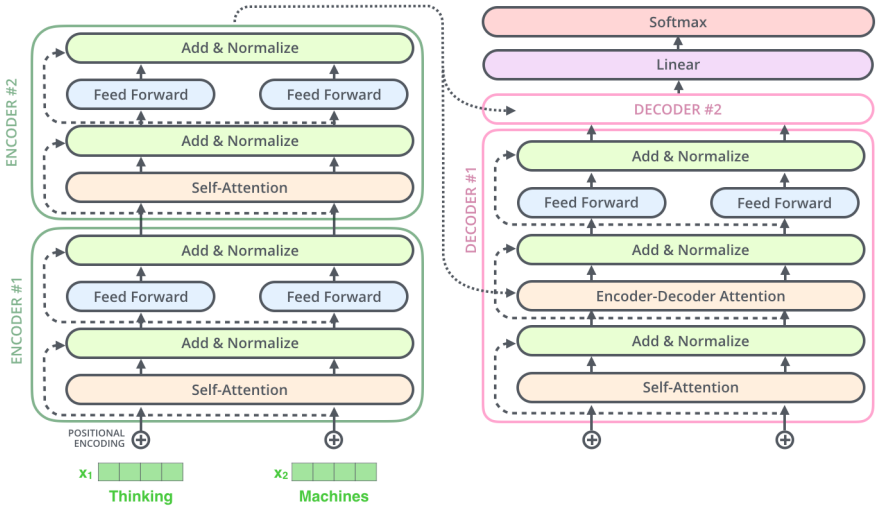


Figure 2.9: Transformer of two stacked encoders and decoders by Jay Alammar [7]

Even though **transformers** have seen multiple applications in different speech processing tasks ranging from ASR to TTS, they are not always the best option for the task at hand. This relates to the fact that most of the most complex Deep Learning (DL) architectures are computationally expensive and require large amounts of data to be trained. This is a problem when dealing with speech data coming from medical field applications where the access to data is limited and the computational resources are scarce.

3

State-Of-The-Art

Contents

3.1 Literature Review 28

3.2 Feature Selection 30

3.3 Conclusions 33

As a consequence of the specific requirements of the topics addressed in this thesis project, the state-of-the-art section for this particular topic unfolds a mosaic of distinct technologies, each individually addressing various aspects pertinent to the subject matter. Unlike conventional scenarios where existing papers contribute to a cohesive body of knowledge, the absence of analogous speech processing tools, to the one proposed in this report, leads to the need for an exploration of topics in a more fragmented manner. The following sections will thus be dedicated to the presentation of how the most relevant technologies that will be used in the development of the proposed system are being implemented in the present day.

3.1 Literature Review

Recently there have been multiple studies conducted in order to understand how speech analysis can be used as a tool for the diagnosis of affective disorders. Coupling this with the evolution in speech processing technologies, in this section, there will be presented some of the most relevant studies that have been conducted in this field in order to identify which are the most commonly used speech features in the literature.

Table 3.1: Research articles found on speech processing for affective disorders, with the respective year of publication, features extracted, evaluation method and computational approaches.

Article	Year	Speech Features	Evaluation Method	Computational Approaches
Yamamoto et al. [52]	2020	Speech Rate, Pause Time and Response Time	Correlation of feature prevalence with HAMD-17 scores	SR technology
Arevian et al. [53]	2020	Linguistic Features, Formants, Pitch, Harmonicity and Pause Time	Correlation of feature prevalence with BASIS-24 and SF-12 scores	ML models based on SVM architecture
Solomon et al. [49]	2015	Pitch, MFCCs, Acoustic Features, Speech rate, Phonation time, Articulation rate and Average syllable duration	Accuracy of the detection of concealed depression using ML models	ML models such as Naive Bayes, kNN, Random Forest and Neural Networks
Seifpanahi et al. [54]	2023	Pitch, Jitter, Shimmer, CPP, HNR, Speed of Speech and Pause Duration	Pearson correlation coefficients between the HAMD scores and speech features	Pearson Statistical Test
Silva et al. [55]	2021	Pitch, Jitter, Shimmer, CPPS and GNE	Association between speech features and BDI-II scores	Multiple Linear Regression model

Continued on next page

Table 3.1: Research articles found on speech processing for affective disorders, with the respective year of publication, features extracted, evaluation method and computational approaches (Continued).

Article	Year	Speech Features	Evaluation Method	Computational Aproaches
Quatieri et al. [56]	2012	Jitter, Shimmer, Pitch, Energy and Degree of Aspiration	Correlation of features with HAMD and QIDS scores	Spearman Correlation
Kiss et al. [57]	2017	Formants, Pitch, Energy, Ratio of Pauses and Transients and Articulation Rate	Significance Test of the correlation between speech features and BDI-II scores	ML models based on SVM architecture
Morales et al. [58]	2016	Pitch, Voicing Probability, Loudness Contours and Speech rate Features	Comparison of MAE and RMSE between text and prosodic features and BDI-II scores	Regression Analysis and SVM models
Hashim et al. [59]	2017	MFCC, Bands of equal PSD, Voice transitions parameters, Probability Functions	Average MAE between predicted and actual HAMD and BDI-II scores	Multiple Linear Regression models

Table 3.1 presents a summary of the most relevant articles found in the literature that address the topic of speech processing for affective disorders. One limitation found while doing this research was the fact that most of the articles found were focused on the analysis of speech features extracted from speech samples recorded in controlled environments with professional audio recording setups, such as the ones found in the SAVEE [60] and RAVDESS [61] databases. This differs from what is intended to be developed in this thesis, which is a tool that can be used in a real-world scenario, where the audio is recorded in a non-controlled environment with conversational speech patterns.

It should also be noted that the objective of the majority of the articles found had a final goal of diagnosing the presence of affective disorders in patients, through the use of machine learning algorithms, having the extraction of speech features as an initial step into creating a data vector that then could be used to train the models for disease classification/identification. This is not the objective of this thesis in which the focus on feature extraction is a lot more pronounced, since it is intended to present the speech data in a way that is easy to understand and analyze in real time.

3.2 Feature Selection

Given the variety of features one can extract from speech data and the fact that not all of them can be applied to the specific problem subjacent to the present thesis, the feature selection for the proposed system was based on the literature review presented in the previous section. The features selected were the ones that were most commonly used in the articles found and that were also the most relevant for the assessment of affective disorders. From that research the following were identified and most important: Pitch, HNR, Jitter, Shimmer, Formants and certain timing measurements such as articulation time, speech rate and pause number.

Even though one of the most used speech features in the literature are the Mel-Frequency Cepstral Coefficients (MFCCs), they were not selected to be studied in this thesis. The basis for their exclusion is that MFCCs are a group of features that not only tend to be computationally expensive to calculate but also are not very intuitive to understand. The most common use of MFCCs is in the field of Speech Recognition (SR), where the features are used to train models to recognize speech patterns, which even though MFCCs are very good at storing the information of the speech signal, they are not very good at representing the information in a way that is easy to understand for someone without deep understanding of speech processing since they tend to have high dimensionality and be abstract in nature.

One other point that was considered not only in the selection of the features but also in the development of the overall system is the fact that the features selected should be able to be calculated in real time. This is important because the system is intended to be used in a real-world scenario, where the audio is recorded in a non-controlled environment with conversational speech patterns. This system concept is different from the ones present in the state-of-the-art where after the audio is recorded it is then processed and segmented into voiced and unvoiced regions, which are then used to calculate the features. The added step of processing the audio before calculating the features is not feasible in a real-world scenario, where the audio is recorded in real time and the features need to be calculated "on the fly", which might lead to higher variability in the feature values calculated.

One path of application that could be followed to bridge the aforementioned gap between this work and the state-of-the-art is the one presented by Quatieri et al. [56] in which speech feature values analysis was done on an audio dataset composed of 20 women and 15 men, that had just started MDD treatment, collected by Mundt et al. [62]. The authors calculated acoustic features over specific sections of speech audio representing vowels (/a/, /e/, /i/, /o/) and compared the results from the first day of treatment with the results from the last day of treatment. Having this in mind one could store feature values over time and give medical professionals the tools to compare how they change from session to session, helping them understand how the patient is evolving over time.

3.2.1 Feature Value Thresholds

Given that the objective of extracting the audio features from the patient's speech is to present them to medical professionals, one other focus of the literature review was to understand if any disease thresholds could be used to classify the feature values into normal and abnormal for someone suffering from a specific affective disorder. The research done in this area was, mainly focused on MDD cases and seems to be inconclusive, with some studies presenting thresholds for some features and others not. When a threshold is presented, as is the case for the study done by Seifpanahi et al. [54], the obtained values seem to be specific to the procedure followed in the study, making the obtained feature data dependent on the severity of the disease of the experimental group and the context and content of the speech data recorded, which makes it difficult to generalize the results. This characterizes a gap in the literature that could be addressed in future studies, in order to provide a more robust and generalizable set of thresholds that could standardize the analysis of the feature values extracted from the patient's speech.

Even though there is a limitation in creating interpatient thresholds for the feature values, one could still use the feature values extracted from the patient's speech to create a baseline for the patient, which could then be used to compare the feature values extracted from the patient's speech in future sessions. So following the example mentioned in the previous section (presented by Quatieri et al. [56]) one could store data from the patient's speech in each session and analyze feature values over time, in order to understand how the patient is evolving. Using an intrapatient approach will also help to mitigate the variability in the feature values extracted from the patient's speech since the feature values will be compared to the baseline of the patient and not to a general threshold that might not be applicable to every case given the complexity of the human voice.

In the search for threshold values, it was found that even though there is no consensus in the literature when it comes to the analysis of speech data from those who suffer from affective disorders, there are studies that try to characterize how the human voice is affected as a consequence of a varied set of pathologies by proposing thresholds for the features extracted from the speech signal of individuals who don't suffer from any speech disturbance or pathology. In the book "A Ciência e a Arte da Voz Humana" [8] the author gathers information on the "normative" values for some of the speech features that are being studied, such as the pitch, jitter and shimmer. Even though these values cannot be used directly in the assessment of affective disorders, they can be used as a reference to verify the quality of the feature extraction process and how well the created system is able to analyze speech data. Even though there was an attempt to standardize feature measurement by calculating it over audio segments of specific vowel phonation, results seem to vary from study to study.

Table 3.2: Normative feature measurements presented in Isabel Guimarães' book [8], organized by study author, age range and gender, vowel phonation analyzed and features extracted.

Article	Age Range & Gender	/u/	/i/	/a/	Feature
Peppard et al. [63]	(16-30) Women	_____	0.036±0.015	_____	Absolute jitter magnitude (msec)
Orlikoff et al. [64]	(26-38) Men	0.046±0.019	0.051±0.020	_____	Absolute jitter magnitude (msec)
Orlikoff et al. [64]	(21-31) Women	0.283±0.062	0.295±0.055	_____	Acoustic Shimmer (dB)
Orlikoff et al. [64]	(26-38) Men	0.361±0.129	0.283±0.062	_____	Acoustic Shimmer (dB)
Sorensen & Horii [65]	Adult Women	204.6	205.5	198.8	F0 (Hz)
Sorensen & Horii [65]	Adult Men	123.2	125.6	110.9	F0 (Hz)

In Table 3.2 there is a select presentation of some of the normative feature measurements presented in Isabel Guimarães' book [8]. Observing the experimental data presented leads to similar conclusions as the ones presented in the background Chapter of this thesis (Section 2.1.2) where is stated that the extraction of speech features is dependent on the sex of the speaker and the phonation of the speech. In addition to this, this data also shows that there are significant differences in speech characteristics as the speaker ages, which is also a factor that should be taken into account when analyzing the feature values extracted from an individual's speech.

Fundamental frequency, or pitch, is not only present in the literature with values for vowel phonation but also with values for speech during conversational and reading tasks. For conversational speaking in Portuguese Guimarães & Abberton [66] obtained a mean pitch value of around 190 Hz for adult women and around 113 Hz for adult men (both with ages comprised between 19 and 40 years old). In the case of HNR, its values range from 7 to 17 dB with a mean value of 11.9 dB [8, p. 183].

3.2.2 Feature Behavior in Affective Disorders

Even though the literature doesn't present a consensus on the thresholds for the feature values extracted from the speech of those who suffer from affective disorders, there are clear patterns that can be identified in the behavior of these features when analyzed in the context of affective disorders. This qualitative analysis of quantitative feature data can be used to assess changes in the patient's speech that might be indicative of degradation or improvement of a person's mood state over time.

In the case of timing-related measurements (prosody) such as speech rate and pause frequency,

their behavior is already well documented in the literature since they are already tracked in the context of a psychiatric evaluation. As presented in Section 1.1 pause number and duration tend to increase in patients with MDD [18]. This is also the case for speech rate, which also tends to decrease in patients with this condition [18].

As for the other features, some studies present identifiable patterns in the behavior of the feature values extracted from the speech of those who suffer from affective disorders. In a 2021 study by Albuquerque et al. [67] the authors investigated the acoustic effects of both depression and anxiety in the speech of patients aged between 35 and 97 years old. They presented a justification for the reduction in F0 parameters being due to the impact Psychomotor Retardation (PMR) has in the larynx's muscle tension, which in turn affects the vocal folds and the pitch of the voice. This tightening of the vocal tract also contributes to the reduction of its acoustic resonance, which leads to a decrease in the different formant frequencies [67, p. 4].

When analyzing changes in jitter, shimmer values different studies support the idea that these features tend to increase in patients with depression independently of sex [68, 69]. These changes seem to also be a consequence of loss of muscle control precision and decreased laryngeal muscle tension which are characteristic of PMR in depression. These physical changes tend to be perceived as "breathy, rough, or hoarse voices" depending on the severity of the disease [70].

As for the HNR, the literature presents varied data, with studies presenting experimental results for the increase, decrease, and no change in the feature values for patients with depression [67]. But, if one considers the effect of PMR in the larynx's muscle tension, it is possible to understand that the HNR values might to decrease in patients with depression, as a consequence of the increased noise in the speech signal imposed by a "more open and turbulent glottis" [56, p. 3].

3.3 Conclusions

The literature review presented in this chapter shows that even though there is a lot of research being done in the field of speech processing, there is still a gap in the literature when it comes to the analysis of speech data from those who suffer from affective disorders. This gap is mainly due to the fact that nowadays research is being geared towards the development of machine learning models that can diagnose the presence of affective disorders in patients, which makes the extraction of speech features a secondary step in the process of developing these algorithms. This leads to the absence of studies that focus solely on the extraction of speech features and how their quantitative analysis can be used to assess the presence of affective disorders in patients.

Despite this gap, it was possible to identify the most relevant speech features that are being used in the literature for the assessment of affective disorders. Given the presented limitations in the identifica-

tion of thresholds for the feature values it was possible to compile a set of normative values for some of the speech features that are being studied, and the behavior of these features in the context of affective disorders.

4

Methods

Contents

4.1 Audio Processing Pipeline 36

4.2 Python GUI 40

In this chapter, the methods used to develop the software and obtain speech features from audio data are described. The overall software development process was divided into two main stages: the development of an audio processing pipeline and then its integration with the pre-existing telemedicine platform. Both stages were built using the Python programming language, which is widely used in the scientific community for data analysis and machine learning tasks. The design of these stages took into account the previous extensive literature research conducted in Chapter 3.

4.1 Audio Processing Pipeline

The crux of the audio processing pipeline is to extract features from speech audio signals and then display them in a user-friendly interface. In order to achieve this goal the audio stream data was captured in real time using the PyAudio library [71] and the features were calculated using the Parselmouth library [72].

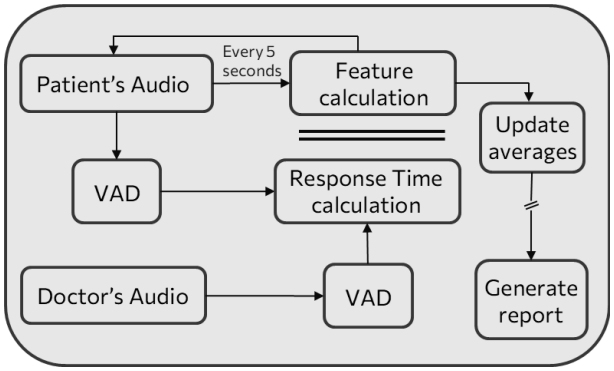


Figure 4.1: Representation of the Audio Processing Pipeline.

The overall structure of the audio processing pipeline can be identified in Figure 4.1. The audio data is captured in real time and then processed in order to extract the selected features and display them.

4.1.1 Audio Capture

Using the PyAudio [71] library, two audio streams are captured in real time: one for the patient and another for the medical professional conducting the teleconsultation. Most of the audio processing is done on the patient's stream, but since one of the most important features to extract is response time (i.e. the time it takes for the patient to respond to a question), the medical professional's stream is also captured.

Since PyAudio captures audio stream data by accessing a computer's microphone, and since the tool created needs to function in a teleconsultation environment, the way the patient's audio stream is captured is by routing the output audio from the software being used for the teleconsultation to the input

audio of the tool. This leads to platform independence, meaning that the tool can be used in conjunction with various videoconference platforms such as Zoom, Microsoft Teams or Google Meet. The audio routing is done using the VB-Audio Virtual Cable[1] software, which supports both Windows and MacOS.

A virtual cable is a software-based audio input and output device that can be used to route audio from one application to another, without the need for physical audio cables. With this device it is possible to use audio from one program as the input for another, "tricking" the second one into treating the audio as if it is coming from a microphone.

The stream data is configured to capture audio at a sample rate of 16000 Hz, with a bit depth of 32 bits and a single channel. The stream data is then gathered in chunks of 512 samples, which corresponds to a duration of 32 ms.

It should be noted that the audio capture process is done in a non-blocking way, meaning that the audio data is captured in a separate thread from the both the feature processing thread and the main thread of the program. This is done to avoid blocking the main thread of the program, which could lead to a lag in the displayed data in the GUI.

4.1.2 Feature Extraction

Given the complexity of extracting features from speech signals in real time without the need for a very powerful computer the choice of the Parselmouth library [72] was due to the fact that it is built on top of the Praat software [73] which is commonly used in the literature as a tool for speech analysis.

Response Time

In order to calculate the response time feature, a simple Voice Activity Detection (VAD) is applied to both audio streams. Two different VADs were tested: the first one is based on a Python interface for the WebRTC Voice Activity Detector developed by Google [74] and the second one is based on the pre-trained Silero model [75]. After testing both VADs, the Silero model was chosen due to its better performance in differentiating speech segments from another type of sounds (e.g. human voice vs. hands clapping).

The response time is calculated by starting a timer when the VAD stops detecting speech in the medical professional's (local) audio stream and stopping it when the VAD detects speech in the patient's (away) audio stream. In order to avoid the introduction of noise in the response time feature, the timer is only started when the medical professional's audio stream is silent for more than 1 second.

Contingency measures were implemented to ensure accurate response time calculation. The timer starts only when the medical professional's microphone has just become inactive, which is achieved by tracking the microphone's activity state from the previous read cycle. Additionally, a reset mechanism

[1] https://vb-audio.com/Cable/

was created in order to prevent multiple timers from being initiated during periods of consecutive silences in the local audio stream when no speech is detected from the patient's audio.

Since the utilization of a virtual cable to route the audio leads to the introduction of a varied amount of delay between the two audio streams the Welford Algorithm[2] for calculating variance was applied in order to calculate the moving average delay between the two streams and then subtract it from delay detected each time a chunk of data is processed. With this approach, one can have a corrected delay value that can be then subtracted from the calculated response time in order to generate more accurate data.

In Figure 4.2 one can observe a representation of how the response time feature is calculated, keeping in mind the existence of variable delay between the two audio streams.

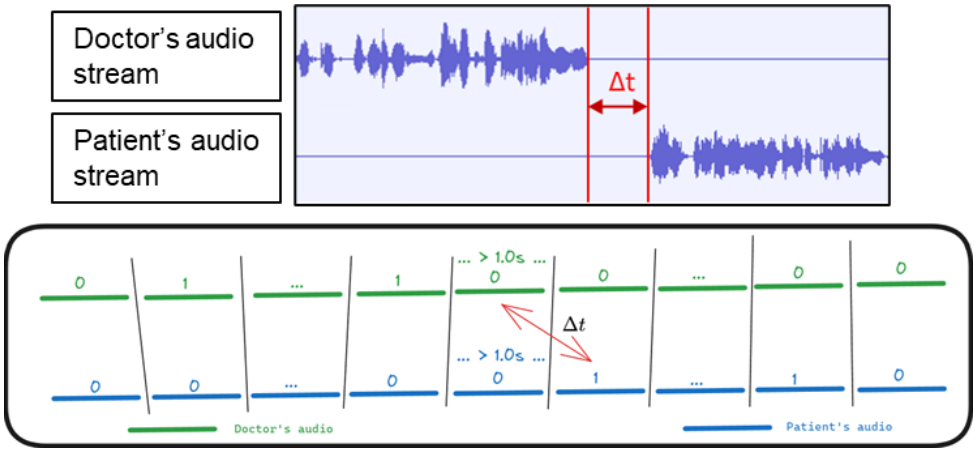


Figure 4.2: Abstract scheme of the Response Time calculation with a binary representation of the VAD output.

Prosodic Features

The other features are calculated over a 5-second window of audio data. The prosodic group of features is obtained by running a Praat script created by Nivja H. de Jong & Ton Wempe [76] in 2009[3] that is able to detect syllable nuclei in voiced segments of audio data through the analysis of intensity variations in the signal. This is possible through the assumption that a vowel within a syllable nucleus is characterized by a higher energy than the surrounding consonants. Having this in mind, the script compares the intensity of multiple successive peaks and selects the ones that are both above a certain defined threshold and preceded by a considerable dip in intensity.

The thresholds used in the script were adjusted to better fit the Portuguese language, leading to the use of a silence threshold of $-20\ dB$ for the detection of voiced segments, a minimum dip near an

[2]For the mathematical formulas of the Welford Algorithm, refer to Appendix A, Section A.3
[3]Even though the study was done in 2009 the software has been updated over the years, most recently in 2019

intensity peak of $2\ dB$ and a minimum pause duration for classification of $0.2\ s$. With these parameters, it is possible to obtain a reasonable approximation of the following prosodic features:

- **Number of Syllables**
- **Number of Pauses**
- **Duration of the audio segment**
- **Phonation Time**
- **Speech Rate**
- **Articulation Rate**
- **Average Syllable Duration**

It should be noted that the script is run on a repeated timer every 5 seconds, this imposes some limitations related to the fact that there might not be enough data to calculate feature values as the use case the tool is intended for leads to an environment where the patient might not be speaking all the time. In order to mitigate this issue in a simple way, if any of the features throws either an undefined value or the difference between the maximum and minimum intensity is below the threshold, the feature set is not calculated as it is assumed that the patient is not speaking, or the reading is not clear enough to extract the features.

Acoustic Features

The rest of the selected features are obtained through the correspondent Parselmouth functions and are divided into two groups: the first group represents features to be displayed in real time in the GUI and the second group represents features that are calculated at the end of the session background and stored in a database for further analysis. The decision to split the features into two groups was made in order to avoid overloading the GUI with too much information and because some features are more relevant for the medical professional to have access to in real time, while others are more relevant for further analysis. This also reduces the computational load on the system, as some features require more processing power to be calculated.

The local jitter and shimmer, HNR, Pitch and its standard deviation, are calculated over the same 5-second window as the prosodic features. In order to obtain these values the following Praat commands are used: *Get local jitter (local, absolute), Get local shimmer (local_dB)*, *To Harmonicity (cc), Get mean*, *To Pitch (cc), Get mean, Get standard deviation*, respectively. These values are then stored on circular buffers, which have a fixed size of 20 elements in order to be displayed in the GUI.

The features that integrate the second group are multiple other methods of measuring jitter and shimmer such as the Period Perturbation Quotient (PPQ) for jitter and the Amplitude Perturbation Quotient (APQ) for shimmer (which can be calculated for varied point systems ranging from 3 to 11 points) and the local calculations represented with percentages. The first, second and third formants are also calculated and stored in the database. This can be achieved by saving into an array the Parselmouth

sound and *pointprocess* objects obtained every read cycle and then calculating the features at the end of the session.

The *sound object* is a representation of the captured audio in one or multiple channels with values varying from -1 to 1 and the *pointprocess object* is a sequence of points in time that is set to capture the acoustic periodicity of the signal as the frequency of a physiological process such as the glottal pulses in vocal fold vibration [73].

As referred in both Section 2.1.2 and 3.1 there are identifiable differences in the pitch range of the voice of individuals depending on their sex and age. Since creating a system for gender and age identification was not the main goal of this project, a pitch range that encapsulates the average conversational pitch range for adult Portuguese speakers, ($187-210\ Hz$ and $109-123\ Hz$ for women and men, respectively [66]), was chosen. This range is from $80\ Hz$ to $300\ Hz$ and is used to calculate every *pointprocess object* needed for feature extraction.

4.1.3 Tool Versatility

It is worth mentioning that even though the tool was developed with the purpose of being used in a teleconsultation environment, it can be used in other scenarios. There is no direct implication for one of the speakers to be an individual with a psychiatric disorder, the tool can be used in any scenario where the extraction of the selected speech features is needed.

A possible use case for this audio processing pipeline is in the field of fatigue detection. Studies have shown that speech features such as jitter, shimmer, and pitch can be used to detect fatigue in individuals [77]. So, using the same interview format as the one used for the teleconsultation environment, the tool can be used to detect fatigue in individuals in a work environment, for example.

4.2 Python GUI

The extracted features are displayed in a Python Graphical User Interface (GUI) using the PyQt5 library [78]. Since the pre-existing telemedicine platform was also developed using PyQt5 the integration of the audio processing pipeline was straightforward. In order to make the data presented more user-friendly, the feature values are each represented in the correspondent mean value updated with every new reading through the Welford Algorithm.

4.2.1 Pre-Existent Graphical User Interface

The pre-existent tool, as mentioned before, was developed by Diogo Ramalho et al. [20, 21] with the objective of processing video data in real time. The GUI was developed using PyQt5 and is divided into

three main sections:

- **Video:** This section displays the video stream of the patient, the time elapsed since the beginning of the teleconsultation and the working frame rate of the system.
- **Features:** This section has a series of text boxes that display the values of the features extracted from the video data. These encapsulate the patient's heart rate, two measures of Heart Rate Variability (HRV) (Standard Deviation of NN Intervals (SDNN) and Root Mean Square of Successive RR interval Differences (RMSSD)), a graph with the percentages of Low Frequency (LF) and High Frequency (HF) of the HRV, the head velocity and total blinks.
- **Signals:** This section displays three graphs: the Eye Aspect Ratio (EAR) signal, the predicted blinks and the pulse and peak detection HRV signals.

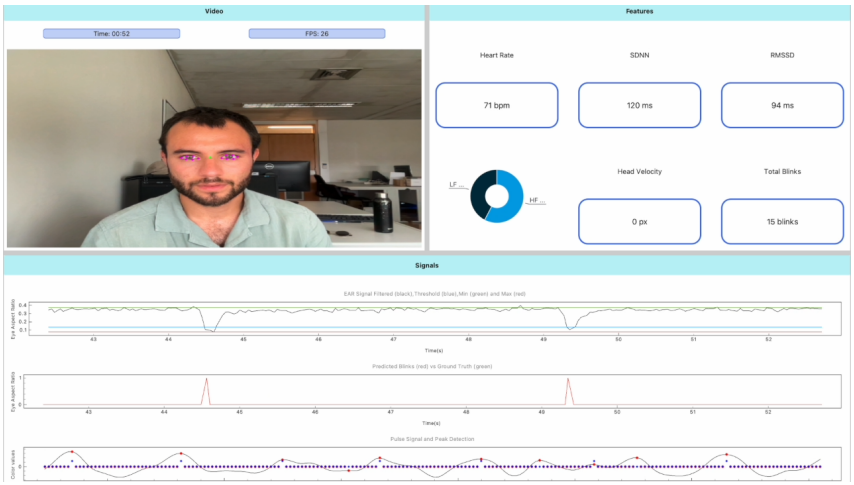


Figure 4.3: Pre-existent Graphical User Interface with real time video processing.

The GUI data is updated every time a new frame is processed, which leads to a very dynamic GUI, but that tends to be dependent on the frame rate of the system. This can be a hindrance when the system is under heavy load or the hardware is not powerful enough to process the video data in real time leading to a lag in the displayed data and a less smooth user experience.

Video Source

This tool was developed with the possibility of processing and displaying video from varied sources, such as a webcam, a video file or a screen capture. The video source is selected by the user at the beginning of the session. The video source is processed in real time and the features are displayed in the GUI as soon as they are calculated.

The webcam and video file sources are processed using the OpenCV library[4] and the screen capture

[4] https://pypi.org/project/opencv-python/

source is processed using the *mss* library[5]. The video file and webcam sources are primarily used for testing purposes, while the screen capture source is used in the final product, as it allows the tool to be used in conjunction with any videoconference platform.

The way Diogo Ramalho et al. [20, 21] envisioned the video capture process for a teleconsultation environment was by recording the screen region where the video of the patient is displayed. Since, as can be seen in Figure 4.3, the GUI is a full screen window, there is a need to have a second screen to display the video of the patient. With this in mind, the user needs to first open a videoconference platform, such as Zoom, and set the video display to gallery view, so that all participants are displayed in a grid. In order to avoid having to select the screen region every time the tool is used, the screen region was "hard coded" to capture the top left corner of the screen, where the video of the patient is displayed in the gallery view.

4.2.2 Multimodal Graphical User Interface

As a way to reduce the amount of information displayed in the GUI the feature values are presented in simple text boxes and the graphs were relocated to a new window accessible with a simple button click. Since some feature values might be hard to interpret "on the fly", a hover event was implemented over each one of them presenting the mean value of that same feature for the last session the current patient had, which can be useful for comparison purposes as can be seen below.

Figure 4.4: Example of a hover event over the Pitch feature.

In order to join the audio processing section with the pre-existent GUI some changes were made to the latter. The **Signals** section was divided into two new sections: **Video Signal Features** and **Audio Signal Features**. The first section presents data that previously was displayed in the **Signals** section through EAR and HRV graphs and the second section presents the selected audio features. In Figure 4.5 one can observe the created GUI with both video and audio processing sections.

[5] https://pypi.org/project/mss/

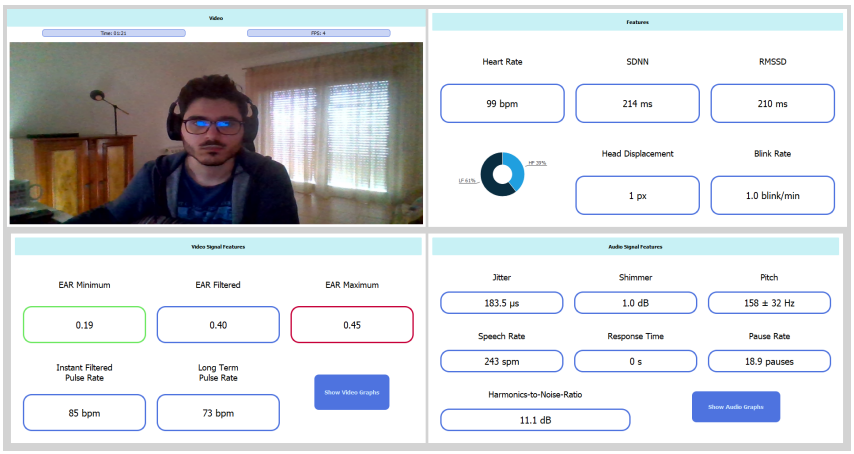


Figure 4.5: Multimodal Graphical User Interface with real time video and audio processing.

By pressing the button on each of the **Signal Features** sections, a new window is opened with the correspondent graph as observed in Figure 4.6. The **Video Signal** window displays the previously mentioned EAR and HRV graphs and the **Audio Signal** window displays the average **Pitch**, **Jitter**, **Shimmer** over time with a red dotted line representing the mean at the end of the previous session. It should be noted that in the case of the **Pitch** graph, the standard deviation is also displayed in order to give a better understanding of the pitch range of the patient.

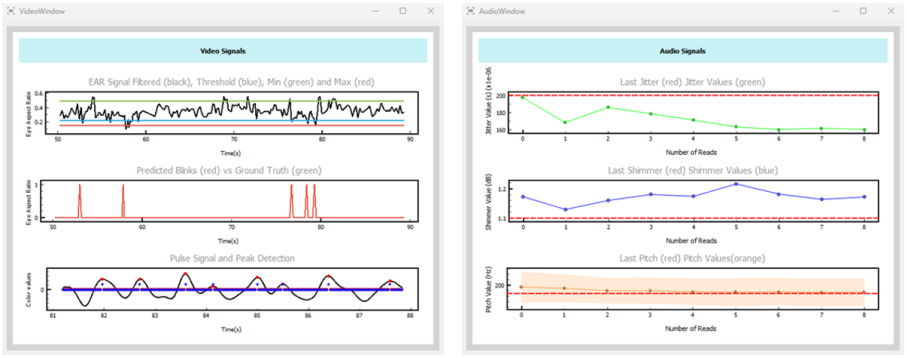


Figure 4.6: Graph windows for the Video Signal Features (left) and Audio Signal Features (right).

At the end of each session, the mean feature values are stored in a data reports for further analysis. This way, two reports are generated: one with the video data and the other with the audio data.[6]

[6] For an example of the generated report refer to Appendix B, Section B.5

5

Results

Contents

5.1 Efficacy Tests 46

5.2 Feature Behaviour with Real-World Data 49

5.3 Conclusions 51

In this chapter, the results of simple efficacy tests are presented. Its objective is to prove the concept of the proposed methodology, by comparing the feature values calculated by the system and comparing them with the corresponding values found in the literature. Here it is also presented the final tool developed, after the implementation of the proposed methodology and the appropriate adjustments.

5.1 Efficacy Tests

In order to verify the efficacy of the proposed system and the Praat scripts developed, some simple tests were performed. These tests were based on the comparison of the values obtained by the system with the values found in the literature and openly available databases.

5.1.1 Syllable Counting

Recurring to the Tatoeba project [1] which collects sentences in multiple languages, a group of American English and Brazilian Portuguese sentences was selected to verify the efficacy of the syllable counting Praat script. The sentences were selected to cover a wide range of syllable counts, from 5 to 15 syllables (which were manually annotated) and durations comprehended between 2 and 3 seconds.

Table 5.1: Syllable counting comparison for English and Portuguese sentences.

Sentence	Annotated Syllable Number	Obtained Syllable Number
“The station is just up ahead”	8	7
“I took a taxi to the station”	9	8
“I went to the station by cab”	8	8
“You can get to the station by bus”	9	9
“I have to go to the station right away”	11	11
“Does this bus go to the station”	8	10
“We’re supposed to meet at the station at five o’clock”	13	12
“That’s the bus to the station”	7	7
“I have to go to the station at three o’clock”	12	10
“The station is ten minutes from here by car”	11	10
“Eu havia comprador na semana anterior”	15	9
“O que causou o problema”	8	4
“O Tom trabalha numa central de atendimento”	15	12
“A Melissa faltou à aula de novo”	12	7
“Como ele descobriu isso”	10	8

[1] https://tatoeba.org/en/

As it can be seen in Table 5.1, the script tends to underestimate the number of syllables in English sentences by one to two syllables. In the case of Portuguese audio, the variation is higher which was expected since the script was developed for English. This limitation supports the need for a language-specific syllable counting script and supports the previously mentioned concept of only using the obtained data for relative comparisons for measurements obtained for the same speaker.

5.1.2 Feature Benchmarking

As presented in the Section 3.2.1 it was possible to find some benchmark values for the features extracted by the developed tool. By processing the audio data of an open-source database of sustained vowel of a varied group of speakers developed by David Venegas[2] **Jitter**, **Shimmer** and **Pitch** values were extracted and compared with the values found in the literature.

The dataset contains 1676 audio files of sustained vowels, in which varied speakers pronounce the vowels /a/, /e/, /i/, /o/ and /u/. There are around 80 files for each vowel sustained by male speakers and 30 files for each vowel sustained by female speakers. The audio files are in the *.wav* format and have varied durations. For the purpose of this test, there were randomly selected 30 audio files for each of the vowels that had correspondence in the Table 3.2 whose duration was around 3 seconds.

It should be noted that the values found in the literature where only used as a reference, since not only the studies in which they were obtained might be outdated, as the research was conducted during the 80s and 90s, but also the database used for the tests was not recorded under controlled conditions and there is a high variability in the speakers' characteristics and the recording conditions.

In relation to the obtained **Pitch** values, it was possible to verify that the values obtained by the system are in the same range as the values found in the literature, since the literature values are in the standard deviation range of the values obtained by the system for both male and female speakers[3].

In the below presented graph (Figure 5.1), it is possible to detect a clear difference between the Pitch values for the different sexes. Specifically, the data reveal that females tend to have higher pitch values compared to males, consistent with the findings reported in the existing literature.

[2] https://www.kaggle.com/datasets/darubiano57/dataset-of-vowels/data
[3] For the obtained graphical data, refer to Appendix B, Section B.1

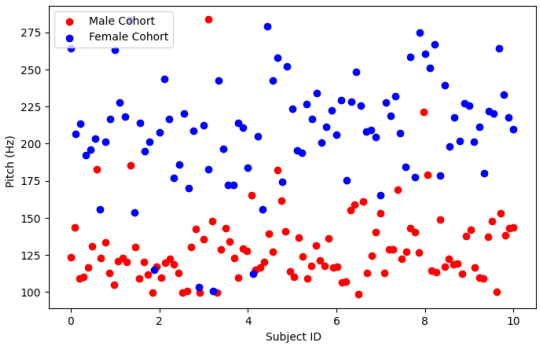


Figure 5.1: Graphical representation of the overall Pitch values obtained in the benchmark separated by sex

Regarding the **Jitter** values obtained, it was possible to verify that the values they have a similar behaviour to the values found in the literature, since the values obtained by detected by the system are mostly in the standard deviation range of the values found in the literature[4].

Even though there is no clear consensus in the literature about the differences in Jitter values between sexes, the obtained data (Figure 5.2) shows a higher jitter value for the male cohort which that correlates with what is presented in the Table 3.2 obtained from the literature.

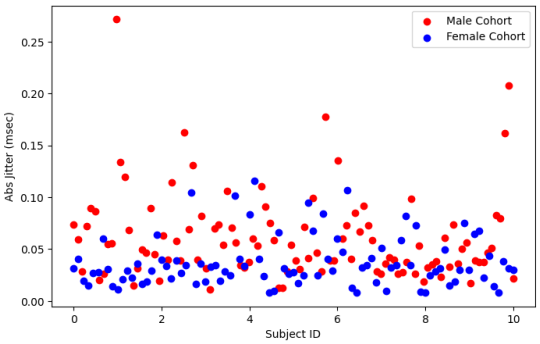


Figure 5.2: Graphical representation of the overall **Jitter** values obtained in the benchmark separated by sex

The **Shimmer** values obtained by the system were also compared with the values found in the literature. In this case, the values obtained by the system are in the same range as the values found in the literature, as the majority or the values obtained by the system for each vowel seem to be extremely dispersed with a lot of outliers, and the comparison between sexes does not show a clear difference[5]. There is, however, a test that behaved differently, the vowel /u/ for the male cohort in which the majority of the obtained shimmer values are present in the standard deviation range of the literature values.

It should be noted that the literature values found in the literature show small difference between different sexes, which might explain the lack of difference in the obtained data. The book “A Ciência e a Arte da Voz Humana” [8] also presents some limitations for the calculation of the Shimmer values, which

[4]For the obtained graphical data, refer to Appendix B, Section B.2
[5]For the obtained graphical data, refer to Appendix B, Section B.3

is highly dependent on the quality of the recording, the characteristics of the speaker and the type of acoustic signal [8, p. 183].

The results of the benchmarking process show that the system is able to extract the features with a similar accuracy to the values found in the literature. This is an overall good indicator that the system is able to extract the features correctly and that the developed code is working as expected.

5.2 Feature Behaviour with Real-World Data

In Section 3.2.2, the predicted behaviour of the selected speech features in individuals who suffer from MDD was presented based on the state-of-the-art information found. In order to verify that the developed system is able to detect the same behaviour, a test was performed using audio data from the **DAIC-WOZ** database [79]. It includes 189 audio files of interviews conducted animated virtual interviewer called Ellie with annotated labels for the depression level of the speaker based on the Eight-Item Patient Health Questionnaire (PHQ-8) depression scale.

The average audio duration of the audio files is 16 minutes, and the audio recording quality is not controlled, which might introduce some noise in the audio files. In order to obtain better results a couple of audio files were removed given the high level of noise present in them, which lead to 56 audio files of patients with MDD and 131 audio files of patients without MDD.

The database documentation also includes the values for first 5 formants calculated for during each interview with a sample rate of $100\ Hz$. Since the developed system is not prepared to extract features from samples with such a small duration ($10\ ms$), formant value comparison was done through the calculation of the moving average and standard deviation of the formant values for the whole audio file.

The moving average of each feature was updated over two second segments of speech data, following the same procedure as when the speech processing tool analyses audio in real time which were described in the previous chapter.

As referred in Section 4.1.2 the developed system only computes the first three formants, which are deemed the most relevant for speech analysis. Comparing the obtained values with the ones present in the database documentation it was possible to verify that the system is able to extract the formant values correctly, since the values obtained by the system are in the standard deviation range of the database documentation values[6]. It was identified a characteristic in the documentation data that limits the possible conclusions that can be drawn from the comparing the different data groups, which is due to the fact that the formant values in the documentation have high standard deviations.

In order to verify if the system is able to detect different feature values for individuals with MDD (experimental group) and individuals without MDD (control group), the same process utilized for the

[6] For the obtained graphical data, refer to Appendix B, Section B.4

extraction of the formant values was extended to the **Jitter**, **Shimmer** and **Pitch**, **HNR** and **Speech Rate** features.

The overall averages for the extracted features for the individuals in the control and experimental groups are presented in Table 5.2.

Table 5.2: Average feature values extracted from speech data of individuals with and without MDD (*$p < .05$).

Feature	Control Group	Experimental Group	T-statistic(p-value)
Jitter (msec)	0.122±0.02	0.125±0.04	0.66 (0.255)
Shimmer (dB)	1.233±0.13	1.217±0.13	-0.75 (0.774)
Mean Pitch (Hz)	229.26±16.29	228.69±13.97	-0.24 (0.406)
Standard Deviation Pitch (Hz)	26.29±6.44	24.58± 6.12	-1.69 (0.048)*
Harmonics to Noise Ratio (dB)	8.35±1.62	8.52±1.76	0.61 (0.270)
First formant (Hz)	750.43±52.62	745.98±60.84	-0.50 (0.308)
Second formant (Hz)	1500.93±49.61	1518.34±53.49	2.13 (0.983)
Third formant (Hz)	2468.97±56.05	2474.28±59.45	0.58 (0.718)
Speech Rate (syll/sec)	3.47±0.33	3.37±0.31	-2.03 (0.021)*

As it can be seen in Table 5.2, the system seems to be able to calculate feature values that demonstrate the expected behaviour for individuals with MDD and individuals without MDD. The **Jitter** values are higher for individuals with MDD compared to individuals without MDD, which is consistent with the literature. The **Pitch** values are lower for individuals with MDD compared to individuals without MDD, which was expected.

The HNR values are higher for individuals with MDD compared to individuals without MDD, which is not consistent with the literature. This might be due to the limitations presented in the previous section, associated with data quality and the different characteristics of the multiple speakers.

In the case of the **Formants** values, only the **Formant 1** values are in agreement with the literature, since the values for individuals with MDD are lower than the values for individuals without MDD. The **Formant 2** and **Formant 3** values are higher for individuals with MDD compared to individuals without MDD, but with a small difference, which imposes the need for further analysis in order to verify if this difference is significant.

The **Speech Rate** values are lower for individuals with MDD compared to individuals without MDD, which is also consistent with the literature.

In the case of the **Shimmer** values, the system was not able to detect the expected behaviour, since the values for individuals with MDD are lower than the values for individuals without MDD. This might be due to the limitations presented in the previous section, associated with the sensitivity of this acoustic

feature to the recording conditions and the characteristics of the speaker. There is also the fact that in this specific test, data is being extracted from speech audio from multiple different speakers, which means the feature comparison is not being done for the same speaker over time as is believed to be the best approach for the assessment of affective disorders.

A simple statistical analysis was performed in order to verify if the differences found in the extracted feature values are significant. The analysis was performed using the *Student's t-test* for independent samples, with a significance level of 0.05. The results demonstrated that of the features presented in Table 5.2, only **SD Pitch** ($p = 0.046$) and **Speech Rate** ($p = 0.021$) values have a significant difference between the control and experimental groups.

It is important to note that the results presented in Table 5.2 are only a preliminary analysis of the system's ability to detect the expected behaviour for individuals with MDD and individuals without MDD. Further analysis is needed in order to verify if the differences found are significant and if the system is able to detect the expected behaviour for the features extracted.

5.3 Conclusions

The results of the efficacy tests show that the developed system is able to extract most of the features correctly and that the extracted values seem to be in the same range as the values found in the literature. The system is also able to detect the plausible feature behaviour for individuals with MDD and individuals without MDD.

As a way to set the startup variables, an initial window was created in order to allow the user to select or add a patient, to select the video input type (camera, file or screen recording) and to confirm if the microphones are correctly set up. Figure 5.3 presents the initial window of the developed tool.

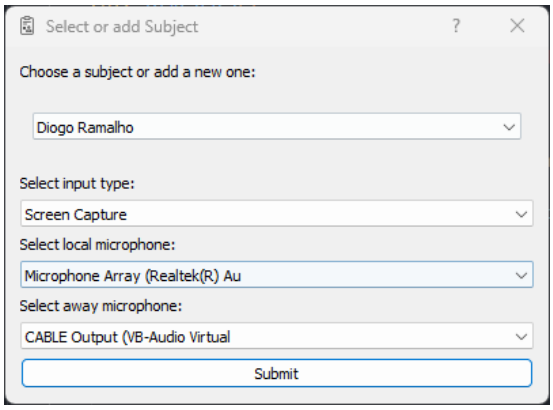


Figure 5.3: Initial window of the developed tool

After the implementation of the proposed methodology some final adjustments where made to the

tool's graphical interface in order to solve a problem presented by a medical professional while testing the tool. The problem was related to the fact that the way the video was being collected required a second screen to be connected to the computer, which was not always possible.

The solution for this problem was solved by changing the overall GUI without removing any of its functionalities. This was done by removing the patient's video from the GUI and making its main window only use half of the screen. This way the medical professional can use the videoconference software in the other half of the screen, as presented in Figure 5.4.

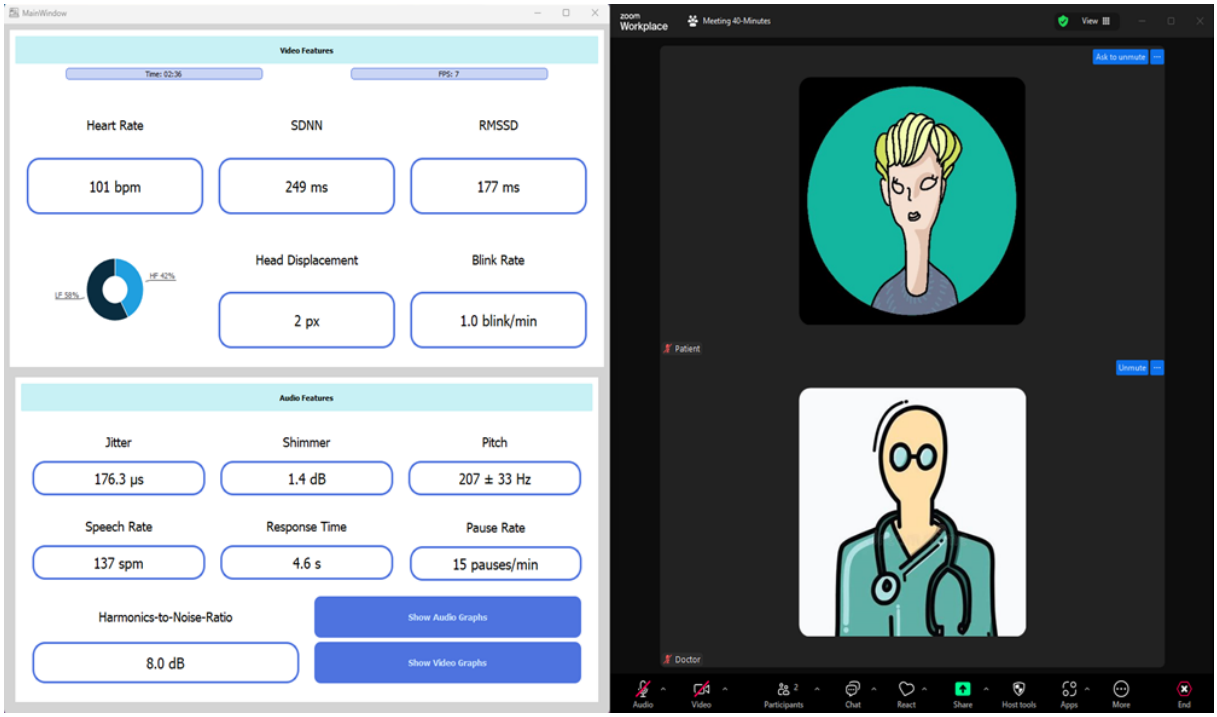


Figure 5.4: Final version of the developed tool's GUI in a split screen configuration with Zoom

A software guide was also devised in order to help first time users of the tool to understand how to install and later use it (refer to Appendix B, Section B.6).

6

Conclusion

Contents

6.1 Future Work 54
6.2 Final Remarks 55

6.1 Future Work

With the development of the proposed tool, there are multiple ways in which it can be improved and expanded upon. Some of the most promising speech features that were not addressed in this project since there isn't a clear consensus on whether they can be used to infer depressive states in a patient suffering from affective disorders are: Teager energy operator coefficients (TEOC), Linear Predictive Coding (LPC) and Cepstral Peak Prominence (CPP). TEOC has the purpose of estimating the energy characteristics of a signal by taking into account how its amplitude and frequency vary over time [80, 81], LPC which is a time-domain feature characterized by representing the speech signal as a linear combination of past samples [43, p. 5–6], and CPP which is a feature that is used to estimate the prominence of the highest peak in the cepstrum which is a pitch-independent method [54].

To build on the work presented in this report, several potential avenues for future research and development have been identified and are listed below:

- **Reevaluate Shimmer extraction:** Reevaluate the way the Shimmer feature is calculated, since the results obtained in this project differ from the ones described in the literature;
- **Test other speech features:** Test other speech features such as the ones mentioned above to see if they can be used on the assessment of psychiatric disorders;
- **Improve application efficiency:** With the support of a computer engineer or a computer scientist, improve the efficiency of the application by optimizing the code and the algorithms used, making it OS-independent;
- **Run a patient study:** Run a patient study to test the application in a real-world scenario and verify the results obtained in this project;
- **Discern areas for improvement based on clinician's opinion:** Gather feedback from clinicians and medical professionals to understand what they think could be improved in the application;
- **Develop a post-consultation analysis tool:** Develop a tool that can be used to analyze the data collected during the consultation and present medical professionals with a more detailed analysis of the patient's speech;
- **Move the application to a web-based platform:** Develop a web-based platform in which the videoconference section is embedded, making it easier for medical professionals to use the application;

Addressing these points will make the application more robust and reliable, and will make it easier for medical professionals to use it in their daily practice.

6.2 Final Remarks

The development of this project was a great opportunity to learn about the different aspects of speech processing and how it can be used to extract information about the mental health state of an individual. The initial proposition of the project was to research the state-of-the-art of speech processing techniques and find out which ones were the most promising for the development of a tool that could be used by medical professionals to help infer PMR characteristics in a patient suffering from affective disorders during a telemedicine consultation, which was effectively achieved.

After the state-of-the-art research was done, it was possible to conclude that changes speech features such as **Jitter**, **Shimmer**, **Pitch**, **HNR** and **Formants** can be used to infer depressive states in a patient suffering from affective disorders. With that in mind a methodology was developed to extract these features in real-time from a patient's speech during a telemedicine consultation and a way of presenting this information to the medical professional was also developed.

Given the two-way nature of the telemedicine consultation, it was also envisioned a way of calculating the patient's time of response to a medical professional's question and or declaration. This feature differs from the others as the way it is extracted is not well-defined in the literature, so an innovative way of calculating it was developed using voice activity data from both the patient and doctor's audio streams.

The feature extraction system was also tested, first through a series of benchmark tests to verify that the features were being extracted correctly, by comparing the results obtained to values described in the literature. The feature behavior was also tested with data from the **DAIC-WOZ** dataset, allowing to verify that most of the features were behaving as expected when used with real data.

The versatility of the audio processing tool is also a highlight of this project, as it can be used in a variety of different situations, not only in telemedicine consultations since it can be used to extract speech features from any audio stream, so if the selected features are appropriate for the task at hand, the tool can be applied to that scenario.

In conclusion, it is possible to say that the project was a success, as the proposed tool was developed and tested, and the results obtained are promising. It was also possible to expand on an existing video analysis tool, by incorporating a speech processing module, making it a more complete tool for medical professionals to infer depressive characteristics in patients with affective disorders during telemedicine consultations.

Bibliography

[1] J. Benesty, M. M. Sondhi, and Y. A. Huang, Eds., *Springer Handbook of Speech Processing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, ISBN: 9783540491279.

[2] R. Jahangir, Y. W. Teh, F. Hanif, and G. Mujtaba, "Deep learning approaches for speech emotion recognition: state of the art and research challenges," *Multimedia Tools and Applications*, vol. 80, no. 16, pp. 23 745–23 812, Jan. 2021.

[3] J. M. Coria, H. Bredin, S. Ghannay, and S. Rosset, "Overlap-aware low-latency online speaker diarization based on end-to-end local segmentation," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 1139–1146. [Online]. Available: https://github.com/juanmc2005/diart

[4] S. G. Koolagudi, Y. V. S. Murthy, and S. P. Bhaskar, "Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition," *International Journal of Speech Technology*, vol. 21, no. 1, pp. 167–183, Feb. 2018.

[5] T. Bäckström, O. Räsänen, A. Zewoudie, P. P. Zarazaga, L. Koivusalo, S. Das, E. G. Mellado, M. B. Mansali, D. Ramos, S. Kadiri, and P. Alku, *Introduction to Speech Processing*, 2nd ed. Zenodo, 2022. [Online]. Available: https://speechprocessingbook.aalto.fi

[6] S. Latif, J. Qadir, A. Qayyum, M. Usama, and S. Younis, "Speech technology for healthcare: Opportunities, challenges, and state of the art," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 342–356, 2021.

[7] J. Alammar, "The illustrated transformer," Jun. 2018, [Online; accessed 16-January-2024]. [Online]. Available: http://jalammar.github.io/illustrated-transformer/

[8] I. Guimarães, *A Ciência e a Arte da Voz Humana*. Escola Superior de Saúde do Alcoitão, Jan. 2007, ISBN: 978-989-95360-0-5.

[9] D. Chandler and R. Munday, *A Dictionary of Media and Communication*. Oxford University Press, 2020.

[10] G. Berrios and I. S. Marková, "The epistemology of psychiatry," *Revista Estudos do Século XX*, no. 19, pp. 59–70, Jun. 2018.

[11] Institute for Health Metrics and Evaluation (IHME). (2024) GBD results. Seattle, WA. [Accessed 17-December-2023]. [Online]. Available: https://vizhub.healthdata.org/gbd-results/

[12] A. P. Association, *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. American Psychiatric Publishing, 2013.

[13] GBD 2019 Diseases and Injuries Collaborators, "Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the global burden of disease study 2019," *The Lancet*, vol. 396, no. 10258, pp. 1204–1222, 2020.

[14] Lancet. (2020, Oct.) The Lancet: Latest global disease estimates reveal perfect storm of rising chronic diseases and public health failures fuelling COVID-19 pandemic. [Accessed 17-December-2023]. [Online]. Available: https://www.healthdata.org/news-events/newsroom/news-releases/lancet-latest-global-disease-estimates-reveal-perfect-storm

[15] L. Brådvik, "Suicide risk and mental disorders," *International Journal of Environmental Research and Public Health*, vol. 15, no. 9, p. 2028, Oct. 2018.

[16] J. Beezhold, L. Borgermans, D. Cozman, M. D. Giannantonio, K. Jones, and R. Nica, "Relatório de indicadores de depressão: Portugal," 2023.

[17] J. Arias-de la Torre, G. Vilagut, A. Ronaldson, A. Serrano-Blanco, V. Martín, M. Peters, J. Valderas, A. Dregan, and J. Alonso, "Prevalence and variability of current depressive disorder in 27 european countries: a population-based study," *Lancet Public Health*, vol. 6, no. 10, pp. 729–738, Oct. 2021.

[18] J. S. Buyukdura, S. M. McClintock, and P. E. Croarkin, "Psychomotor retardation in depression: Biological underpinnings, measurement, and treatment," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 35, no. 2, pp. 395–409, Mar. 2011.

[19] J. Shaver, "The state of telehealth before and after the covid-19 pandemic," *Primary care*, vol. 49, pp. 517–530, Dec. 2022.

[20] D. Ramalho, P. Constantino, H. Plácido, M. Constante, and J. Sanches, "An augmented teleconsultation platform for depressive disorders," *IEEE Access*, vol. 10, pp. 130 563–130 571, 2022.

[21] Diogo Alexandre Boasinha Ribeiro Ramalho, "Telemedicine platform for monitoring and assessment of depressive disorder," *Instituto Superior Técnico, Universidade de Lisboa*, 2021, Master's Thesis.

[22] B. Ellenbroek and J. Youn, *Gene-Environment Interactions in Psychiatry: Nature, Nurture, Neuroscience*. Academic Press, Aug. 2016.

[23] P. Harrison, P. Cowen, T. Burns, and M. Fazel, *Shorter Oxford Textbook of Psychiatry*. Oxford University Press, Oct. 2017, ISBN: 9780198747437.

[24] O. J. Bienvenu, D. S. Davydow, and K. S. Kendler, "Psychiatric 'diseases' versus behavioral disorders and degree of genetic influence," pp. 33–40, May 2010.

[25] American Medical Association, *ICD-10-CM 2024 the Complete Official Codebook*. American Medical Association, 2023.

[26] A. Picardi and P. Gaetano, "Psychotherapy of mood disorders," *Clinical Practice Epidemiology in Mental Health*, vol. 10, no. 1, pp. 140–158, Nov. 2014.

[27] E. Khawam, G. Laurencic, and Donald Malone Jr., "Side effects of antidepressants: an overview." *Cleveland Clinic Journal of Medicine*, vol. 73, no. 4, pp. 351–353, Apr. 2006.

[28] C. Volkmann, T. Bschor, and S. Köhler, "Lithium treatment over the lifespan in bipolar disorders," *Frontiers in Psychiatry*, vol. 11, May 2020.

[29] J. De Fruyt, E. Deschepper, K. Audenaert, E. Constant, M. Floris, W. Pitchot, P. Sienaert, D. Souery, and S. Claes, "Second generation antipsychotics in the treatment of bipolar depression: a systematic review and meta-analysis," *Journal of Psychopharmacology*, vol. 26, no. 5, pp. 603–617, 2012.

[30] K. Strimbu and J. Tavel, "What are biomarkers?" *Current Opinion in HIV and AIDS*, vol. 5, no. 6, pp. 463–466, Nov. 2010.

[31] G. Fagherazzi, A. Fischer, M. Ismael, and V. Despotovic, "Voice for health: The use of vocal biomarkers from research to clinical practice," *Digital Biomarkers*, vol. 5, no. 1, pp. 78–88, Apr. 2021.

[32] S. B. G. (editor), *Fundamentals of Telemedicine and Telehealth*. Elsevier, 2019.

[33] Rui Pedro Salgueiro Caceiro, "Telehealth in portugal: current state and physician satisfaction," *Universidade de Coimbra*, 2022, Master's Thesis.

[34] E. Castela, "Coimbra Telemedicine Service Improves Access to Pediatric Cardiology in Cape Verde," *Acta Médica Portuguesa*, vol. 30, no. 4, pp. 253—-254, Apr. 2017. [Online]. Available: https://www.actamedicaportuguesa.com/revista/index.php/amp/article/view/9034

[35] H. Martins, M. Monteiro, P. Loureiro, and M. Cortes, "National Strategic Telehealth Plan 2019-2022 (PENTS)," 2019. [Online]. Available: https://www.isfteh.org/files/media/PENTS_English_Version.pdf

[36] S. Chakrabarti, “Usefulness of telepsychiatry: A critical evaluation of videoconferencing-based approaches,” *World Journal of Psychiatry*, vol. 5, no. 3, p. 286, 2015.

[37] B. M. Washington, A. Robinson, T. C. Mike, M. Ruley, and A. Coustasse, “Telepsychiatry: Access in rural areas,” *International Journal of Healthcare Information Systems and Informatics*, vol. 16, no. 4, pp. 1–14, Jan. 2022.

[38] P. B. de Oliveira, T. M. Dornelles, N. P. Gosmann, and A. Camozzato, “Efficacy of telemedicine interventions for depression and anxiety in older people: A systematic review and meta-analysis,” *International Journal of Geriatric Psychiatry*, vol. 38, no. 5, May 2023.

[39] J. Lokkerbol, D. Adema, P. Cuijpers, C. F. Reynolds, R. Schulz, R. Weehuizen, and F. Smit, “Improving the cost-effectiveness of a healthcare system for depressive disorders by implementing telemedicine: A health economic modeling study,” *The American Journal of Geriatric Psychiatry*, vol. 22, no. 3, pp. 253–262, Mar. 2014.

[40] L. Gutiérrez-Rojas, M. A. Alvarez-Mon, Á. Andreu-Bernabeu, L. Capitán, C. de las Cuevas, J. C. Gómez, I. Grande, D. Hidalgo-Mazzei, R. Mateos, P. Moreno-Gea, T. D. Vicente-Muñoz, and F. Ferre, “Telepsychiatry: The future is already present,” *Spanish Journal of Psychiatry and Mental Health*, vol. 16, no. 1, pp. 51–57, jan 2023.

[41] A. Klein, J. Clucas, A. Krishnakumar, S. S. Ghosh, W. Van Auken, B. Thonet, I. Sabram, N. Acuna, A. Keshavan, H. Rossiter, Y. Xiao, S. Semenuta, A. Badioli, K. Konishcheva, S. A. Abraham, L. M. Alexander, K. R. Merikangas, J. Swendsen, A. B. Lindner, and M. P. Milham, “Remote Digital Psychiatry for Mobile Mental Health Assessment and Therapy: MindLogger Platform Development Study,” *Journal of Medical Internet Research*, vol. 23, no. 11, p. e22369, 2021.

[42] C. Zucco, B. Calabrese, and M. Cannataro, “Sentiment analysis and affective computing for depression monitoring,” in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2017, pp. 1988–1995.

[43] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, “A review of deep learning techniques for speech processing,” *Information Fusion*, vol. 99, p. 101869, Nov. 2023.

[44] D. D. DeSouza, J. Robin, M. Gumus, and A. Yeung, “Natural language processing as an emerging tool to detect late-life depression,” *Frontiers in Psychiatry*, vol. 12, Sep. 2021.

[45] V. Shuman, D. Sander, and K. R. Scherer, “Levels of Valence,” *Frontiers in Psychology*, vol. 4, 2013.

[46] J. P. Teixeira and P. O. Fernandes, “Jitter, shimmer and hnr classification within gender, tones and vowels in healthy voices,” *Procedia Technology*, vol. 16, pp. 1228–1237, 2014, cENTERIS 2014 -

Conference on ENTERprise Information Systems / ProjMAN 2014 - International Conference on Project MANagement / HCIST 2014 - International Conference on Health and Social Care Information Systems and Technologies.

[47] P. Boersma, “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” pp. 97–110, 1993.

[48] Y. Xu, “Prosody, Tone and Intonation,” *The Routledge Handbook of Phonetics*, pp. 314–356, 2019, ISBN: 9780429509803.

[49] C. Solomon, M. F. Valstar, R. K. Morriss, and J. Crowe, “Objective methods for reliable detection of concealed depression,” *Frontiers in ICT*, vol. 2, Apr. 2015.

[50] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang, “A comparative study on transformer vs rnn in speech applications,” in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, Dec. 2019.

[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017.

[52] M. Yamamoto, A. Takamiya, K. Sawada, M. Yoshimura, M. Kitazawa, K.-c. Liang, T. Fujita, M. Mimura, and T. Kishimoto, “Using speech recognition technology to investigate the association between timing-related speech features and depression severity,” *PLOS ONE*, vol. 15, no. 9, p. e0238726, Sep. 2020.

[53] A. C. Arevian, D. Bone, N. Malandrakis, V. R. Martinez, K. B. Wells, D. J. Miklowitz, and S. Narayanan, “Clinical state tracking in serious mental illness through computational analysis of speech,” *PLOS ONE*, vol. 15, no. 1, pp. 1–17, Jan. 2020.

[54] M. Seifpanahi, T. Ghaemi, A. Ghaleiha, D. Sobhani-Rad, and M.-K. Zarabian, “The association between depression severity, prosody, and voice acoustic features in women with depression,” *The Scientific World Journal*, vol. 2023, pp. 1–8, Dec. 2023.

[55] W. J. Silva, L. Lopes, M. K. C. Galdino, and A. A. Almeida, “Voice acoustic parameters as predictors of depression,” *Journal of Voice*, vol. 38, no. 1, pp. 77–85, Feb. 2021.

[56] T. Quatieri and N. Malyska, “Vocal-source biomarkers for depression: A link to psychomotor activity,” *Proceedings of Interspeech*, vol. 2, pp. 1059–1062, Jan. 2012.

[57] G. Kiss and K. Vicsi, “Mono- and multi-lingual depression prediction based on speech processing,” *International Journal of Speech Technology*, vol. 20, no. 4, pp. 919–935, Sep. 2017.

[58] M. R. Morales and R. Levitan, “Speech vs. text: A comparative analysis of features for depression detection systems,” in *2016 IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 136–143.

[59] N. W. Hashim, M. Wilkes, R. Salomon, J. Meggs, and D. J. France, “Evaluation of voice acoustics as predictors of clinical depression scores,” *Journal of Voice*, vol. 31, no. 2, pp. 256.e1–256.e6, Mar. 2017.

[60] S. Haq and P. Jackson, *Machine Audition: Principles, Algorithms and Systems*. Hershey PA: IGI Global, Aug. 2010, ch. Multimodal Emotion Recognition, pp. 398–423, iSBN-13: 978-1615209194.

[61] S. R. Livingstone and F. A. Russo, “The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english,” *PLOS ONE*, vol. 13, no. 5, pp. 1–35, May 2018. [Online]. Available: https://doi.org/10.1371/journal.pone.0196391

[62] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts, “Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology,” *Journal of Neurolinguistics*, vol. 20, no. 1, pp. 50–64, Jan. 2007.

[63] R. C. Peppard, D. M. Bless, and P. Milenkovic, “Comparison of young adult singers and nonsingers with vocal nodules,” *Journal of Voice*, vol. 2, no. 3, pp. 250–260, Jan. 1988.

[64] R. F. Orlikoff, “Vocal stability and vocal tract configuration: An acoustic and electroglottographic investigation,” *Journal of Voice*, vol. 9, no. 2, pp. 173–181, Jun. 1995.

[65] D. Sorensen and Y. Horii, “Cigarette smoking and voice fundamental frequency,” *Journal of Communication Disorders*, vol. 15, no. 2, pp. 135–144, Mar. 1982.

[66] I. Guimarães and E. Abberton, “Fundamental frequency in speakers of portuguese for different voice samples,” *Journal of Voice*, vol. 19, no. 4, pp. 592–606, Dec. 2005.

[67] L. Albuquerque, A. R. S. Valente, A. Teixeira, D. Figueiredo, P. Sa-Couto, and C. Oliveira, “Association between acoustic speech features and non-severe levels of anxiety and depression symptoms across lifespan,” *PLOS ONE*, vol. 16, no. 4, p. e0248842, Apr. 2021.

[68] D. M. Low, K. H. Bentley, and S. S. Ghosh, “Automated assessment of psychiatric disorders using speech: A systematic review,” *Laryngoscope Investigative Otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.

[69] K. Vicsi, D. Sztahó, and G. Kiss, “Examination of the sensitivity of acoustic-phonetic parameters of speech to depression,” in *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*, 2012, pp. 511–515.

[70] M. Farrús, J. Hernando, and P. Ejarque, “Jitter and shimmer measurements for speaker recognition,” in *Proc. Interspeech 2007*, 2007, pp. 778–781.

[71] H. Pham, “Pyaudio,” 2006. [Online]. Available: https://people.csail.mit.edu/hubert/pyaudio/

[72] Y. Jadoul, B. Thompson, and B. de Boer, “Introducing Parselmouth: A Python interface to Praat,” *Journal of Phonetics*, vol. 71, pp. 1–15, Nov. 2018.

[73] P. Boersma and D. Weenink, “Speak and unspeak with PRAAT,” *Glot international*, vol. 5, pp. 341–345, Jan. 2001. [Online]. Available: https://www.fon.hum.uva.nl/praat/

[74] J. Wiseman, “Python interface to the WebRTC Voice Activity Detector (VAD),” https://github.com/wiseman/py-webrtcvad, 2021.

[75] Silero Team, “Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier,” https://github.com/snakers4/silero-vad, 2021.

[76] N. H. de Jong and T. Wempe, “Praat script to detect syllable nuclei and measure speech rate automatically,” *Behavior Research Methods*, vol. 41, no. 2, pp. 385–390, May 2009.

[77] X. Gao, K. Ma, H. Yang, K. Wang, B. Fu, Y. Zhu, X. She, and B. Cui, “A rapid, non-invasive method for fatigue detection based on voice information,” *Frontiers in Cell and Developmental Biology*, vol. 10, Sep.

[78] M. Fitzpatrick, *Create GUI Applications with Python & Qt5*. PyQt5, 2016. [Online]. Available: https://www.pythonguis.com/pyqt5-book/

[79] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. R. Traum, A. A. Rizzo, and L.-P. Morency, “The distress analysis interview corpus of human and computer interviews,” in *International Conference on Language Resources and Evaluation*, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:14488823

[80] A. K. Alimuradov, “Speech/pause segmentation method based on teager energy operator and short-time energy analysis,” in *2021 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT)*, 2021, pp. 0045–0048.

[81] F. Wang and X. Shen, “Research on speech emotion recognition based on teager energy operator coefficients and inverted mfcc feature fusion,” *Electronics*, vol. 12, no. 17, p. 3599, Aug. 2023.

[82] P. Heckbert, “Fourier transforms and the fast fourier transform (fft) algorithm,” *Computer Graphics*, pp. 15–463, Feb. 1995.

[83] B. P. Welford, “Note on a method for calculating corrected sums of squares and products,” *Technometrics*, vol. 4, no. 3, pp. 419–420, 1962.

A

Support Algorithms

A.1 Fourier Transform (FT)

The FT is a mathematical operation characterized by the decomposition of a continuous time signal into a complex-valued function of frequency [82]. The FT of a continuous time signal $f(t)$ is defined as:

$$\hat{f}(k) = \int_{-\infty}^{\infty} f(t) e^{-2ikt} dt \tag{A.1}$$

A.1.1 Discrete Fourier Transform (DFT)

For periodic and discrete signals the FT can be expressed as a summation of complex exponentials. This is known as the DFT and is defined as:

$$A_k = \sum_{n=0}^{N-1} e^{-i2\pi kn/N} a_n \tag{A.2}$$

A.1.2 Fast Fourier Transform (FFT)

The FFT is an algorithm that computes the DFT of a sequence, or its Inverse Discrete Fourier Transform (IDFT). The FFT is a fast algorithm that reduces the number of computations needed for the DFT from $O(2N^2)$ to $O(2N \log n)$, where N is the number of samples in the sequence [82].

A.2 Long Short-Term Memory (LSTM)

Below are presented the equations that define the **LSTM** architecture, as described in the article "A review of deep learning techniques for speech processing" [43, p. 10].

$$t_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i), \tag{A.3}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f), \tag{A.4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \tag{A.5}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o), \tag{A.6}$$

$$h_t = o_t \odot \tanh(c_t), \tag{A.7}$$

where x_t is the input vector, h_t is the hidden state vector, c_t is the cell state vector, i_t, f_t, o_t are the input, forget and output gates, respectively, and σ is the sigmoid function, characterized by $\sigma(x) = \frac{1}{1+e^{-x}}$. W and b represent the weight matrices and bias vectors, respectively. The $\odot$ is the element-wise multiplication operator [43, p. 10].

A.3 Welford Algorithm

The Welford's Method is a simple algorithm for calculating the variance and mean of a stream of data. The algorithm is characterized by B. P. Welford as "n iteration formula for deriving the corrected sum of squares for n values from the corrected sum of squares for the first $(n-1)$ of these" [83].

Based on the definition of the sum of squares $S = \sum_k^{i=1} (x_i - \bar{x})$, where $\bar{x} = \sum_k^{i=1} x_i/k$ is the mean of the data, the Welford defines:

$$m_n = \sum_n^{i=1} x_i/n,\ n = 1,\ \ldots, k \tag{A.8}$$

$$S_n = \sum_n^{i=1} (x_i - m_n)^2,\ n = 1,\ \ldots, k \tag{A.9}$$

With that in mind, the Welford's algorithm can be defined as:

$$m_n = m_{n-1} + \frac{x_n - m_{n-1}}{n}, \tag{A.10}$$

$$\sigma_n^2 = \sigma_{n-1}^2 + \frac{(x_n - m_{n-1})(x_n - m_n) - \sigma_{n-1}^2}{n}, \tag{A.11}$$

$$s_n^2 = s_{n-1}^2 + \frac{(x_n - m_{n-1})^2}{n} - \frac{s_{n-1}^2}{n-1} \tag{A.12}$$

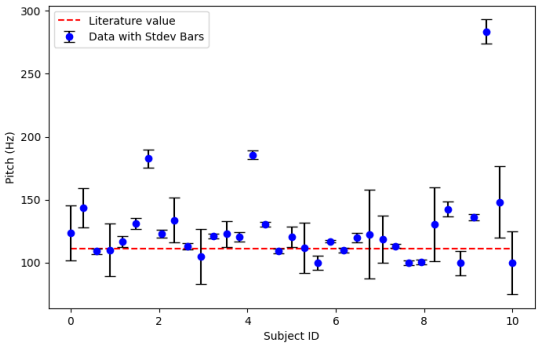
Where m_n is the mean of the first n values, σ_n^2 is the biased sample variance and s_n^2 is the unbiased sample variance.

B

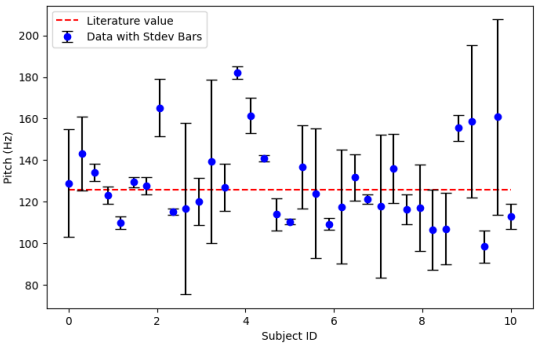
Supplementary Material

B.1 Pitch Benchmarking Figures

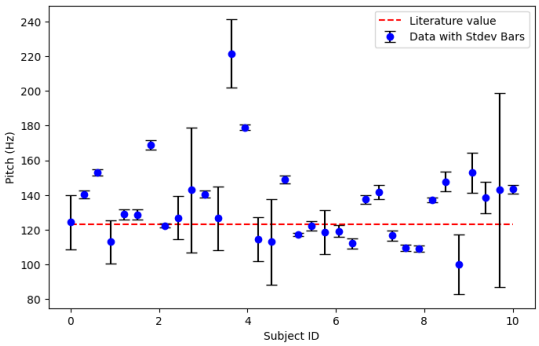
Here are presented graphs related to the feature benchmarking process for the testing of the system in the calculation of the Pitch over sustained vowels.



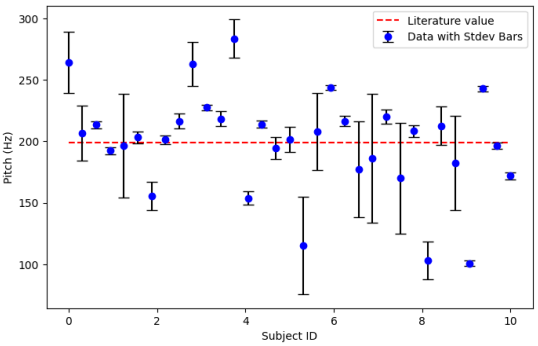
(a) Pitch values for male sustainment of the vowel /a/



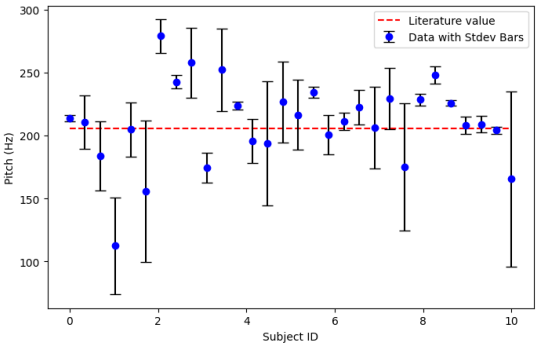
(b) Pitch values for male sustainment of the vowel /i/



(c) Pitch values for male sustainment of the vowel /u/



(d) Pitch values for female sustainment of the vowel /a/

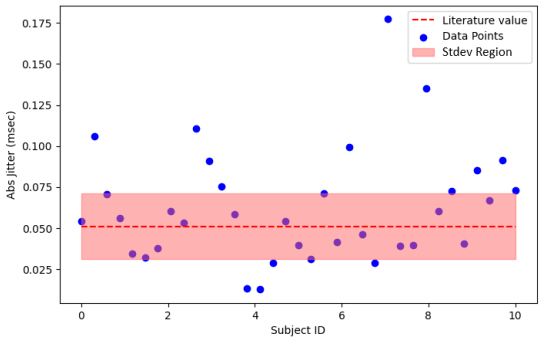


(e) Pitch values for female sustainment of the vowel /i/

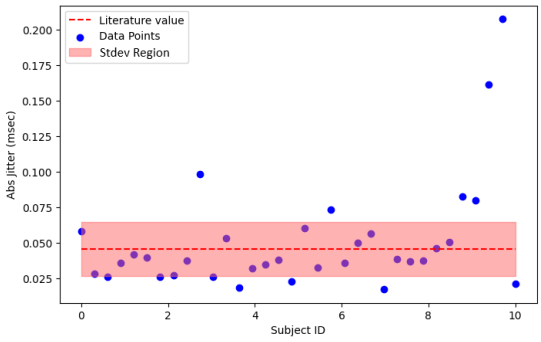
Figure B.1: Graphs of the Pitch values obtained in the benchmark

B.2 Jitter Benchmarking Figures

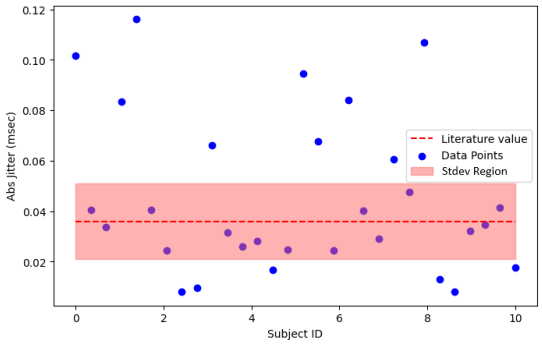
Here are presented graphs related to the feature benchmarking process for the testing of the system in the calculation of the Jitter over sustained vowels.



(a) Jitter values for male sustainment of the vowel /i/



(b) Jitter values for male sustainment of the vowel /u/

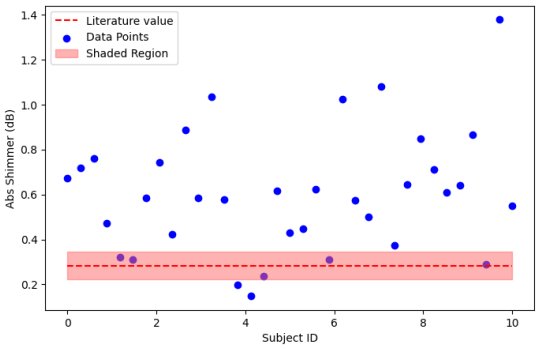


(c) Jitter values for female sustainment of the vowel /i/

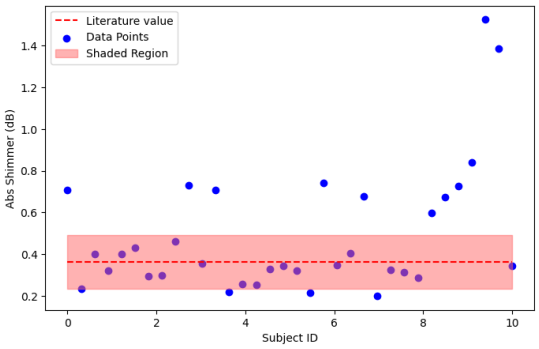
Figure B.2: Graphs of the Jitter values obtained in the benchmark

B.3 Shimmer Benchmarking Figures

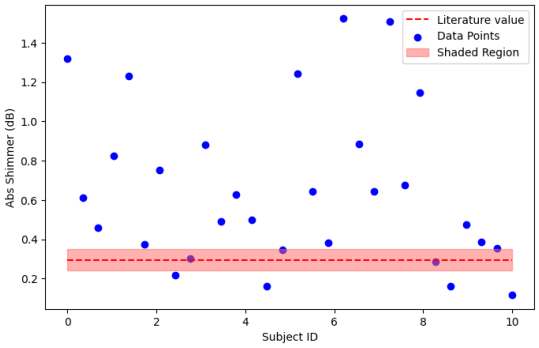
Here are presented graphs related to the feature benchmarking process for the testing of the system in the calculation of the Shimmer over sustained vowels.



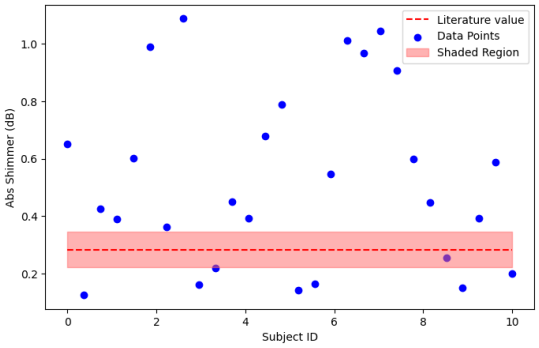
(a) Shimmer values for male sustainment of the vowel /i/



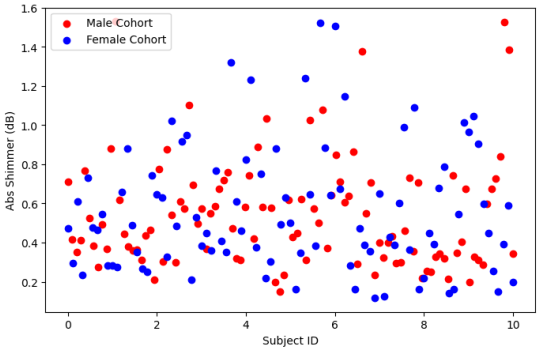
(b) Shimmer values for male sustainment of the vowel /u/



(c) Shimmer values for female sustainment of the vowel /i/



(d) Shimmer values for female sustainment of the vowel /u/

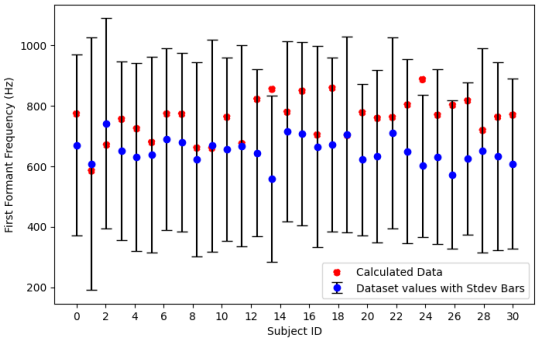


(e) Shimmer value comparison between sexes

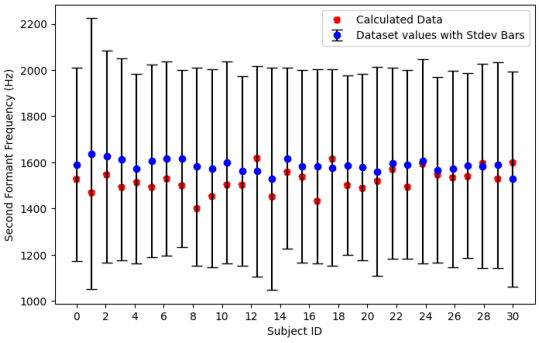
Figure B.3: Graphs of the Shimmer values obtained in the benchmark

B.4 Formant Benchmarking Figures

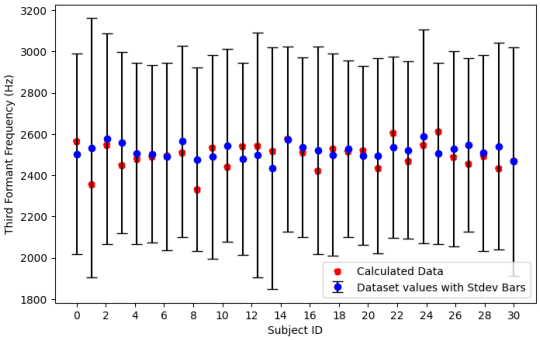
Here are presented graphs related to the formant extraction process for the testing of the system in the calculation of the formants over the speech data from varied interview sessions.



(a) First Formant values for 30 different speakers



(b) Second Formant values for 30 different speakers



(c) Third Formant values for 30 different speakers

Figure B.4: Graphs of the Formant values obtained by the system and the database documentation

B.5 Example Report

Table B.1: Example of an audio data report generated by the audio section of the multimodal tool.

Features	Session1	Session 2	Session 3
Avg Syllables (syll)	48	61	44
Total Pauses	66	31	47
Avg Measure Duration (s)	14	15	14
Avg Phonation Time (s)	10	13	11
Avg Speech Rate (syll/sec)	3.3	4.2	3
Avg Articulation Rate (syll/sec)	4.7	4.8	4.1
Avg Syllable Duration (s)	0.21	0.21	0.24
Avg Pitch	150	160	160
Avg Pitch Stdev	41	50	47
Avg HNR (dB)	7.7	9.1	8.1
Avg Local Absolute Jitter (s)	0.00026	0.00021	0.00025
Avg Local Shimmer (dB)	1.1	1.2	1.4
Avg Local Jitter	0.038	0.03	0.033
Avg Rap Jitter	0.013	0.012	0.015
Avg PPQ5 Jitter	0.017	0.015	0.018
Avg ddp Jitter	0.04	0.037	0.045
Avg Local Shimmer	0.12	0.12	0.13
Avg APQ3 Shimmer	0.044	0.047	0.057
Avg APQ5 Shimmer	0.068	0.067	0.081
Avg APQ11 Shimmer	0.13	0.12	0.14
Avg dda Shimmer	0.13	0.14	0.17
Avg First Formant (Hz)	440	520	520
Avg Second Formant (Hz)	1600	1600	1600
Avg Third Formant (Hz)	2400	2500	2500
Avg Response Time (s)	20	42	30
Duration Time (s)	61.0	203.0	305.0

B.6 Software Guide

The software developed for this Thesis is an executable file that can be run on any Windows machine.

B.6.1 Miminum System Requirements

- Windows 10 (64-bit)
- 2.0 GHz Intel Core i7 or equivalent
- 16 GB of RAM
- 2 GB of free disk space

B.6.2 Installation

- Move the tool's **project.zip** file to an appropriate and accessible folder and extract it – make sure to not move any of the files inside it;
- Go to the VB-Audio Software website (https://vb-audio.com/Cable/) and download the VB-Cable Windows driver following its installation instructions;
- On the desired video conference software select **CABLE Input** (VB-Audio Virtual Cable) as the speaker;
- Go to Settings $>$ Sound then either *More Sound Settings* (Win 11) or *Sound Control Panel* (Win 10) and on the *Recording Tab* find the **CABLE Output** device, right click to access its *Properties* and on the *Listen Tab* check the "Listen to this device" box;
- After this one should be able to run the **project.exe** executable if the starting conditions below presented are met.

B.6.3 Starting Conditions

- Project folder structure unchanged since the **project.zip** file was extracted;
- Start the video conference meeting and select **CABLE Input** (VB-Audio Virtual Cable) as the speaker
- On the video conference app set the *View* of different speakers to the *Gallery* mode leaving the patient's camera above the doctor's

- Move the video conference window to the left the screen and as tool window will appear on the right;
- Execute the **project.exe** file and wait until the subject selection window appears in the foreground
- In the Subject Selection window add or create subject and choose *Screen Capture* as the input type;
- In the same window confirm that the local microphone is correctly set and the away microphone is set to **CABLE Output** (VB-Audio Virtual Cable);
- A new white window will appear at the center of the screen, wait until it moves to the left region of the screen and data starts to be shown.

It should be noted that after the tool is closed some data might still be processed in the background, so it is advised to wait a few minutes before starting a new session. This data can be found in the data folder of the project, in the form of *.csv* files.

If any issues arise, please contact the author or refer to the project's GitHub repository for further information.