Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Control and Computer Engineering (36^{th}cycle)

Voice as the reservoir of valuable clinical information: a diagnosis and monitoring support for speech-affecting diseases

By

Federica Amato

******

Supervisor(s):
Prof. G.Olmo, Supervisor

Doctoral Examination Committee:
Prof. J.I. Godino Llorente (Referee), Universidad Politécnica de Madrid, Spain
Prof. H. Gamboa (Referee), NOVA School of Science and Technology, Portugal
Prof. M. Grangetto, Università degli Studi di Torino, Italy
Prof. G. Dimauro, Università degli Studi Aldo Moro di Bari, Italy
Prof. M. Violante, Politecnico di Torino, Italy

Politecnico di Torino
2024

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Federica Amato
2024

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

A tutte le persone a me care, fonte inesauribile di sostegno e ispirazione

Abstract

In the contemporary healthcare landscape, Artificial Intelligence emerges as a revolutionary paradigm with unprecedented potential to transform clinical practice. Vocal biomarkers, extracted from the rich information embedded in the human voice, have garnered substantial interest for their ability to provide valuable insights into various aspects of health. This dissertation delved into the multifaceted applications of vocal analysis within the healthcare domain, with a primary focus on Parkinson's Disease.

The research explored the entire pipeline of vocal analysis, encompassing data collection, development of automated analytical models, and comparative assessment of professional recording equipment versus more economical alternatives. Various speech tasks, including sustained phonation, isolated words, and text reading, were examined to identify relevant acoustic features for speech analysis. The study also investigated the influence of external co-factors and aimed to develop robust methodologies supporting diagnosis, monitoring, and follow-up of speech-affecting disorders.

A significant portion of the work was dedicated to the analysis of acoustic parameters, involving a comprehensive literature review, comparison of algorithms for parameter extraction, and exploration of new acoustic measures and analysis techniques. The research considered the dependency of these parameters on speaker characteristics, language, and the severity of the condition, as well as the recording setup. Statistical techniques and automatic classification algorithms were employed to evaluate algorithm effectiveness and propose novel pipelines for the analysis of speech samples of patients with Parkinson's Disease.

In addition, the dissertation investigated the effects of concurring pathologies, such as Gastroesophageal Reflux Disease and obesity, on vocal production. It explored the potential correlation between speech and poor sleep quality, shedding light on how temporary conditions may alter vocal patterns. The impact of transitory

alterations from alcohol consumption on speech signals was also examined, laying the foundation for assessing psychological changes, particularly in-car contexts.

The study demonstrated the potential effectiveness of voice analysis across diverse fields, addressing neurodegenerative diseases, transient conditions, and the simultaneous presence of multiple pathologies. Experiments also highlighted the variability in speech samples due to individual characteristics and propose mitigating solutions, including the incorporation of covariates among acoustic parameters and the use of domain adversarial networks.

In conclusion, this dissertation emphasized the importance of constructing specialized models tailored to specific applications, mitigating the influence of confounding factors. This approach enhances the reliability, applicability, and interpretability of generated models, laying the foundation for the effective implementation of voice analysis techniques in real-world scenarios.

Contents

List of Figures

List of Tables

Acronyms

A Alcohol-Intoxicated.

ADA AdaBoost.

ADH Alcohol Dehydrogenase.

ADL Activities of Daily Living.

ADSV Analysis of Dysphonia in Speech and Voice.

AI Artificial Intelligence.

ALC Alcohol Language Corpus.

ALDH Aldehyde Dehydrogenase.

ANN Artificial Neural Network.

ANTHEA-PDSS Anthea Parkinson's Disease Speech Samples Corpus.

APQ3 Three-point Amplitude Perturbation Quotient.

APQ5 Five-point Amplitude Perturbation Quotient.

ASV Absolute Systematic Variation.

AT Amplitude Tremor.

AUC Area Under the Curve.

BAC Blood Alcohol Concentration.

BBE Bark Band Energy.

BMI Body Mass Index.

BrAC Breath Alcohol Test.

CCC Concordance Correlation Coefficient.

CFS Correlation Feature Selection.

CNS Central Nervous System.

CPP Cepstral Peak Prominence.

CV Cross Validation.

CYP2E1 Cytochrome P450 2E1.

D2 Correlation Dimension.

DARTH-VAT DARTH Voice Analysis Toolbox.

DBS Deep Brain Stimulation.

DCT Discrete Cosine Transform.

DFA Detrended Fluctuation Analysis.

DL Deep Learning.

DPI Duration of Pause Intervals.

DR Duration Ratio.

DT Decision Tree.

DWT Discrete Wavelet Transform.

ECG Electrocardiography.

EEG Electroencephalography.

EMG Electromyography.

EOG Electrooculography.

ETS Energy Transition Slope.

F0 Fundamental Frequency.

F1 First Formant.

F2 Second Formant.

F3 Third Formant.

FT Frequency Tremor.

GERD Gastro-Esophageal Reflux Disease.

GNE Glottal to Noise Excitation Ratio.

GP Gaussian Process.

GUI Graphical User Interface.

H Hurst Exponent.

H&Y Hoen and Yahr.

HC Healthy Controls.

HED Heavy Episodic Drinking.

HNR Harmonic to Noise Ratio.

ID Intensity Difference.

IEDCC Instantaneous Energy Deviation Coefficients.

IG Information Gain.

IPVS Italian Parkinson's Voice and Speech Corpus.

KNN k-Nearest Neighbors.

L-DOPA Levodopa.

LDA Linear Discriminant Analysis.

LES Lower Esophageal Sphincter.

LLE Largest Lyapunov Exponent.

LOSO Leave One Subject Out.

LPC Linear Prediction Coding Coefficients.

LPCC Linear Prediction Cepstral Coefficients.

LUHS Lithuanian University of Health Sciences.

MDS-UPDRS Movement Disorder Society-sponsored UPDRS.

MFCC Mel-Frequency Cepstral Coefficients.

ML Machine Learning.

MPT Maximum Phonation Time.

mRMR Minimum Redundancy Maximum Relevancy.

mRMRS Spearman Coefficient mRMR.

NA Non-Intoxicated.

NB Naive Bayes.

NNE Normalized Noise Energy.

NREM Non-Rapid Eye Movement Sleep.

OP Obese Patients.

OPR Obese Patients with Concomitant GERD.

OSA Obstructive Sleep Apnea.

PD Parkinson's Disease.

PDP Patients with Parkinson's Disease.

PET Positron Emission Tomography.

PLP Perceptual Linear Prediction Coefficients.

PPI Proton Pump Inhibitors.

PPQ5 Five-point Period Perturbation Quotient.

PR Patients with GERD.

PSG Polysomnography.

PSQI Pittsburgh Sleep Quality Index.

PTS Pitch Transition Slope.

RAP Relative Average Perturbation.

RASTA-PLP Relative Spectral Transform -Perceptual Linear Prediction.

RBD REM sleep Behaviour Disorder.

REM Rapid Eye Movement.

RF Random Forest.

RPDE Recurrence Period Density Entropy.

RST Rate of Speech Timing.

SD Subspace Discriminant.

SPECT Single Photon Emission Computed Tomography.

SPI Soft Phonation Index.

sPSQI Shortened Pittsburgh Sleep Quality Index.

SQ Sleep Quality.

STE Short Time Energy.

SVM Support Vector Machine.

SWS Slow-Wave Sleep.

TKEO Teager-Kaiser Energy Operator.

TNR True Negative Rate.

TPR True Positive Rate.

TR Transient Regions.

UCI University of California Irvine.

UPDRS Unified Parkinson's Disease Rating Scale.

VTI Voice Turbulence Index.

WA Web Application.

WHO World Health Organization.

XGB Extreme Gradient Boosting.

ZCR Zero Crossing Rate.

Chapter 1

Introduction

1.1 Overview and Motivation

In the contemporary landscape of healthcare, the integration of Artificial Intelligence (AI) emerged as a transformative paradigm, offering unprecedented potential to change the provision of medical services. AI-based solutions received significant attention due to their capacity to enhance diagnostic accuracy, remote monitoring, and personalized patient care.

The frontiers of technological progress in this area are continually expanding and are now reaching domains that were previously deemed accessible only to human experts. This expansion is notably attributable to the widespread availability of wearable sensors and devices which can be leveraged for the collection of physical signals serving as data sources for AI algorithms. In addition to the historically employed chemical, physiological, or electrical inputs, vocal signals, acquired through the recording of spoken tasks by individuals, are gaining importance as machine learning-based voice analysis progresses and researchers explore the effects of various pathologies on voice characteristics. Beyond the potential of vocal samples to reveal physio-pathological information, such an approach offers non-invasive, real-time, and cost-effective assessments, which could be of crucial importance in various stages of diseases'study.

Voice production, defined as the process of translating thoughts into audible sound, involves several stages. These latter encompass a *conceptual stage*, where the idea to be expressed is identified in an abstract form; a *syntactic stage*, where a

specific structure for expressing the idea is selected; a *lexical stage*, where speech units are chosen and organized; a *phonological stage*, where abstract information is transformed into a speech-like form; and a *phonetic stage*, where the selected sentence is ultimately converted into vocal emissions through a series of instructions sent to dedicated anatomical regions [1, 2]. The execution of this process involves the interaction of numerous systems and sub-systems (e.g., lungs, glottis, oral cavity, nasal cavity, trachea), all governed by brain activity. In this burgeoning field, vocal analysis harnesses the computational capabilities of AI to analyze and interpret vocal patterns, providing comprehensive insights into various aspects of an individual's health conditions.

In this context, this dissertation represented a comprehensive investigation into multifaceted applications of vocal analysis within the healthcare domain. These applications encompassed a spectrum that includes the early detection and monitoring of neurological disorders, specifically Parkinson's disease (PD), along with the assessment of conditions such as Gastro-Esophageal Reflux (GERD) and Obesity. In addition, the analysis of temporary fluctuations in overall health status arising from factors like poor sleep quality or alcohol intoxication were also studied.

1.2 Objectives and Significance of The Study

The primary objective of this dissertations was to develop robust and effective methodologies for the analysis of vocal signals, with the ultimate goal of creating models that can assist both medical professionals and patients in diagnosing and monitoring various medical conditions.

The core of this research focused on applications related to PD, where AI-based speech analysis proved exceptionally effective, given that nearly 90% of affected individuals presents significant alterations in speech production. Furthermore, it is essential to note that PD exhibits a prodromal phase characterized by ongoing neurodegeneration, that can arise up to 10 years earlier than cardinal motor manifestation. Within this context, the studies here described encompassed the entire pipeline of vocal analysis, from data collection to the development of automated analysis models. Comparative analysis between professional recording equipment and cost-effective alternatives like microphones embedded in smartphones were performed together with the exploration of unsupervised data collection environments. The analysis also

involved the investigation of different speech tasks employed solo or fused together, including sustained phonation, isolated words, sentence, or text reading. Additional crucial objectives involved the identification of relevant biomarkers for PD speech analysis, the examination of the influence of external co-factors, and the development of robust methodologies to support clinical practice in the realms of PD diagnosis, monitoring, and follow-up.

A significant portion of the study was dedicated to the analysis of acoustic parameters. This involved an extensive literature review to identify the most effective acoustic features, comparisons of algorithms used for parameter extraction, and the exploration of new acoustic measures and analysis techniques. The research also investigated the dependency of these parameters on speaker characteristics, such as language and the severity of the condition, as well as the recording setup. Throughout these analyses, statistical techniques and automatic classification algorithms were tested and compared to provide valuable insights into algorithm effectiveness and propose novel pipelines for the analysis of speech samples of patients with PD (PDP).

Beyond PD, this dissertation also investigated the effects of concurring pathologies on the intricate process of speech production. Specifically, the study examined GERD and obesity, analyzing both the effects of the two isolated diseases on the vocal signal and those arising from their concurrent presence. Indeed, given the complexity of the speech production process, involving various anatomical regions, the simultaneous presence of two or more alterations may lead to distinct and non-linear characteristics of the generated signals. These latter, if properly studied, can contribute to a more precise assessment of the patient's health condition.

With a similar objective, preliminary findings demonstrating a potential correlation between speech and poor sleep quality are presented to shed light on how temporary conditions may alter the speech alterations typically associated to more persistent diseases.

In its final sections, this dissertation examined the impact of temporary alterations stemming from alcohol consumption on speech signals. In-depth analyses were also conducted in order to test the feasibility of the automatic identification of the altered state through a model which is at the same time independent from subject's characteristics and the specific task performed. Also, due to the increased numerosity of the dataset employed for this latter task, significant attention was dedicated to exploring the efficacy of a deep learning architecture, aiming to a more resilient

and adaptable model, capable of handling the increased complexity and variability present in the dataset. This investigation lays the foundation for assessments of psychological changes, particularly to be used in-car contexts, where monitoring such alterations is of paramount importance.

1.3 Thesis Related Publications

1.3.1 Journal

1. Robust and language-independent acoustic features in Parkinson's disease / Scimeca, Sabrina; Amato, Federica; Olmo, Gabriella; Asci, Francesco; Suppa, Antonio; Costantini, Giovanni; Saggio, Giovanni. - In: FRONTIERS IN NEUROLOGY. - ISSN 1664-2295. - 14:(2023). [10.3389/fneur.2023.1198058]

2. Deep learning and wearable sensors for the diagnosis and monitoring of Parkinson's disease: A systematic review / Sigcha, Luis; Borzi', Luigi; Amato, Federica; Rechichi, Irene; Ramos-Romero, Carlos; Cárdenas, Andrés; Gascó, Luis; Olmo, Gabriella. - In: EXPERT SYSTEMS WITH APPLICATIONS. - ISSN 0957-4174. - ELETTRONICO. - 229, Part A:(2023). [10.1016/j.eswa.2023.120541]

3. Machine learning- and statistical-based voice analysis of Parkinson's disease patients: A survey / Amato, F.; Saggio, G.; Cesarini, V.; Olmo, G.; Costantini, G.. - In: EXPERT SYSTEMS WITH APPLICATIONS. - ISSN 0957-4174. - 219:(2023), p. 119651. [10.1016/j.eswa.2023.119651]

4. Artificial Intelligence-Based Voice Assessment of Patients with Parkinson's Disease Off and On Treatment: Machine vs. Deep-Learning Comparison / Costantini, Giovanni; Cesarini, Valerio; Di Leo, Pietro; Amato, Federica; Suppa, Antonio; Asci, Francesco; Pisani, Antonio; Calculli, Alessandra; Saggio, Giovanni. - In: SENSORS. - ISSN 1424-8220. - 23:4(2023). [10.3390/s23042293]

5. How resistant are levodopa-resistant axial symptoms? Response of freezing, posture and voice to increasing levodopa intestinal infusion rates in Parkinson's disease / Imbalzano, Gabriele; Rinaldi, Domiziana; Calandra-Buonaura,

Giovanna; Contin, Manuela; Amato, Federica; Giannini, Giulia; Sambati, Luisa; Ledda, Claudia; Romagnolo, Alberto; Olmo, Gabriella; Cortelli, Pietro; Zibetti, Maurizio; Lopiano, Leonardo; Artusi, Carlo Alberto. - In: EUROPEAN JOURNAL OF NEUROLOGY. - ISSN 1351-5101. - ELETTRONICO. - (2022). [10.1111/ene.15558]

6. Speech impairment in Parkinson's disease: acoustic analysis of unvoiced consonants in Italian native speakers / Amato, F.; Borzi', L.; Olmo, G.; Artusi, C. A.; Imbalzano, G.; Lopiano, L.. - In: IEEE ACCESS. - ISSN 2169-3536. - ELETTRONICO. - 9:(2021), pp. 166370-166381. [10.1109/ACCESS.2021.3135626]

7. An algorithm for Parkinson's disease speech classification based on isolated words analysis / Amato, Federica; Borzì, Luigi; Olmo, Gabriella; Orozco-Arroyave, Juan Rafael. - In: HEALTH INFORMATION SCIENCE AND SYSTEMS. - ISSN 2047-2501. - ELETTRONICO. - 9:32(2021), pp. 1-15. [10.1007/s13755-021-00162-8]

1.3.2 Conference

1. Hallmarks of Parkinson's disease progression determined by temporal evolution of speech attractors in the reconstructed phase-space / Amato, Federica; Cesarini, Valerio; Pietrosanti, Luca; Costantini, Giovanni; Olmo, Gabriella; Saggio, Giovani. - (2023), pp. 270-274. (Work presented at MetroInd4.0&IoT 2023, Brescia (IT), 06-08 June 2023) [10.1109/MetroInd4.0IoT57462.2023.10180199].

2. Obesity and Gastro-Esophageal Reflux voice disorders: a Machine Learning approach / Amato, Federica; Fasani, Maria; Raffaelli, Glauco; Cesarini, Valerio; Olmo, Gabriella; Di Lorenzo, Nicola; Costantini, Giovanni; Saggio, Giovanni. - (2022), pp. 1-6. (Work presented at 2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Giardini Naxos, Messina, 22-24 June 2022) [10.1109/MeMeA54994.2022.9856574].

3. Sleep Quality through Vocal Analysis: a Telemedicine Application / Amato, F.; Rechichi, I.; Borzi', L.; Olmo, G.. - ELETTRONICO. - (2022), pp. 706-711. (Work presented at 2022 IEEE International Conference on

Pervasive Computing and Communications Workshops and other Affiliated Events, PerCom Workshops 2022, Pisa (ITA), 2022) [10.1109/PerComWorkshops53856.2022.9767372].

1.3.3 Submitted papers

1. Beyond Breathalyzers: AI-Powered Speech Analysis for Alcohol Intoxication Detection / Amato, F.; Cesarini, V., Olmo, G., Saggio, G., Costantini G.. - (2023) - EXPERT SYSTEMS WITH APPLICATIONS.

Chapter 2

Speech

2.1 Biological Foundations of Speech Production

The process of speech production is a complex and highly organized physiological phenomenon that relies on a continuous and precise interactions among several biological structures within the vocal tract and the central nervous system.

Although the detailed analysis of the anatomy of the sound-producing apparatus goes beyond the scope of this dissertation, it is worth providing a brief overview of its primary aspects, to offer a more complete vision of the intricate mechanism that underlies the process of speech production.

2.1.1 Anatomy of the Speech Apparatus

The anatomy of the speech apparatus consists of a remarkable set of structures responsible for human voice production. From a macroscopic perspective, it can be divided into three main functional units: the generation of air pressure, the regulation of vibration, and the control of resonators [3].

The essential airflow required for speech results from the functions of the respiratory systems that regulates the lung air pressure during a prolonged expiration phase and a short inhalation through the synergistic action of diaphragm and intercostal muscles.

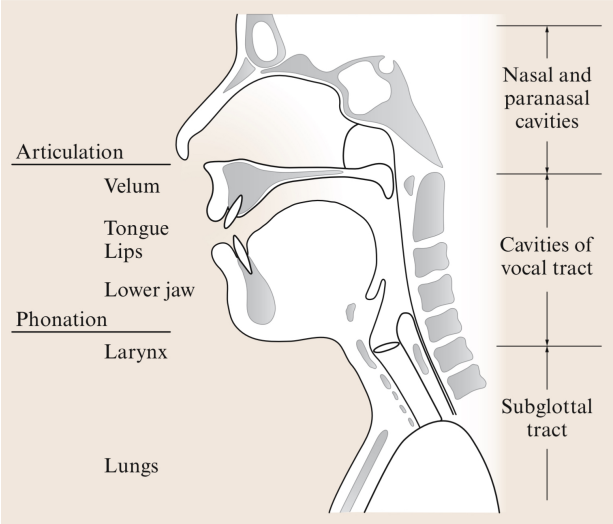


Fig. 2.1 Overview of the speech apparatus, sourced from[3]

The larynx is the central component of the phonatory system. It houses the vocal folds, which are made up of layers of muscle and connective tissue. Observing the larynx anatomy from an anterolateral perspective, the entire skeleton of the vocal duct is composed by cartilages, that ensure to the organ robustness and pliability. The cartilages are in turn interconnected by ligaments, that provide the elastic component. The overall structure is then completed by muscles that act on the movable cartilages to perform the various larynx tasks. These muscles are grouped in antagonistic pairs and can be divided into two classes: those controlling the glottis opening, and those regulating the tension of the vocal bands. The sectional view of the larynx reveals the vocal cords at the top of the trachea. They are composed of twin folds of mucous membrane placed horizontally across the larynx. Their outer edges are bounded to the laryngeal tube, whereas the inner margins are free to move [4]. Figure 2.2 reports a schematic of the vocal folds anatomy and their phases of vibration.

The actual resonant cavity is represented by a series of concamerations located above the vocal cords, namely the laryngeal vestibule, the upper portion of the pharynx, the mouth and the nasal cavities, all in all refereed as the articulatory system. The articulatory system includes various structures such as the tongue, lips, teeth, palate, and jaw that collaborate to shape the vocal tract into different configurations, thereby producing distinct speech sounds, whose single unit is referred to as *phoneme*. The positions and movements of these articulators are controlled by intricate neural networks.

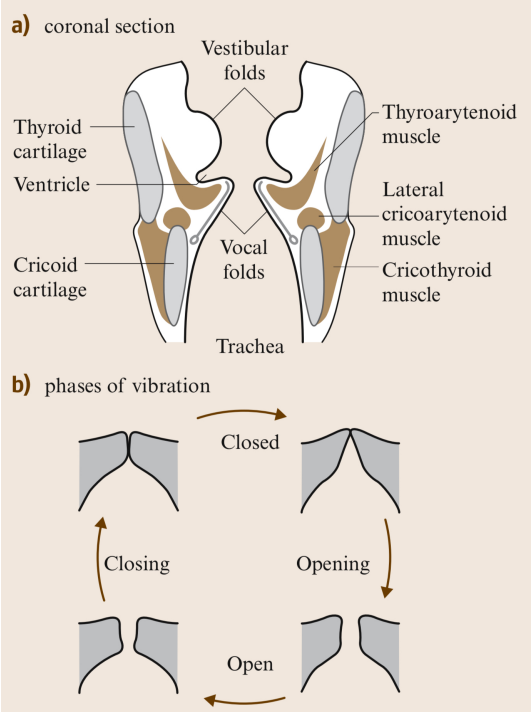


Fig. 2.2 Coronal sections of the vocal folds and their pattern of vibration, sourced from [5]

2.1.2 Neural Control of the Speech Apparatus

The central nervous system, primarily comprising the brain and the spinal cord plays a fundamental role in converting abstract linguistic concepts into articulate sentences through an intricate control of the speech apparatus.

Central to this process is the motor cortex, a region located in the frontal lobe of the brain mainly devoted to planning and executing complex sequences of muscle movements. During speech production, this area operates in close coordination with other brain regions to ensure the execution of motor programs required for vocal production [6].

At a cellular level, motor neurons located within the motor cortex are instrumental in the transmission of neural signals to the muscles involved in speech production. These neurons form synapses with the muscle fibers of various muscle groups, including those responsible of controlling the articulatory and phonatory systems. Through a finely tuned interplay of excitatory and inhibitory signals, these motor neurons thus control the contraction and relaxation of specific muscle groups and

allow for the precise and coordinated movements required to produce specific speech sounds.

In the realm of language production, Broca's area, situated in the dominant hemisphere of the brain, takes also center stage [7]. This region is mainly associated with higher-level speech planning and syntactic processing and serves a linguistic control center, ensuring that sentences are constructed with grammatical precision and syntactical coherence. Broca's area plays a crucial role in assembling the selected lexical items and syntactic structures into well-formed sentences, ultimately facilitating the fluency and intelligibility of speech.

2.2 The Speech Production Mechanism

2.2.1 Phonatory Mechanism

The term *Phonation* generally refers to the voice production process that occurs during the passage of air through the vocal cords, causing them to vibrate and produce sounds [8].

According to the classical theory of voice production, the respiratory pattern that is typically employed during quiet breathing undergoes a transformation during speech. This transformation involves a lengthened expiratory phase and a shortened inspiratory phase. Consequently, this alteration results in an airflow stream passing through the vocal apparatus, which serves as the driving force for speech execution.

The anatomy of the vocal folds placed into the larynx is then responsible for the actual voice production: their excitation, generally refereed to as *glottal excitation* [9], is indeed the fundamental element of vocal production, and can be voiced, unvoiced, or a mixture of both [10]. In the first case, the sound is produced by forcing air through the vocal folds, which vibrate and generate a quasi-periodic signal. In the second case, there is no vibration of the vocal folds, and the airflow arrives unaltered to the articulating elements

In further detail, when the underneath pressure originated from the lungs increases, if the vocal folds are abducted, they are forced to separate in order to let air flow; the high velocity immediately produces a lowered pressure due to the Bernoulli effect, which brings the vocal cords in the original position [11]. As a result of this

passage, the membranes start to vibrate generating sounds. The vocal fold vibrations repeats four phases within a cycle: *closed phase*, *opening phase*, *opened phase*, and *closing phase*. The frequency of the oscillation and the volume of the passing airflow are determined by the stiffness and mass of the vocal folds, the width of the glottal area, and the pressure difference across the larynx [3].

On the other hand, in case vocal folds remains in the closed phase for the entire duration of intended sound production, a turbulent flow characterized by a non-periodic behaviour is generated.

2.2.2 Articulatory Mechanism

The term *Articulation* generally refers to the precise movements of the articulatory structures to shape the vocal tract and generates phonemes, syllables, and words sequences.

The excitation source generated from the passage of air from the larynx propagates through the upper part of the vocal apparatus prior the emission of the final sound from the oral cavity. Depending on the reciprocal positions of the articulators the vocal apparatus presents different resonant properties that modulates the airflow thus leading to the production of different sounds [3, 11].

The tongue is the most important articulator organ characterized by intrinsic and extrinsic muscles that can deformate their shape and, as a consequence, the airflow prior to its emission. Deformation of the whole tongue determines the vowel quality and produces palatal and velar consonants. Moreover, depending on the reciprocal position between the tongue apex and the teeth, *dental* or *alveolar* consonants can be differentiated [3]. Similarly, based on the reciprocal position of the tongue with the upper jaw, dental, alveolar, and *palatal* consonants can be distringuished.

The lips and the velum also plays a pivotal role in speech sounds production. Under the influence of several muscles and other connected articulators, the lips can undergo three different deformation that eventually shape the vocal signal: opening/closing, rounding/spreading, protrusion/retraction. As for the velum, or soft palate, it controls opening and closing of the velopharingeal porta, thus allowing for the distinction between *nasal* and *oral* sounds [3, 11].

2.3 The Speech Signal

The acoustic pressure waveform constituting the human vocal signal is the result of the joint action of phonatory and articulatory mechanism. In this process, according to the classical *source-filter theory* [12], the lungs and the larynx act as a source generator, while the upper region of the vocal tract plays the role of an acoustic filter that modulates the source sounds and emits it through the lips.

For voiced sounds, the excitation source is a quasi-periodic train of airs pulses generated by the rapid oscillations of the vocal folds. Their frequency of vibration, typically referred as fundamental frequency (F0), is the lowest harmonic component of the vocal signal. It is characteristic of the single speaker, although it can be modified by an alteration in the tension of the vocal cords. Indeed, depending on the amount of energy imposed and on the air pressure generated underneath, the pulse frequency of the cords and their force of collision can vary, resulting in different intensities of the generated sound. Moreover, it is also the result of some anatomical characteristics which are dependent on the speaker's sex and age. The possible range of F0 in adult speakers is about 80–400 Hz for male speakers and 120-800 Hz for female speakers [3, 13]. As for unvoiced sounds, due to the static position of the vocal folds during their production, the associated produced signals is characterized by a non-periodic signal with a noise-like behaviour.

The supra-laryngeal tract acts as a time-varying filter for the excitation source with the articulators that continuously change their reciprocal position while speaking and thus resulting in time-varying resonant properties of the vocal tract. During this process, the different vocal tract shape let more acoustic energies through a set of formant frequencies, while attenuating others. Conventionally, F1 refers to the first formant, F2 to the second, and F3 to the third one [13].

Chapter 3

Automated Health Assessment Through Vocal Analysis

3.1 Rationale Behind Vocal Analysis

Speech production is a multifaceted process, reliant upon an intricate interplay of several articulatory and phonatory mechanisms (Section 2.1). Thereafter, any pathology that afflicts the vocal apparatus directly or through indirect system influences, can manifest as alterations in the generated speech signals. In the realm of clinical practice, these modifications can be measured, analyzed, and employed for diagnostic and monitoring support purposes through integration into machine learning algorithms (ML).

The fields of application are diverse and encompass the monitoring of various conditions, such as neurodegeneration [14–16], psychiatric and psychological disorders [17], cardiovascular problems [18], and vocal tract diseases [19] among the others. Recent evidence also suggests potential associations between eating disorders and voice alterations [20, 21] .

Neurological alterations can give rise to difficulties both in speech ideation and production, depending on the impaired area. In the former case, which is often more readily studied by asking the patient to engage in spontaneous speech, alterations may become perceivable as altered timing of vocal production that can manifest as an abnormal number of pauses and hesitations. Alterations in semantics may be also

possible, with recent studies showing that individuals affected by specific diseases tend to avoid terms related to their disability, such as PDPs avoiding verbs associated with movement [22] .

In cases where language ideation remains intact but the issue pertains to the execution of the required movements, the nature of the alterations varies depending on the affected region. As for muscular impairments, they can lead to issues related to both articulation and phonation, depending on the sub-districts involved. In the case of respiratory alterations, they result into reduced airflow generation, with consequences for intensity, sustainable vocal endurance, sub-glottal pressure, and glottal source control. Reduced or asymmetrical regulation of vocal fold vibrations can give rise to irregular and repetitive patterns; incomplete vocal fold closure can increase turbulent airflow, leading to additional noise within the produced sound.

Additionally, although certain diseases do not constitute explicit obstacle to overall language production, they can impede the execution of fine and precise movements. The lack of coordination mainly seen in various neurological diseases can give rise to phenomena like *voicing leakage* [23, 24]. In this case, patients with impaired glottal control face difficulties in interrupting vocal fold movements after the production of a voiced sound, resulting in partial vibrations, in lieu of an interruption of the phonation.

Another consequence is the *spirantization*, a speech impediment occurring due to incomplete vocal fold closure, causing air to escape during what should be a silent interval. This leads to noticeable distortions in unvoiced consonants, such as a /t/ sounding more like an /s/ [25]. In these cases, a comprehensive understanding of the complex interplay of neurological and physiological factors is crucial for assessing and managing vocal pathologies effectively.

Regardless the specific dimension being investigated, this innovative fusion of medicine and technology can provide clinicians valuable tools, facilitating the assessment and diagnosis of several pathologies. Furthermore, the technical advantages of vocal analysis are manifold: it is non-invasive, operator-independent, and necessitates minimal setup and little expensive instrumentation. Voice recordings can be made with diverse equipment, ranging from professional microphones and recorders to readily accessible devices such as smartphones or laptops. This flexibility enhances the accessibility of vocal analysis and its potential be employed for widespread remote-monitoring. This gives rise to a plethora of further benefits. Foremost among

these is the potential to increase the frequency of clinical assessments, which, at present, often occur once or twice a year, limiting the comprehensiveness of the clinical picture. Moreover, sporadic, in-person clinical evaluations frequently do not adequately account for the psychological aspect and emotional state of the patient, which can influence the assessment both negatively and positively. In this context, the increased monitoring frequency would enable more precise access to the patient's daily conditions, in a home environment and in the absence of external factors. Indeed, the accumulation of multiple measurements can minimize the influence of sporadic events, such as poor sleep quality, which, in the case of a single visit, could significantly impact the overall evaluation. Finally, if suitable biomarkers measuring vocal impairment are assessed, it is also possible to precisely quantify the source and level of alteration, allowing for a more accurate assessment of the patient's condition, a reduction in operator dependency, and increased precision from visit to visit.

3.2 Speech Analysis

From an engineering perspective, a speech signal is a complex waveform that carries information about the phonatory and articulatory settings of the vocal apparatus. Consequently, tracking the time-varying nature of the frequency content of the speech pressure waveform and deriving patterns that describe the type of variations is of pivotal importance in the field of speech research [11].

Within this context, several methods have been proposed to study the speech signal and extract objective features to quantify the information it contains. These methods encompass time-domain representation, spectrographic analysis, and cepstral analysis among the others. In this section, a detailed description of the principal techniques currently employed in the field of vocal analysis is provided, as well as the biomarkers typically extracted.

3.2.1 Periodicity Analysis

Speech signal is often characterized by its periodic nature. This periodicity is associated with the *Fundamental Frequency* (F0), often referred to as *pitch* (i.e., its perceptual counterpart). F0 represents the vibration of the vocal folds and can be therefore computed only for voiced sounds.

Short-term amplitude and temporal variation of F0, namely *Shimmer* and *Jitter* are generally employed to gather information about irregularites of periodic behaviour that may stem from a combination of factors, including biomechanical influences (such as vocal fold asymmetry), neurogenic alterations (involuntary actions of laryngeal muscles), and aerodynamic impairments (fluctuations in airflow and subglottal pressure) [3].

Shimmer is employed to detect changes in waveform amplitude within a period. According to [26], four different measures of this parameter have been largely applied in voice impairment detection: *Local Shimmer*, *Absolute Shimmer*, *APQ3* (three-point Amplitude Perturbation Quotient), and *APQ5* (five-point Amplitude Perturbation Quotient). Denoting as A_i the amplitude of the i_{th} frame and N the total number of frames (i.e., a section of the signal having length equal to a period), the first three parameters can be evaluated according to Equations 3.1, 3.2 and 3.3, respectively.

$$Shimmer(dB) = \frac{1}{N-1}\sum_{i=1}^{N-1}|20\log(A_{i+1}/A_i)| \tag{3.1}$$

$$Shimmer(relative) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|(A_{i+1}-A_i)|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \tag{3.2}$$

$$APQ3 = \frac{\frac{1}{N-2}\sum_{i=2}^{N-1}|0.5\,(A_{i-1}+A_{i+1})-A_i|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \tag{3.3}$$

APQ5 is computed similarly to APQ3 but takes into account the four closest neighbors.

Similarly, Jitter quantifies the variation of F0 among subsequent frames. Three different measures of Jitter are generally evaluated according to Equations 3.4, 3.5, and 3.6, representing *Absolute Jitter*, *Relative Jitter*, and the *Relative Average Perturbation* (RAP), respectively. The variable T_i in the formulas corresponds to the inverse of F0, whereas N is the total number of frames. Similar to the case of Shimmer, the *Five-point Period Perturbation Quotient* (PPQ5) can be calculated as the RAP while considering the influence of the four nearest neighbors.

$$Jitter(absolute) = \frac{1}{N-1}\sum_{i=1}^{N-1}|(T_i-T_{i+1})| \tag{3.4}$$

$$Jitter(relative) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|(T_{i+1} - T_i)|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \tag{3.5}$$

$$RAP = \frac{\frac{1}{N-2}\sum_{i=2}^{N-1}|0.5\,(T_{i-1} + T_{i+1}) - T_i|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \tag{3.6}$$

Moreover, to gather more insights into the temporal variation of F0, classical statistical moments are generally employed. Among these, the standard deviation of F0, namely *Monopitch*, is typically computed to quantify the speaker's ability to maintain a consistent frequency of vocal fold vibration throughout sustained phonation. This measure is particularly relevant in various pathologies, such as PD, where vocal fatigue often leads to difficulties in mantaining a steady phonation.

3.2.2 Noise Analysis

Noise analysis in speech processing is a critical aspect of understanding and quantifying the voice quality of the speakers, as it is mainly due to incomplete closure of the vocal folds [24].

While the computation of noise components can be challenging due to the complex nature of real-world signals which often lack clear separation between speech constituent units, several features have proven effective in characterizing non-normophonic speakers.

Harmonic to Noise ratio (HNR) is a measure based on the assumption that a speech signal consists of a periodic component and additive noise. It calculates the ratio between the energy of the periodic structure and the energy of the additive noise, which is influenced by voice impairments. Similarly, *Glottal to Noise Excitation Ratio* (GNE) quantifies the ratio between excitation originating from the vocal folds and excitation caused by turbulence in the speech signal. *Normalized Noise Energy* (NNE) computes the noise energy within each F0 period and estimates the overall contributions as the sum of non-harmonic portions of the voice spectrum [27].

Voice Turbulence Index (VTI) and *Soft Phonation Index* (SPI) involve predefined bandwidths. VTI measures the average ratio of energy in the 2800-5800Hz and 70-

4500Hz frequency bands, providing information about turbulence due to incomplete vocal fold closure. SPI measures the average energy ratio in the 70-1600Hz and 16000-4500Hz bands, capturing information about the closing phase of the vocal folds [24].

3.2.3 Spectral Analysis

Spectral analysis plays a crucial role in the field of speech analysis for characterizing phonation and articulation in both normophonic and non-normophonic speakers.

Primarily, researchers have extensively examined the speech spectrum using measures of *Formant Frequencies*. These latter represent the resonance frequencies of the oropharyngeal tract and provide valuable insights into the positions of the tongue, jaw, and lips during speech production [8]. Generally, the first three formants together with their bandwidths are employed.

Furthermore, classical parameters, including *mean*, *variance*, *skewness*, *kurtosis*, *crest*, *flux*, and *roll-off point* among the others, are usually employed to synthesize the overall behavior of the speech signal in the spectral representation. For instance, the more the spectrum exhibits a flat trend, the more the speech signal can be likened to white noise. Consequently lower values of spectral flatness may suggest increased noise due to incomplete closure of the vocal folds.

Energy measures are frequently utilized in speech analysis, often in the form of *Short Time Energy* (STE), which is employed to capture fluctuations in the energy contour of a speech signal over time. Additionally, energy ratios between voiced and unvoiced regions are employed to quantify the accuracy of vocal fold vibrations when a task with a predefined prompt is presented.

3.2.4 Cepstral Analysis

Cepstral Analysis is a widely employed techniques that allow separation of the effects of the vocal tract and excitation in speech processing [11]. The method is based on the fundamental assumption of the *source-filter*, which views speech as the result of convolving an excitation source with the vocal tract filter. It relies on two key mathematical properties: the *convolution in the time domain corresponds*

to multiplication in the frequency domain, and *the sum of the logarithms of two numbers is equal to the logarithm of their product.*

To apply cepstral analysis, the speech signal is first Fourier-transformed, which inherently involves two convolved signals, then the logarithm is applied. Transforming back to the time domain allows for the separation of the signal into its excitation and vocal tract components (Equation 3.7-3.11) [11].

$$y(t) = x(t) * h(t) \tag{3.7}$$

$$Y(f) = X(f) \cdot H(f) \tag{3.8}$$

$$log(Y(f)) = log(X(f) \cdot H(f)) \tag{3.9}$$

$$log(Y(f)) = log(X(f)) + log(H(f)) \tag{3.10}$$

$$\mathfrak{F}^{-1}\{log(Y(f))\} = \mathfrak{F}^{-1}\{log(X(f))\} + \mathfrak{F}^{-1}\{log(H(f))\} \tag{3.11}$$

The prominence of the main peak in the cepstrum, namely *Cepstral Peak Prominence* (CPP), has proven effective in characterizing both normophonic and non-normophonic speakers. CPP quantifies the prominence of the primary harmonic, corresponding to F0. Higher CPP values indicate a more distinct peak in the cepstrum, which can be associated with clearer and more stable pitch information.

In the specific case of vocal signal, an additional transformation is introduced to enhance the machine capability to deal with non-linearities, as in the human auditory system. This is achieved by mapping the power spectrum onto the *Mel scale* using a filter bank composed of overlapping triangular windows. The center frequencies and bandwidths of these windows are determined by a constant Mel-frequency interval [28].

This new frequency scale, known as the Mel scale, exhibits linear dependence from the traditional frequency scale when values are below the 1 kHz threshold and logarithmic dependence otherwise (Equation 3.12). This transformation aims

to replicate the behavior of the human cochlear region, where the perception of pitch is not linearly related to physical frequency. Indeed, the Mel is a unit of pitch defined such that pairs of sounds that are equally spaced in the perceived frequency domain are separated by an equal number of Mels. This transformation enhances the representation of speech features in a way that aligns more closely with human auditory perception.

$$f_{Mel} = \begin{cases} f & \text{if } f \leq 1 \text{ kHz} \\ 2595 \cdot \log\left(1 + \frac{f}{700}\right) & \text{if } f > 1 \text{ kHz} \end{cases} \tag{3.12}$$

Indeed, as stated in the Weber-Fechner law (Equation 3.13), in human systems the perceived intensity from a stimulus is not linearly proportional to the actual intensity generated. Instead, the ratio between these two quantities gives rise to a statistical parameter known as the *Weber proportionality term* (K). The concrete evidence supporting this physical formulation lies in the observation that loudness perception is not constant. Humans exhibit greater sensitivity to lower-frequency regions due to the fact that frequency resolution decreases as frequency increases.

$$\Delta R_{perc} = K \cdot R \tag{3.13}$$

In this context, the *Mel-frequency cepstral coefficients* (MFCC) are defined as the discrete cosine transform (DCT) of the log of the Mel spectrum. These coefficients, which enhance the machine capability to mimic effectively the behaviour of the human ear, provide a compact representation of the short-term spectrum and have been demonstrated to be effective in modeling irregular movements within the vocal tract [24]. Typically, the feature vectors used for analysis include the original MFCCs, their delta coefficients, and delta-delta coefficients concatenated together. Indeed, first-order derivatives, or ΔMFCCs, offer insights into the speed of spectral features, capturing short-term variations in the signal. On the other hand, second-order derivatives, or ΔΔMFCCs, delve into the acceleration of spectral features, providing information about higher-level dynamics.

3.2.5 Linear Prediction Analysis

Linear prediction analysis of speech samples is a widely employed technique in the field of speech processing and analysis [11]. This approach is based on the premise that a speech signal can be effectively represented as a linear combination of its past values and can be conceptualized as the outcome of applying a filter to the excitation source. The coefficients of these filters, referred to as *Linear Prediction Coding Coefficients* (LPC), are designed to emulate the behavior of the vocal tract. Consequently, they can serve as valuable tools for achieving a precise characterization of the entire vocal apparatus [24, 11, 29].

Given the set of LPC coefficients, it is possible to further improve the representation by applying the Cepstrum transformation. This latter allows to incorporate information about the human auditory system response to sound, yielding what are known as *Linear Prediction Cepstral Coefficients* (LPCC) [24].

Furthermore, if a proper adequate compression and a smoothing step are applied to the speech signal, *Perceptual Linear Prediction Coefficients* (PLP) are obtained. Similarly to MFCC, these coefficients are designed to model the perception of sound by the human auditory system [24, 29].

One noteworthy variant of PLP is *Relative Spectral Transform - Perceptual Linear Prediction* (RASTA-PLP). RASTA-PLPs are specifically engineered to enhance the robustness of PLP features, particularly in the presence of noise or adverse acoustic conditions. Its core innovation lies in the application of RASTA (Relative Spectral Transform) filtering to the PLP feature vectors. RASTA filtering employs a straightforward temporal filtering operation that effectively smooths the PLP coefficients within each frame independently. The term *relative* in RASTA-PLP denotes that this filtering operation subtracts the mean value of each PLP coefficient within a frame from the coefficient itself [30, 24].

3.2.6 Complexity Analysis

The presence of non-linear phenomena during voice production, primarily stemming from pressure flow in the glottis, stress-strain characteristics of vocal fold tissues, and vocal fold collision is a typical aspect of physiological and pathological speech. However, these complexities may be further exacerbated by compensatory move-

ments commonly observed in patients affected by speech disorders, who, conscious of their condition, try to hidden their motor dysfunctions. To analyze these non-linear phenomena and quantify related speech impairments, vocal signals are often represented in the *state space* and parameterized by means of specific features able to measure the complexity of the system.

The reconstruction of the speech signal in the state space is typically performed by means of an embedding procedure. The embedding theory proposed by Takens et al. in [31] represents the set of *diffeomorphic attractors* in the state space by the Equation 3.14, which is the solution of a system of nonlinear differential equations that define the speech production process.

$$X(k) = x(k), x(k+\tau), ...(x(k+(\theta-1)\tau) \tag{3.14}$$

In Equation 3.14, the variable X represents a collection of points within the attractors, denoted as $X(k)$. The signal $x(k)$ corresponds to the original time signal, while τ is the time delay, which is estimated to ensure minimal correlation among state variables. Additionally, θ stands for the dimension of the embedding space [32]. The determination of the optimal τ value, suited for reconstructing the vocal signal effectively, is typically achieved through a method relying on mutual information. This method defines the time delay as the point where the mutual information function exhibits its first minimum [32, 24].

Among the most common features employed to quantify the complexity and the non-linearity of the vocal signal starting from the reconstructed attractors in n dimensions, we mention: (i) *Correlation Dimension* (D2), *Largest Lyapunov exponent* (LLE), (iii) *Hurst Exponent* (H), (iv) *Recurrence Period Density Entropy* (RPDE), and (v) *Detrended Fluctuation Analysis* (DFA).

D2 quantifies the system complexity by assessing the self-similarity of an embedded attractor. In chaotic systems, D2 tends to be higher, indicating a larger-dimensional subset of the system state space related to the speech signal [33, 24].

LLE provides insights into a system sensitivity to initial conditions by calculating the average divergence rate of neighboring trajectories. Chaotic systems typically exhibit positive Lyapunov exponents, whereas non-chaotic systems often have zero or negative exponents [33, 24].

H measures long-term dependencies in a time series, reflecting how past values influence future ones. Values of H > 0.5 indicate strong long-term dependencies, while H < 0.5 suggests reduced persistence of signal characteristics. H = 0.5 signifies a complete lack of correlation between past and present values, as in the case of Brownian motion [33, 24].

RPDE assesses signal irregularity and changes in vocal fold oscillation periodicity. It examines the density distribution of recurrence periods, with higher RPDE values indicating more complex and less regular trajectories in the state space, characteristic of chaotic and less predictable systems [34, 33, 24].

DFA is employed to analyze the stochastic and fractal properties of a signal, helping to understand the underlying dynamics of complex systems. In vocal analysis, DFA can reveal deviations from periodicity and the influence of stochastic components, such as aspiration noise. DFA = 0.5 corresponds to highly chaotic and random systems, while DFA values smaller or greater than 0.5 indicate correlations and strong self-similarities, respectively [34, 33, 24].

3.2.7 Timing Analysis

Timing Analysis of complex speech tasks such as monologue and sentence reading, plays a crucial role in distinguishing between normophonic and non-normophonic speakers. Indeed, rhythmic organization contributes significantly to speech fluency and smoothness with deviations from typical rhythmic patterns denoting underlying speech disorders or neurological conditions.

To effectively characterize speech, particularly when tasks with predefined prompts are employed, various ease-to-compute metrics have been proposed. Among these, spectral moments (e.g., mean, median) extracted from voiced and unvoiced intervals lengths within a long text can quickly provide insights into abnormal speech patterns [35, 36]. Additionally, the *Rate of Speech Timing* (RST), which quantifies the rate of voiced, unvoiced, and paused intervals, is also a valuable metric. It is calculated as the slope of the regression line of total interval count over time and can help quantify reduced intervals due to impaired muscle control within the vocal apparatus [37].

As for sustained phonation tasks, strong correlation to vocal fold impairment is yielded by*Maximum Phonation Time* (MPT), which measures the total duration of

sustained phonation and can reveal airflow insufficiency if speakers are instructed to sustain the task as long as possible [36].

In addition to analyzing voiced and unvoiced regions, the *Duration of Pause Intervals* (DPI) is very important in speech characterization. Prolonged initial pauses may indicate difficulties in initiating speech, while frequent pauses, especially in free speech, can suggest challenges in speech conceptualization or vocal fatigue due to complex speech impairments.

3.3 Tools for Speech Analysis

To offer a comprehensive overview of the most common approaches used for evaluating acoustic features from speech samples, the following section provides a brief description of the available libraries and software tools. Additional, more detailed information can be found in the associated original papers.

- *Praat.* It is a C-based software package designed for speech analysis, enabling the extraction of diverse metrics. It is possible to integrate this software into Python using the Parselmouth library [38], [39].
- *Dysarthria Analyser.* It is a system that conducts automated acoustic analysis of different dysarthric speech patterns, employing specific algorithms to detect features extracted from tasks such as sustained phonation, syllables, and connected speech [35, 40].
- *pyAudioAnalysis library.* It is a comprehensive Python library designed for both feature extraction and the creation of classifiers, as well as facilitating automatic segmentation [41].
- *DARTH Voice Analysis Toolbox (DARTH-VAT).* It is a toolbox operating within MATLAB which has been primarily validated in settings involving the sustained vowel /a/. It predominantly comprises features related to F0, Jitter, Shimmer, MFCC, RPDE, DFA, and glottal modeling [42–44].
- *BioMetR©Tools.* It is a graphical user interface (GUI)-based toolbox used for extracting high-level glottal features through the modeling of speech signals [45].

- Voice Sauce. It is a MATLAB toolbox designed for extracting frequency and, more notably, harmonic-related content from audio signal [46, 47].
- *OpenSmile.* It is a comprehensive software developed by Audeering, enabling the extraction of over 6000 parameters according to custom configuration files[48–50].
- *Neurospeech.* It is a software platform specifically designed for conducting speech analysis on individuals with neurodegenerative disorders, with a particular focus on Parkinson's Disease. It calculates various measures to assess phonation, articulation, prosody, and intelligibility [51].
- *APARAT Toolbox.* It is a software package designed for use within the MATLAB environment. It incorporates glottal inverse filtering and multiple time-based parameters of the voice source, all presented through a user-friendly graphical interface [52].
- *Analysis of dysphonia in speech and voice (ADSV).* It is a software designed to extract dysphonia-related spectral and cepstral features from various speech tasks, such as sustained phonation, sentences, or syllables.

3.4 Challenges and Limitations

Properly validated speech-based AI tools hold promise in mitigating potential subjectivity biases and providing insights into the speaker's health status. However, the analysis and parameterization of speech samples, as complex signals, are not without limitations.

3.4.1 Variability in Vocal Characteristics

The primary limitation to consider pertains to the variability of the vocal signal. As discussed in the previous sections (Sections 2.1, 3.1), vocal production is the result of the intricate coordinated activity of various anatomical regions. While the analysis of this complexity allows for the extraction of objective parameters related to specific pathologies, it also embeds the influence of individual-specific characteristics on the produced signal as well as time changes within a single-speaker.

Notably, a critical factor is the speaker's gender, as vocal mechanics are directly dependent on vocal cord size and length, resulting in distinct acoustic features between genders, typically characterized by higher frequencies in females compared to their male counterparts. Similarly, speaker's age significantly affects the produced signal. Consequently, although the use of highly-dimensional datasets can potentially mitigate these factors, the typically limited size of biomedical corpora requires careful consideration of these aspects or appropriate stratification.

In addition to physiological characteristics, further confounding factors arise from the co-occurrence of multiple pathologies, whose impact on the vocal signal may vary depending on the specific region being analyzed. Consequently, while research studies usually focus on carefully selected participant populations by excluding individuals with concurrent pathologies that could potentially impact the collected data, this approach is commonly restricted to specific categories of disorders, often overlooking individuals with less severe conditions. On the contrary, recent studies have shown that even pathologies with a minor impact on the overall health status (e.g., gastro-esophageal reflux, obesity...) or even temporary alterations (e.g., emotions fluctuations, sleep quality disturbances..) can exert a noticeable influence on the produced signal. Thereafter, it is of crucial importance the necessity of thoroughly analyzing the nature of the alterations or developing algorithms that can account for such variability.

Finally, though the influence is limited to the analysis of more complex language production tasks, the educational level of the subjects also warrants consideration. Differing levels of linguistic vocabulary can indeed lead to the production of texts with varying degrees of complexity, which are independent of the subject's health status.

It is also worth noting that, conversely to other biomedical signals used in conjunction with AI techniques, information derived from vocal samples are profoundly influenced by the speaker's language. Despite this limitation is partially mitigated when the analysis focuses on sustained vowels or speech timing, it becomes of fundamental importance when the focus of analysis revolves around specific sounds or linguistic phenomena unique to certain languages.

In the pursuit of robust vocal signal analysis, some key guidelines should be adhered to during the data collection process in order to reduce as far as possible the beforementioned criticalities. To address the intricate variability inherent in vocal

production, it is of paramount importance to ensure a diverse and representative participant pool that encompasses various demographics such as age, gender, and educational backgrounds. Longitudinal data collection strategies should be employed to capture temporal changes within individual speakers, allowing for a comprehensive understanding of vocal variability over time.
As for the impact of concurring pathologies, a more inclusive approach should involve incorporating participants with varying severity levels of conditions, enabling a comprehensive perspective on the impact of health-related factors on vocal signals. Moreover, while extracting language-neutral features could help minimize biases related to linguistic phenomena specific to certain languages and enhance the generalizability of findings, the analysis of language-specific sound should not be neglected. Within this context, the collaboration with linguists and language experts becomes paramount to gain insights into these aspects and align the analysis framework with linguistic nuances.
To conclude, it is crucial to emphasize that, while all the mentioned approches can contribute to an effective data collection process, investing in significantly larger biomedical corpora remains the most effective technique to minimize potential biases and contribute to better statistical power and reliability.

3.4.2 Influence of Recording Conditions

The influence of background noise and, more generally, the recording conditions is of crucial importance to derive significant and robust evidence from the analyses of vocal signals. Although it is possible to mitigate the extent of the influence by employing uni-directional microphones, which are capable of effectively capturing the speaker's voice while minimizing ambient noise, a significant portion of the literature in vocal signal analysis aims to leverage the ease of remote, low-cost, and unsupervised data collection. In such contexts, there is a strong increase in the likelihood of inappropriate microphone placement, continuous or sporadic background noise capture, and related issues. Moreover, smartphones and laptops typically embed omni-directional microphones, which are more likely to include also unwanted information.

To address these concerns, it is essential to educate users about proper recording procedures and develop applications capable of guiding subjects during data collection. Subsequently, during the data analysis and utilization phase, it is necessary to

apply appropriate pre-processing techniques and select parameters that are minimally influenced by background noise and recording conditions, while being aware that any pre-processing applied to the signal could potentially alter its characteristics and the extracted information.

Chapter 4

Application I: Parkinson's Disease

4.1 Parkinson's Disease

4.1.1 Incidence and Prevalence

Parkinson's disease (PD) is a chronic and progressive neurodegenerative disorder, which affects approximately 1% of individuals over the age of 60 [53]. The incidence of PD increases with age, being rare before the age of 50 and more common in men than in women [54, 55]. The global prevalence of PD is expected to rise significantly due to the overall aging population, increasing from 6.9 million cases in 2015 to an estimated 12 million cases by 2040 [56]. However, even when age-related factors are taken into account, PD's incidence is still projected to increase [57], indicating a more complex and as yet not fully understood scenario.

4.1.2 Pathophysiology

After the onset of PD, patients experience a progressive decline in their ability to perform Activities of Daily Living (ADL). This progressive disability is the result of a complex interplay of factors, including the aggregation of aberrant α-synuclein, dysfunction of cellular components like mitochondria, lysosomes, and vesicle transport, problems with synaptic transmission, and neuroinflammation [57].

The primary pathological characteristic of PD is the gradual loss of dopaminergic neurons in the substantia nigra pars compacta region of the midbrain. Clinical-pathological correlation studies have indicated that this ongoing degeneration is likely responsible for motor symptoms like bradykinesia, tremor, and rigidity, which are observed in both early and mid-advanced stage patients [58]. Although the exact cause of this neurodegeneration remains unclear, it is known to involve the formation of Lewy pathology, a result of abnormal aggregation of α-synuclein proteins. Indeed, in their misfolded state, α-synuclein becomes insoluble and forms intracellular inclusions within cell bodies [58]. Consequently, these aggregations can disrupt the normal functioning of the brain, ultimately causing dysfunction and death of particular neuronal groups [10, 58].

4.1.3 Etiology

The etiology of PD is still a matter of debate, however two main factors are acknowledged as relevant for PD onset: genetics and environment [57].

Genetic factors are estimated to contribute approximately 25% to the overall risk of developing PD and mostly interests mutations of the genes SNCA, LRRK2, PRKN, PINK, and GBA [59, 57, 58]. Despite PD presentation and progression is acknowledged to present a marked heterogeneity, it is still possible to identify some clusters based on the type of genetic alteration. SNCA mutations usually imply earlier age of disease onset and faster progression of both motor and non-motor symptoms. LRKK2 mutations accounts for the majority of familial PD cases but are also observed in non-genetic patients, thus increasing the complexity of a precise diagnosis. PRKN and PINK mutations are mainly responsible for juvenile PD and early non-motor symptoms with a generally slow progression which is rarely characterized by dementia. On the contrary, GBA-linked PD presents a more severe course with a rapid cognitive decline [57].

As for *Environmental factors*, they mainly include pesticide exposure, head injuries, rural living, *beta*-blocker use, and agricultural occupation. Additionally, despite the underline associations are still elusive, tobacco smoking, coffee drinking, non-steroidal anti-inflammatory drug use, alcohol consumption, and calcium channel blocker use were found to have a negative correlation with PD arising [58].

4.1.4 Symptoms

After PD arising, patients usually face a broad variety of symptoms that significantly impacts their ADLs. The cardinal and more evident manifestations include motor symptoms such as rigidity, tremor at rest, bradykinesia, and postural instability [60]. However, the clinical spectrum also contains many less visible components, including non-motor features, such as olfactory impairment, orthostatic hypotension, constipation, sleep disturbances, and vocal impairment. Behavioral problems, depression, and anxiety frequently occur, and dementia is quite common in the advanced stages of the disease [58]. In the following a list of the most frequent PD symptoms is reported.

- *Bradykinesia* refers to the slowness of movement and is considered to be the most characteristic clinical feature of PD. It involves challenges related to planning and executing movements, as well as difficulties with carrying out tasks sequentially or simultaneously. Other manifestations include reduced reaction times, difficulties in performing fine movements, and loss of spontaneous gesturing [60, 57, 61].

- *Tremor* in PD refers to involuntary muscle contractions. The most common and easily identifiable form is rest tremor, which typically involves shaking in the hands, lips, chin, jaw, and legs. This tremor is often unilateral, affecting one side of the body, and it usually occurs at frequencies between 4 and 6 Hz. Some PDPs may also experience postural tremor, which typically emerges when an individual assumes an outstretched horizontal position [60, 57, 61].

- *Rigidity* refers to an increased resistance to passive movements, which hinders the range of motion in terms of flexion, extension, or rotations. This rigidity can manifest proximally, affecting areas like the neck and shoulders, or distally, involving the wrists and ankles. Muscle rigidity in PD is often characterized by the emergence of pain, making it one of the most common and recognizable features of the condition, even though it can be challenging to perform an accurate differential diagnosis [60, 57, 61].

- *Postural deformities* are primarily linked to rigidity and encompass atypical axial postures that tend to develop in the advanced stages of PD. They may

involve conditions like scoliosis, alterations in the trunk (such as Pisa Syndrome), and deformities affecting the limbs due to the involvement of the striatum [60].

- *Freezing* is a type of akinesia, and along with postural instability, it is one of the primary factors contributing to falls in PDPs. This phenomenon typically occurs when initiating walking or during specific actions like making a turn or passing through narrow passages, resulting in a sudden and temporary inability to move [60].

- *Cognitive and neurobehavioural abnormalities* are believed to affect a significant proportion of PDPs and may manifests as cognitive decline, dementia, depression, or other neuropsychiatric comorbities. Furthermore, despite the inherent mechanism is not well understood, PDPs may exhibit features of obsessive-compulsive and impulsive behavior, which have been linked to the development of dopamine dysregulation syndrome due to the use of dopaminergic drugs [60].

- *Sleep alterations* can manifest in various forms, with REM Sleep Behavior Disorder (RBD) being the most common. RBD presents as a parasomnia characterised by lack of physiological muscle atonia during REM Sleep [62]; it is estimated to affect approximately 2% of the elderly population globally [63]. Recent research suggests that sleep disturbances are among the earliest prodromal symptoms of α-synucleinopathies, with RBD having a remarkable 90% conversion rate when observed over a 14-year follow-up period [64].

- *Speech alterations* in PD are often categorized as *hypokinetic dysarthria* and are primarily characterized by difficulties in articulation and breathing, along with a voice quality often described as trembly and unstable. Approximately 90% of PDPs experience these symptoms, with the first manifestations that can manifest up to a decade earlier than the cardinal motor symptoms of PD, making it one of the early prodromal signs of the condition [65–67].

4.1.5 Diagnosis and Complicating Factors

With the exception of genetic tests for individuals with a family history of PD, a definitive diagnosis of the disease can only be confirmed post-mortem through the

identification of neuropathological changes in the brain [57]. Nevertheless, clinical practice universally recognizes and applies a diagnosis based on a comprehensive neurological assessment, a thorough review of the patient's medical history, and a clinical evaluation of both motor and non-motor symptoms [68]. Positron Emission Tomography (PET) and Single Photon Emission Computed Tomography (SPECT) can also be employed to quantify the dopaminergic reduction in the substantia nigra pars compacta region of the midbrain [58]. However, the dopamine imaging approaches may not be sufficient for an accurate diagnosis since they do not allow for a reliable differentiation between PD and other parkinsonian syndromes [58].

The diagnostic process for PD typically begins with the observation of symptoms, and a confirmed diagnosis is usually only made after neuropathological examinations. This process can be complex due to the similarities and overlaps in signs and symptoms among different conditions with diverse underlying causes. Thereafter, in addition to assessing the classic motor symptoms, clinicians also consider potential clinical markers to better understand the patient's condition. These markers can encompass the presence of olfactory impairment, RBD alterations, as well as behavioral and vocal changes reported by the patient or observed by their caregivers [58]

PD monitoring and follow-up visits are typically conducted during scheduled medical visits, occurring every 6 to 9 months [69]. However, these periodic visits make it challenging for neurologists to detect short-term changes in a patient's condition. Additionally, despite validated protocols are employed, the subjective nature of clinical examinations can introduce bias into the assessment [70, 71]. These limitations make it difficult to implement appropriate therapeutic adjustments and can reduce the overall effectiveness of therapy [72].

4.1.6 Clinical Scales

The Unified Parkinson's Disease Rating Scale (UPDRS) [73] is the most widely used scale for diagnosis and monitoring of PD. It employs a structured scoring system that ranges from 0 (indicating no impairment) to 4 (representing severe impairment), with intermediate scores reflecting various degrees of severity. Its revised form, the Movement Disorder Society-sponsored UPDRS (MDS-UPDRS), is a comprehensive

clinical tool designed to evaluate the severity and progression of Parkinson's disease. It consists of four parts, each addressing different aspects of the condition.

- *Part I - Non-Motor Aspects of Daily Living.* This initial section is dedicated to evaluating the impact of PD on everyday activities that are not directly linked to movement. It encompasses assessments of cognitive function, behavior, mood, and ADLs.
- *Part II -Motor Aspects of Daily Living.* In this section, the focus shifts to assessing how PD impacts motor functions during daily activities. It involves evaluating elements such as speech, swallowing, handwriting, cutting food, dressing, personal hygiene, turning in bed, walking, and other routine tasks.
- *Part III -Motor Examination.* The third section guides trained clinicians to systematically evaluate the patient's motor function in this section. It encompasses various subdomains, including tremor, rigidity, bradykinesia, gait, posture, and other motor features.
- *Part IV -Motor Complications.* The last section places its focus on assessing motor complications that arise due to treatment, including dyskinesias (involuntary movements) and fluctuations in medication response. It evaluates the presence and impact of these complications on the patient's ADLs providing valuable information for treatment optimization.

The Hoen and Yahr (H&R) clinical scale [74] provides a simplified and practical assessment of PD progression, focusing primarly on motor symptoms and functional disability. It categorizes patients into five stages based on their motor symptoms and functional disability.

- *Stage 1.* In the initial stage, the patient exhibits unilateral involvement only. The symptoms are typically mild, encompassing features like tremor, rigidity, and bradykinesia; the patient's posture and balance remain unaffected.
- *Stage 2.* In stage 2, the disease affects both sides of the body. However, the patient can still maintain an upright posture and balance. Although symptoms become more pronounced, the patient can generally perform daily activities without significant impairment.

- *Stage 3.* At this stage, the disease further progresses, resulting in moderate to severe motor impairments. The patient experiences bilateral involvement and postural instability. Balance is compromised, and falls may occur, necessitating assistance. Despite these challenges, some independence in performing daily activities is retained.

- *Stage 4.* In stage 4, the symptoms becomes severe, significantly limiting the patient's mobility. Assistance or assistive devices may be required for walking, and independent living is no longer possible. However, the patient can still stand or walk without assistance.

- *Stage 5.* The final stage represents the most advanced and debilitating phase of PD. The patient is typically either wheelchair-bound, necessitating full-time support. Symptoms are extremely severe, and the patients may experience substantial fluctuations in their response to medication.

4.1.7 Treatment

Currently, there is no cure for PD, and the approach to treatment involves a multimodal strategy aimed at mitigating symptoms, preserving residual motor function, and improving the patient's quality of life. This comprehensive approach typically encompasses pharmacological treatment, possibly surgical interventions, and supportive therapies.

In terms of *pharmacological treatment*, the primary class of medications used includes dopaminergic agents. These medications are designed to supplement or replicate the diminished dopamine levels in the brain. Notably, drugs like Levodopa (L-Dopa) and dopamine agonists are considered the most effective treatments for alleviating the symptoms of PD [61, 75]. These medications are particularly effective in managing motor symptoms, especially in the early stages of the disease. However, it is important to note that they do not interrupt the process of neurodegeneration, disease progression, or the development of disability [76]. Furthermore, it is worth mentioning that prolonged use of these medications is related to motor complications, such as dyskinesias [77]. To manage specific symptoms or enhance the efficacy of L-Dopa, adjunctive medications like anticholinergics and catechol-O-methyltransferase inhibitors may be prescribed [78, 79].

Among *surgical interventions*, deep brain stimulation (DBS) is the most used surgical approach and it considered when medication alone prove insufficient in managing PD symptoms or when motor complications become problematic. The procedure involves implanting electrodes into specific brain structures, like the subthalamic nucleus or the globus pallidus internus, tailored to the patient's characteristics and symptoms [80]. Advanced imaging techniques guide electrode placement; thereafter, once correctly positioned, they are connected to a pulse generator [81]. This generator delivers controlled electrical impulses to the brain, normalizing neural activity. Healthcare professionals fine-tune stimulation parameters to optimize symptom control. DBS significantly improves motor symptoms, reduces medication needs, and enhances overall quality of life. Some studies suggest it might even have a neuroprotective effect, potentially slowing disease progression [80, 81].

In addition to medication and surgery, *supportive therapies* play a crucial role in managing PD. Physical therapy aims to improve mobility, gait, balance, and overall physical function. Occupational therapy focuses on enhancing the patient's ability to perform ADLs and maintain independence. Speech and swallowing therapy is designed to address challenges in speech production and swallowing function, which are frequently encountered in PD. Additionally, exercise programs have shown beneficial effects in maintaining motor function and improving overall well-being [58].

4.2 Related Literature

The investigation of vocal changes associated with PD has gained growing interest in recent decades. Notably, various types of algorithms, ranging from simple to complex pipelines, have been proposed with excellent results for the automatic assessment of the presence of the disease and its staging.

In the realm of *acoustic features analysis*, several recent reviews studies investigated the current panorama of relevant parameters. Among these, the authors in [82] conducted a comprehensive investigation of computational approaches within the neurodegenerative spectrum, encompassing both motor and non-motor symptoms in PD. They underscored the complexity in selecting the optimal set of features due to diverse potential applications, including disease assessment and progression monitoring, along with the different dimensions of speech, such as phonation, articulation, or

prosody. Nevertheless, they identified a list of common features, including time and frequency measures. Gómez-García et al. in [83] reviewed prevalent measures used in automatic vocal analysis tools and summarized routines for their extraction in a freely available toolbox, namely AVCA. They highlighted amplitude perturbation, frequency perturbation, and noise measures as the most common features. Similar findings were reported in [84] and [85]. The former study delved into phonatory and prosodic changes in PD, exploring the pathophysiology behind vocal alterations and the features used to describe them. The authors reported that voice disturbances often resulted from larynx asymmetric rigidity and incomplete glottic closure, which was confirmed through laryngoscopy and stroboscopic investigations. Commonly used features included HNR, jitter, shimmer, intensity, and F0. In their review of 86 articles on the acoustic analysis of PD, Brabenec et al. [85] focused on early diagnosis, monitoring, functional imaging studies, and the impact of dopaminergic medication and brain stimulation. While conventional features like jitter, shimmer, and F0 remained prominent, it was noted that these features, while interpretable, may not be sufficient for complex mechanisms or advanced analysis. More extensive techniques such as DFA, D2, RPDE have proven effective. More recently, Moro-Velazquez et al. in [86] conducted an extensive analysis comparing articulatory and phonatory aspects. Evidence from 192 reviewed papers emphasized the significance of articulatory analysis, with commonly adopted features such as amplitude and frequency perturbation values, noise, complexity, timing features, and sets of coefficients like MFCC and LPC.

Regarding the *speech tasks* typically employed, the literature review revealed a predominant focus on phonation, particularly using the sustained phonation of the vowel /a/ , which is employed to assesses the ability to control the airflow from the lungs and the glottal source vibration [24, 87]. While sustained phonation is reliable, it lacks the complexity of natural speech. Some studies suggest that analyzing the articulatory process occurring during running speech, reading, or word repetitions can provide more comprehensive insights, with classification accuracy ranging from 80% to 95% [88**?** –90]. In addition to phonation and articulation, prosody impairments are also observed, affecting speech rate, pauses, and intonation [24]. However, these analyses may be influenced by external factors such as anxiety and alertness, which can impact PD symptoms, making recordings less reflective of real-life vocal changes [91]. The strong discriminatory potential of articulation and prosodic analysis comes at the expense of high complexity due to language-

dependent variations in phoneme pronunciation, with a universal set of phonemes strongly correlated with the disease yet to be established [92, 88].

Indeed, several studies explored the effectiveness of the different phonemes in specific languages. Among these, [**?**] demonstrated the potential of *fricatives* sounds parameterized by means of duration, intensity, and spectral moments to model the co-articulation capability and the ability of the patient to perform a complex sequencing of movements, which are typically impaired in PDPs [93]. In [94], the authors explored the use of *occlusive* consonants for early PD identification employing temporal and spectral parameters from voice-onset-time segments. Using a dataset of Spanish speakers, they achieved a classification accuracy of 94.4% in leave-one-out Cross Validation (CV), with the consonant /k/ exhibiting the highest discrimination capability. The importance of fricatives and occlusive were eventually confirmed in [23], in which the authors performed preliminary cross-language experiments employing Czech and Spanish participants yielding accuracy ranging from 72% to 94%. In a subsequent work [88], the authors introduced features extracted from relevant articulation moments, such as bursts, transitions between vowel and consonants, or the beginning and end of the glottal activity and demonstrated the importance of studying the transition between specific phonemes.

While many articles explored speech analysis in PD with participants from various nationalities, only a few works directly addressed the *influence of the language* of the speakers on the proposed models. Among these, in [95], the authors examined continuous speech samples from five nationalities (Czech, English, German, French, and Italian), identifying changes in voice quality, articulation, and speech speed. More recently, [96] conducted an analysis involving five datasets in five different languages (Italian, Hebrew, English, Czech, and Spanish). In this study the authors performed cross-language experiments achieving a 75% classification accuracy using an Extreme Gradient Boosting (XGB) classifier. However, the reported performance was based on a 10-fold CV, and no additional tests were conducted on previously unseen samples.

Beside language, an increasing number of works is currently studying the panorama of external factors that can potentially influence the analysis of speech samples, ranging from environmental factors to subjects-specific characteristics. Among these, the effect of medication on speech production is still a matter of debate, with results ranging from no effect at all [97] to a significant effect that however

depends on the off-medication speech disfluency [98]. Some authors also referred to differential alterations depending on the speech dimension and the phonemes specifically investigated [99].

Recent works have also highlighted the potential impact of recording modalities on the features extracted from collected speech samples. Indeed, several studies [100, 42, 101, 35] tested the feasibility of automatically assessing PD-related alterations by means of samples collected via smartphones. These studies generally agree that satisfactory results can be obtained even with low-quality recordings. However, only a few of them directly compared the same signals simultaneously recorded using both modalities. In this context, recent analyses also explored the influence of the recording modality, whether under the supervision of an operator or completely unsupervised, where participants record in their home environment. For instance, Carron et. al in [102] compared the performance achieved using data acquired with smartphones in supervised and unsupervised conditions. Their findings indicated that recording conditions have a more significant impact than the recording equipment itself, with unsupervised recordings leading to an overall performance decrease. Furthermore, cross-corpus experiments revealed a partial improvement when the algorithm is trained on the worst-quality dataset and then applied to the remaining samples.

Lastly, an open debate regards the choice between shallow ML models and more advanced approaches including deep-learning (DL) algorithms. As discussed in a recent review study [103], the difference is mainly conceptual and revolves around the differences in data-driven and model-driven nature. While DL has demonstrated its effectiveness, the use of low-interpretability models may raise concerns among clinicians who demand high-level evidence in clinical practice. This, in turn, can lead to overfitting issues and a lack of generalization, especially in presence of corpus with limited cardinality, which often characterize biomedical applications.

4.3 Corpora for PD voice analysis

In this section, an overview of the corpora utilized in this study is presented, both private and freely available, encompassing voice samples from PDPs. Each dataset is referred to by its original name as reported in the corresponding publication. In

cases where the original name is not available, the denomination is based on the name of the first author.

4.3.1 Italian Parkinson's Voice and Speech Corpus

The Italian Parkinson's Voice and Speech Corpus (IPVS) is a publicly accessible dataset distributed under the Creative Commons Attribution License (CC BY 4.0), as detailed in the work by Dimauro et al. [104]. This resource comprises vocal recordings from 65 Italian native speakers, categorized into three groups: 15 young Healthy Controls (HCs), 22 elderly HCs, and 28 PDPs. Importantly, none of the HCs reported any vocal or language disorders, and all PDPs were under their usual anti-parkinsonian treatment.

As for the disease severity, the majority of PDPs had a H&Y score below 4, with only a few exceptions: one classified as stage 5 and two as stage 4. Additionally, the MDS-UPDRS Part III (motor examination) scores were reported, with 11 patients at stage 0, 9 at stage 1, 5 at stage 2, 1 at stage 3, and 2 at stage 4. For a more comprehensive overview of participant demographics, please refer to Table 4.1.

Table 4.1 Demographic details of participants in the IPVS corpus. Measures are reported in terms of mean ± standard deviation. HC: Healthy Controls; PDP: Patients with Parkinson's Disease

	HC (young)	HC (elderly)	PDP
Age	20.8 ± 2.65	67.09 ± 5.16	67.21 ± 8.73
Gender	13M, 2F	10M, 12F	19M, 9F

All voice recordings were conducted under supervised conditions, employing professional microphones (16 KHz with 16-bit resolution) placed at 15-25 cm from the mouth in a quiet, echo-free environment. Participants were instructed to perform a series of vocal tasks, encompassing: (i) reading of a phonetically balanced text, (ii) articulation of the syllables /pa/ and /ta/, (iii) sustained phonation of the five vowels in Italian language, (iv) reading of a list of phonetically balanced words, and (v) reading of a list of phonetically balanced sentences. The chosen tasks were designed to be challenging, requiring participants to breath with effort and featuring complex phonetics in close proximity. For a comprehensive list of the speech exercises conducted, along with their English translations, please consult Table 4.2.

Table 4.2 List of tasks in the IPVS dataset with corresponding English translations

	Prompt (IPA Translation)	Translation to English
Text	iːl ramˈarɥo ˈdɛlːa dʒˈiːa.iːl papˈa (ʾɔː iːl babbo kˈome diʾtʃe iːl piʾkːoːlo dˈaːdo) ˈɛra sul ˈlɛto. soʾto di lui, akˈkanto ˈal ʾlaːgo, sedɛʾva dʒiʾdʒi, dɛtːo tʃˈitʃːo, ˈkɔːko della ˈmamːa eː ˈeː della ˈnɔnna. viʾtʃino ˈad un ˈsatsːo tʃiʾɛrɑ ˈuna rɔːza rɔsːo ʾvivo ˈeː lɔː ˈʃoːko, vedenʾdola, la ʾvɔlle per la ˈðʒia. la ˈðʒia luʾlʊ tʃerʾkava dʒanʾdzare per il suo ramaʾro, ma dato kɛ ˈɛra dʒuʾŋŋo (ʾɔː ʾuljo nɔn sɔː ʾbɛːne) nɔn nɛ troʾvava. troʾvɔː inʾvetʃ e ʾuna ʾrana kɛ salʾtando daʾla ʾstraːda finʾɪ ʾnɛl ʾlaːgo kon ʾun ʾgrande ʾsprutso. saj kɛ ʾfifa, la ʾðʒia! lɔ ʾskitso baŋŋʾɔː il suo komʾplɛto ʾrɔːza kɛ diʾvɛnne ʾdʒallo kɔːme ʾun ʾtaksi. pasʾsava di li ʾun siŋʾɔːre kozmoːpoʾlita di ʾnoːme ʾsardanapalo nabukodoʾnosoːr kɛ si innamoʾraː dɛlla ʾðʒia eː la ʾpɔrtɔː kɔn sɛː ʾin afgaʾniːstan.	The aunt's lizard. Dad (or daddy, as little Dado says) was on the bed. Under him, next to the lake, sat Gigi, also known as Ciccio, the darling of Mom and Grandma. Near a stone, there was a bright red rose, and the foolish one, seeing it, wanted it for his aunt. Aunt Lulù was looking for mosquitoes for her lizard, but since it was June (or maybe July, I'm not sure), she couldn't find any. Instead, she found a frog that, jumping from the road, ended up in the lake with a big splash. You know, the aunt was scared! The splash wet her pink suit, turning it yellow like a taxi. A cosmopolitan gentleman named Sardanapalo Nabucodonosor passed by and fell in love with the aunt, taking her with him to Afghanistan.
Phrases	ˈɔdʒˈːi ˈɛː ˈuna bˈɛlla dʒornˈata pˈer ʃiˈare. vˈoʎo ˈuna mˈaʎa dˈi lˈana kolˈor ˈɔkra iˈɛlle mototʃiklˈista atːraversˈɔ ˈuna strˈada strˈetːa dˈi montˈaŋŋa. patrˈitsia ˈa prantsˈato ˈaː kˈaza dˈi fˈabio.	Today is a beautiful day for skiing. I want an ochre wool sweater. The motorcyclist crossed a narrow mountain road. Patrizia had lunch at Fabio's house.

Table 4.2 continued

	kwˈesto ˈɛ ˈiːl tˈuo kapːˈɛllo?	Is this your hat?
	dˈopo vjˈɛni ʾaː kʾaza?	Are you coming home later?
	lˈa televiziˈone fʊntsiʾona?	Is the television working?
	nˈon pˈɔsso ajʊtˈarti?	Can't I help you?
	mˈarko nˈon ˈɛː partˈito.	Marco didn't leave.
	ˈiːl mˈɛdiko nˈon ˈɛː impeŋˈato.	The doctor is not busy.
Words	pˈipa, buko, tɔpo, dado, kˈaza,	pipe, hole, mouse, nut, house, cat,
	gˈato, fˈilo, vˈazo, mˈuro, nˈɛve,	wire, vase, wall, snow, moon, net,
	lˈuna, rˈete, dzˈɛro, sˈia, tʃˈao,	zero, wake, hello, lap, sun, man,
	dʒˈiro, sˈole, wˈɔmo, jˈuta, ŋˈɔmo,	jute, gnome, him, well, broth,
	ŋˈɛlo, pˈotso, brˈɔdo, plˈadʒo,	plagiarism, train, class, gray, fleet,
	trˈɛno, klˈase, grˈidʒo, flˈɔta,	clay, dragon, friar, shopping,
	krˈɛta, drˈago, frˈate, spˈeza,	stove, ladder, sled, shines, road,
	stˈufa, skˈala, zlˈitsa, splˈɛnde,	writes, spray, rudeness,
	strˈada, skrˈive, sprˈʊtso, zgrˈido,	disfigurement, displeasure,
	sfrˈɛdʒo, zdrˈaio, zbrˈigo, prˈɔva,	displeasure, test, calendar,
	kalendˈario, aʊtobjografˈia,	autobiography, monotonous,
	monˈɔtono, perikolˈɔzo,	dangerous, mountainous,
	montaŋˈozo, prestidʒˈɔzo	prestigious.

4.3.2 Anthea Parkinson's Disease Speech Samples Corpus

The Anthea Parkinson's Disease Speech Samples corpus (ANTHEA-PDSS) refers to two private sub-datasets recorded by ourselves at the Anthea research group within the Polytechnic of Turin. Participants were enrolled at *A.O.U Città della Salute e della Scienza di Torino* and *Associazione Amici Parkinsoniani Piemonte Onlus*. The inclusion criteria were a clinical diagnosis of PD exhibiting vocal signs and symptoms, while lacking significant cognitive impairment or any other conditions that might hinder task completion.

To evaluate the effectiveness of the proposed algorithms in real-world scenarios, both corpora were recorded under sub-optimal conditions with low-cost equipment

such as laptops and smartphones. In the following a detailed description of the two sub-corpus included is reported .

- *ANTHEA-PDSS1*. This sub-dataset involves samples collected in a non-supervised manner through the user's personal computer or smartphone. A tailored web application was employed to guide participants through the same set of tasks included in the IPVS corpus (Table 4.2).
- *ANTHEA-PDSS2*. The second dataset consists of samples recorded in a quiet room under the supervision of an operator. Participants were instructed to sit in a relaxed position with their backs and arms resting comfortably on the back- and arm-rest. They were asked to perform sustained phonation of the vowel /a/ at a comfortable volume, with a smartphone and a high-definition equipment positioned approximately 5 cm from their mouths. An iPhone 12 was employed together with an audio recorder (H4n Zoom, Zoom Corporation, Tokyo, Japan) connected to a Shure WH20 Dynamic Headset Microphone (Shure Incorporated, USA).

Both data collection processes adhered to ethical standards, following the principles outlined in the Declaration of Helsinki. The Ethics Committee of the A.O.U Città della Salute e della Scienza di Torino approved these collections (approval number 00384/2020). Participants were provided with comprehensive information about the study objectives and procedures. Informed consent was obtained, and all demographic and clinical data were recorded anonymously.

Further detail about the participants demographics characteristics are reported in Table 4.3. It is worth noting that, despite all the enrolled patients received detailed neurological examination by experts clinicians, no very recent information regarding the progression and the level of the disease is available.

Table 4.3 Demographics of participants in the ANTHEA-PDSS corpora. Measures are reported in terms of mean ± standard deviation. HC: Healthy Controls; PDP: Patients with Parkinson's Disease; C1: ANTHEA-PDSS1; C2: ANTHEA-PDSS2

	HC (C1)	PDP (C1)	HC (C2)	PDP (C2)
Age	63.62 ± 5.80	70.35 ± 7.23	59.93 ± 15.15	70.38 ± 7.7
Gender	8M, 5F	12M, 5F	11M, 4F	11M, 4F

4.3.3 PC-GITA Corpus

The PC-GITA corpus is a private set of vocal recordings featuring 100 Colombian Spanish speakers. It comprises 50 PDPs (UPDRS III: 29.2±9.11) and 50 HCs balanced in age and gender [105]. All voice samples were recorded while patients were in the ON-state (i.e., no more than 3 hours after taking their morning medication). None of the HC subject presented symptoms associated with PD or any other neurological disease.

The recordings took place in controlled noise conditions employing high-quality audio equipment, including a professional microphone and a Fast Track C400 sound card. The sample rate is 44.1 kHz, with 12-bit resolution. Participants were instructed to perform a series of vocal tasks encompassing: (i) reading of a phonetically balanced dialogue, (ii) articulation of the syllables /pa/ and /ta/, (iii) sustained phonation of the five Spanish vowels, (iv) reading of three list of phonetically balanced words, (v) reading of a list of phonetically balanced sentences, and (vi) performing a spontaneous speech monologue describing a day. The typical recording protocol was designed in order to evaluate phonation, articulation, and prosody by forcing the speaker to use specific muscles whose control is generally impaired in PD. For a comprehensive list of the speech exercises conducted, along with their English translations, please consult Table 4.5.

Additionally, to address the challenge of replicating optimal recording conditions in real-life scenarios, the authors of PC-GITA provided a second dataset, herein referred to as PC-GITA2 [105]. It includes 18 Spanish PDP and 19 Spanish HC individuals. The recordings took place in a quiet room using regular headset and a laptop (16 kHz with a 16-bit resolution) and participants were instructed to perform the same series of speech exercises as in the PC-GITA.

All participants in the two corpora signed an informed consent which was revised and approved by the Ethical Committee of the Research Institute in the Faculty of Medicine at the University of Antioquia (approval 19-63-673). Detailed demographics characteristics about the participant are reported in Table 4.4. It is worth noting that detailed information on the clinical scores of patients in the second corpus is not available, however the samples were collected in such a way as to present distributions similar to the main corpus.

Table 4.4 Demographic information of participants in the PC GITA corpora. Measures are reported in terms of mean ± standard deviation. HC: Healthy Controls; PDP: Patients with Parkinson's Disease; C1: PC-GITA; C2: PC-GITA2

	HC (C1)	PDP (C1)	HC (C2)	PDP (C2)
Age	60.90 ± 6.80	61.14 ± 7.78	53.5 ± 12.25	66.5 ± 8.25
Gender	50M, 50F	50M, 50F	9M, 10F	8M,10F

Table 4.5 List of tasks in the PC-GITA dataset with corresponding English translations

	Prompt (IPA Translation)	Translation to English
Text	P: ajj'er fw'i 'al m'e d'iko. D: k'e' l'e p'asa? m'e preɣ'unto. P: ʏo le dije: ɑy doctor! ðonde pongo el dedo me duele. D: ʝ'o l'e d'ixe: 'aɪ dokt'or! d'onde p'oŋgo 'el d'eðo m'e dw'ele. P: s'i. D: pw'es ʝ'a saβ'emos k'e 'es. d'exe s'u tʃ'eke 'a l'a sal'iða.	P:Yesterday I went to the doctor. D:What happened to you? He asked me. P:I said him: ah doctor! Where I put my finger it pains me. D:Do you have the nail broken? P:Yes. D:Then we now know what is happening. Leave your check at the exit.
Phrases	b'iste l'as not'iθjas? ʝ'o b'i ganˈar l'a með'aʌa d'e pl'ata 'en p'esas. 'ese mutʃ'atʃo tj'ene m'utʃa fw'erθa!	Did you see the news? I saw to win the silver medal in Weightlifting. That boy is very strong!
	xw'an s'e rr'ompjo 'una pj'erna kw'ando 'iβa 'en l'a m'oto.	Juan broke his leg when he was driving his motorcycle
	est'oɪ m'ujj tr'iste, ajj'er b'i morˈi' 'a 'un amˈiɣo	I am very sad, yesterday I saw a friend die
	est'oɪ mw'i pr,eokup'aðo, k'aða b'eθ m'e 'es m'as d,iffiθ'il aβl'ar	I am very concerned, it is increasingly more difficult to talk
	m'i k'asa tj'ene tr'es kw'artos.	My house has three rooms.
	om"aR, k"e b"iBe T"eRka, t""axo mj"el.	Omar, who lives near, brought honey
	l'aʊra s'uβe 'al tr'en k'e p'asa.	Laura gets on the passing train
	l'os l'iβros nw'eβos n'o k'aβen 'en l'a m'esa d'e l'a ,ofiθ'ina.	The new books do not fit in the office's table

Table 4.5 continued

	ɾɾosˈita nˈin ˈo, kˈe pˈinta bjˈen, dˈono sˈus kwˈaðros ajjˈeɾ. luˈisa ɾɾˈeɪ kˈompɾa ˈel kˈoltʃon dˈuɾo kˈe tˈanto lˈe ɣˈusta.	Rosita Nino, who paints well, donated her paintings yesterday Luisa Rey buys the hard mattress that is so fond her
Words	petˈaka, boðˈeɣa, pˈato, ˈapːto, kampˈana, pɾˈesa, plˈato, bɾˈaθo, blˈusa, tɾˈato, atlˈeta, dɾˈama, ɣɾˈito, ɣlˈoβo, kɾˈema, klˈaβo, fɾˈuta, flˈetʃa, bjˈaxe, ʎuˈeβe, kaʊtʃo, ɾɾˈeɪna, nˈame, kˈoko, ɣˈato	petaka, cellar, duck, suitable, bell, dam, dish, arm, blouse, deal, athlete, drama, cry, balloon, cream, clove, fruit, arrow, trip, rains, rubber, queen, yam, coconut, cat
	ˌakaɾiθjˈaɾ, aplaʊðˈiɾ, ˌaɣaɾɾˈaɾ, dˌiβuxˈaɾ, pˌataleˈaɾ, pˌisoteˈaɾ, tɾotˈaɾ, sonɾˈeɪɾ, sopˈlaɾ, mˌastikˈaɾ.	stroke, clap, grab, draw, stamp, trample, jog, smile, blow, chew
	bˈaɾko, bˈoske, θjuðˈad, estˈaβlo, ospitˈal, lˈuna, montˈana, nˈuβe, pwˈɛnte, traktˈoɾ.	ship, forest, city, stable, hospital, moon, mountain, cloud, bridge, tractor.

4.3.4 Hlavnicka Corpus

The Hlavnicka corpus is a publicly available dataset released under the Creative Commons Attribution License (CC BY 4.0). This dataset involves a total of 83 Czech participants including 22 PDPs and 22 HCs. The remaining subjects presented other pathologies which are outside the aim of this work, including Multiple System Atrophy, and Progressive Supranuclear Palsy. As for PD, the disease duration was estimated from self-reported first motor symptom occurrence. A neurologist evaluated all patients and assessed their motor abilities using standardized scales. The UPDRS III score was 15.9 ± 7.6, with a median disease duration of 9.3 ± 5.5. All the PD participants were examined while on medication after at least 4 weeks of stable medication. Detailed information aboout the subject's demographics are reported in Table 4.6

Table 4.6 Demographics of participants in the Hlavnicka corpus. Measures are reported in terms of mean ± standard deviation. HC: Healthy Controls; PDP: Patients with Parkinson's Disease

	HC	PDP
Age	63.6 ± 10.10	64.4 ± 9.6
Gender	11M, 11F	10M, 12F

During the data collection procedure, each participant received instructions from a trained specialist to produce prolonged vowel sounds, specifically /a/ and /i/ as long and steadily as possible. The recordings took place in a low-ambient-noise room using a headset condenser microphone (Opus 55, Beyerdynamic, Heilbronn, Germany) positioned approximately 5 cm from their lips. Recordings were digitized at 16-bit resolution and a 48 kHz sampling frequency.

All participants provided written, informed consent. The study received ethical approval from the Ethics Committee of the General University Hospital in Prague (approval number 67/14 Grant VES AZV 1.LFUK) and adhered to approved guidelines. For further insights and detailed information, please refer to the source publication [106].

4.3.5 Suppa Corpus

The Suppa corpus is a private dataset detailed in the publication by Suppa et al. in 2022 [107]. The dataset comprises a total of 115 Italian native speakers diagnosed with PDPs and 108 age-matched HCs. Participants were recruited from the *IRCCS Neuromed Institute and the Department of Systems Medicine at Tor Vergata University in Rome, Italy*. The inclusion criteria were being 18+ non-smokers, lacking significant cognitive impairment, or any pathology affecting the vocal apparatus. The clinical diagnosis of PD was performed by expert clinicians following established clinical criteria, and using the H&Y and UPDRS-III scales.

The study cohort PDP was thoughtfully structured to encompass two distinct subgroups of PDP, including *Early* and *Mid-Advanced* PDPs. The former consisted of 57 individuals at an early stage of the disease (H&R score ≤ 2). Notably, these patients had not yet been treated with L-Dopa for their condition at the time of the study. The second group included 58 patients who had progressed to a mid-advanced

stage of PD (H&R score ≥ 2). These patients were undergoing chronic treatment with L-Dopa. Additionally, a subset of 31 out of the 58 mid-advanced-stage patients were further assessed in ON and OFF therapy condition (i.e., after 12 hours from the last medication intake and 1 to 2 hour after administration). Demographic and clinical details of the participants can be found in Table 4.7. It is worth nothing that despite the three subgroups features the same proportion of male and female participants, the exact cardinality of the subgroups is not available.

The data collection procedure was conducted by instructing participants to perform specific vocal tasks using their usual voice intensity, pitch, and quality. These tasks included the sustained phonation of the vowel /e/ for at least 5 seconds and the utterance of two standardized Italian proverbs ˈa kaβˈal donˈato nˈon sˈi ɣwˈarða ˈin bˈokka (i.e., Don't look a gift horse in the mouth) and mˈeɣlio sˈoli tʃˈe mˈale ˌakkompaɣnˈati (i.e., Better alone than in bad company). Voice recordings were captured using a high-definition audio recorder (H4n Zoom, Zoom Corporation, Tokyo, Japan) connected to a Shure WH20 Dynamic Headset Microphone (Shure Incorporated, USA). The microphone was positioned approximately 5 cm from the participants' mouths. Voice samples were recorded in linear PCM format (.wav) at a sampling rate of 44.1 kHz, with a 16-bit sample size. Participants gave written informed consent, which was approved by the institutional ethics committee (0026508/2019), according to the Declaration of Helsinki.

Table 4.7 Demographics of participants in the Suppa corpus. Measures are reported in terms of mean ± standard deviation. HC: Healthy Controls; PDP: Patients with Parkinson's Disease; n.r.: not reported

	HC	PDP-Early	PDP Mid-Advanced
Age	68.2± 9.2	64.2 ± 8.6	72.1 ± 8.1
Gender	n.r.	n.r.	n.r.

4.3.6 LUHS Corpus

The Lithuanian University of Health Sciences (LUHS) is a publicly accessible dataset distributed under the Creative Commons Attribution License (CC BY 4.0), as detailed in [108]. This resource comprises vocal recordings from 99 Lithuanian native speakers, categorized into two groups: 64 PDPs and 35 HCs, with no symptoms

associate to neurodegenerative disorders nor diseases that could affect the vocal apparatus.

The audio samples were recorded using two channels simultaneously: an acoustic cardioid microphone (AKG Perception 220) and a smartphone (Samsung Galaxy Note 3). Both microphones were positioned approximately 10 cm from the mouth. The audio format is mono PCM .wav (16-bit at a sampling frequency of 44.1 kHz). During the data collection procedure, each participant was instructed to execute the sustained phonation of the vowel /a/ as long as possible for three times and to pronounce a phonetically balanced phrase in Lituanian languge. For a more comprehensive overview of participant demographics, please refer to Table 4.8.

The dataset is distributed in the form of pre-extracted set of features computed by means of different toolboxes, namely OpenSmile, Essentia, MPEG7AudioEnc, jAudio, YAFEE, and DARTH. In Table 4.9 an overview of the extracted sets is reported.

Table 4.8 Demographics of participants in the LUHS corpus. Measures are reported in terms of mean ± standard deviation. HC: Healthy Controls; PDP: Patients with Parkinson's Disease

	HC	PDP
Age	41.74 ± 17.11	64.95 ± 9.56
Gender	11M, 24F	30M, 34F

4.3.7 Additional Corpora

Despite not directly used within the experiments conducted in the present dissertation, this section briefly reports additional corpora previously employed in similar studies for the analysis of individuals with PD. The purpose is to offer readers comprehensive insights into the available materials for future research.

References to the original papers, which provide detailed information about these datasets, are included in order to allow readers to access more information. As previously done, if a specific dataset lacks a designated name, the name of the first author in the corresponding paper is used as a reference for the corpus. It is important to note that the datasets from the Center for Machine Learning and Intelligent Systems at the University of California Irvine (UCI) (Little, Sakar18,

Table 4.9 Overview of the feature sets included in the LUHS corpus

No	Feature set	ID	Toolbox
1	Avec2011	AV1	OpenSMILE toolkit [109]
2	Avec2013	AV2	OpenSMILE toolkit [109]
3	Emo_large	EL	OpenSMILE toolkit [109]
4	Emobase	EM1	OpenSMILE toolkit [109]
5	Emobase2010	EM2	OpenSMILE toolkit [109]
6	Essentia_descriptors	ED	Essentia [110]
7	IS09_emotion	IS1	OpenSMILE toolkit [109]
8	IS10_paraling	IS2	OpenSMILE toolkit [109]
9	IS10_paraling_compat	IS3	OpenSMILE toolkit [109]
10	IS11_speaker_state	IS4	OpenSMILE toolkit [109]
11	IS12_speaker_trait	IS5	OpenSMILE toolkit [109]
12	IS12_speaker_trait_compat	IS6	OpenSMILE toolkit [109]
13	IS13_ComPare	IS7	OpenSMILE toolkit [109]
14	jAudio_features	JA	jAudio [111]
15	MPEG7_descriptors	MP	MPEG7AudioEnc [112]
16	Tsanas	TS	Tsanas [44]
17	YAAFE_features	YA	YAAFE toolbox [113]

Sakar13, Naranjo, Tsanas) and the Vaiciukynas corpus are provided in the form of pre-extracted feature vectors.

- *Little.* First introduced by Little et al. in 2009 [114], this corpus consists of six recordings of sustained vowel phonation (/a/) from 23 individuals with PD and 8 HCs. The recordings were captured using a head-mounted microphone (AKG C420) placed approximately 8 cm from the speakers' lips within a controlled acoustic environment. The voice signals were directly recorded onto a computer using the Computerized Speech Laboratory (CSL) 4300B hardware by Kay Elemetrics. Notably, this dataset does not provide information regarding any therapy sessions or interventions.

- *Sakar18.* First introduced by Sakar et al. in 2019 [115], this corpus comprises three recordings of sustained vowel phonation (/a/) from a total of 188 individuals with PD and 64 HC. Unfortunately, this dataset does not provide details regarding the type of microphones used for the recordings.

- *Sakar13.* First introduced by Sakar et al. in 2013 [116], this dataset consists of recordings of sustained vowel phonation (/a/ , /o/, /u/), utterances of numbers from 1 to 10, short sentences, and single words. The dataset includes samples from 34 individuals with PD and 34 HC. Recordings were made using a Trust MC-1500 low-end computer microphone positioned 10 cm from the subjects.

- *Naranjo* First introduced by Naranjo et al. in 2016 [117], this dataset comprises three recordings, collected weekly over a period of six months, of sustained vowel phonation (/a/) from 40 individuals with PD and 40 HC. Unfortunately, the microphone details used for the recordings are not provided in the available information.

- *Tsanas* First introduced in Tsanas' work in 2010 [42], this dataset comprises multiple recordings from 42 individuals with PD).

  These recordings were made over the course of a 6-month trial and involve sustained vowel phonation (/a/). The data was collected using the Intel Corporation at-home testing device (AHTD), a telemonitoring system equipped with various sensors, including a high-quality microphone headset.

- *m-Power-* Initially introduced in the works of Bot et al. in 2016 [118] and Prince et al. in 2018 [119], this dataset consists of audio data sourced from volunteers in English. The data was recorded using iPhones and specifically focuses on the vowel /a/ .

4.4 Experimental Findings: Variability in Vocal Tasks

As evident from the analysis of the related literature (Section 4.2), various vocal tasks have been employed over time to study vocal alterations in individuals with PD. In this context, the effectiveness of sustained vowel analysis is well-established. However, an increasing body of evidence suggests the need to complement this analysis with a more comprehensive examination of the subject's articulatory and prosodic features to delve further into the complexity of vocal alterations. Within this field, there is a wide variety of vocal tasks employed, including the rapid repetition of specific phonemes, as well as the repetition of word sequences, phrases, or entire texts. Without a proper selection, the diversity of exercises can be burdensome,

leading to fatigue and excessive time required for data collection, which may reduce patient's compliance, especially in a home environment.

In this context, the initial experiments presented in this dissertations focus on the comparison of various sets of vocal exercises to identify the vocal tasks that can better capture vocal alterations in PDPs while minimizing the effort required to the subjects. More specifically, the first study (Section 4.4.1) considers isolated words and examines the advantages and disadvantages of different techniques for using multiple words, both in terms of classification accuracy and computational efficiency. Building on these findings, the second study (Section 4.4.2) systematically assesses effective exercise sets, including vocalizations and sentence repetitions. To ensure robust and widely applicable results, the study is conducted on diverse datasets from various conditions and nationalities. The results are then compared to yield strong, general conclusions. The experiments described in this sections are published in [14, 120]

4.4.1 Analysis of Isolated Word Speech Task

This study aimed to assess the effectiveness of a classification model based on a single vocal task, involving the utterance of multiple isolated words. A multi-level approach was employed, where features were extracted from various segments of the vocal signal, including the entire signal, voiced segments, and transition areas between voiced and unvoiced regions. The goal was to study the effectiveness of different types of acoustic parameters in modeling the subject's articulatory capability.

In the proposed experiment, 25 vocal signals were employed for each subject, with each sample corresponding to a different word. These vocal signals underwent pre-processing and feature extraction steps. Subsequently, two distinct feature fusion techniques, known as *early-fusion* and *late-fusion*, were explored and compared to merge the extracted features. This procedure allowed to assess which fusion approach was the most effective in terms of classification performance and the most efficient in terms of computational requirements.

Materials

In this work, the two distinct datasets provided by the GITA group were employed, namely PC-GITA and PC-GITA2 (Section 4.3.3). For each corpus 25 samples per subject corresponding to 25 phonetically balanced words were used (Table 4.5).

Data analysis was carried out in Matlab, leveraging customized pre-processing and feature extraction routines based on *Audio* and *Signal Processing* Toolboxes. Praat (by Paul Boersma and David Weenink, Phonetic Sciences, University of Amsterdam) [121] was also employed during the pre-processing steps.

Methods

Pre-processing. The audio signals were initially low-pass filtered to reduce distortion and eliminate background noise. To maintain minimal phase distortion within the pass-band, a 10^{th}-order zero-lag Butterworth low-pass filter with a cutoff frequency of 3750 Hz was employed, as described in [122]. In addition, detrending was applied to remove slow fluctuations in the signal that might be attributed to the recording system.

The Praat software was used for the identification of voiced and unvoiced regions within the audio data. Following signal labelling, voiced regions were merged and each sample was segmented into 20 ms windows with 50% overlap, as in prior studies involving the same task [123, 124]. Furthermore, to gain a deeper understanding of transient-type sounds, which result from the abrupt release of previously blocked airflow, particularly in occlusive consonants (as discussed in Section 4.2), windows centered on the edges of each voiced region were identified and analyzed. Based on previous research by Vasquez-Correa et al. [125], the window length was set to 160 ms.

Feature Extraction. In the process of vocal parameterization, a multi-level approach was adopted, incorporating a total of 126 features extracted from different parts of the audio signal, namely the entire signal, voiced segments, and the onset/offset regions.

The feature set included widely recognized parameters for vocal analysis in PD patients. This encompassed measures such as F0, HNR, Formant Bandwidths

STE, and 10 spectral parameters (comprising flux, skewness, entropy, rest, latness, slope, roll-off, spread, centroid, and kurtosis), along with DFA. Furthermore, to capture articulatory dynamics, the set featured 13 MFCCs with their first and second derivatives, which convey information about the velocity and acceleration of the articulators. Additionally, 3 LPC coefficients, 25 Bark Band Energy (BBE) coefficients, and 6 Instantaneous Energy Deviation Coefficients (IEDCC) proposed in [123] were included to provide a comprehensive representation.

To better address the alterations at the onset and offset of voiced sound regions and to model the phenomenon of *voicing leakage* which is often perceivable in PDPs due to their difficulty in executing precise and rapid articulatory movements (as discussed in Section 3.1), two novel features were introduced: the Pitch Transition Slope (PTS) and Energy Transition Slope (ETS). These features were derived from the evaluation of pitch and energy contours in the transitional regions of each word using a first-order polynomial. The slope of the resulting curve served as a metric to quantify vocal alterations. In pathological voices, this curve is expected to flatten when there is a failure in the change of F0 and energy between voiced and unvoiced regions. To ensure consistency and facilitate subsequent analyses, range normalization was applied to the entire feature set.

For a more comprehensive view of the extracted features and their categorization, please refer to Table 4.10. Specifically, the Table elucidates the region from which each metric was computed. It is noteworthy to mention that distinct features were extracted from various regions of the vocal signal to ensure the inclusion of adequate parametrizaion techniques for both quasi-periodic and non-periodic segments of the vocal signal.

Feature Selection. A customized feature selection process was employed on the PC-GITA database to identify features that exhibited a strong correlation with the class while ensuring they were non-redundant. Indeed, it was imperative to employ a specific and meticulous feature selection procedure. Given the exploratory nature of the experiment, which involved extracting a comprehensive set of parameters to investigate potential correlations, minimizing the risk of the curse of dimensionality problem was crucial. This was especially challenging considering the size of the dataset under examination. To avoid overfitting and ensure generalization capability, the corpus was randomly divided into two subsets: 70% for training and

Table 4.10 Overview of the extracted features, categorized by the domain of analysis

Region	Feature Name	Information retrieved
Entire signal	IEDCC(1–6)	Vocal tract and vocal folds abnormalities
	Zero crossing rate	Voice activity
Voiced	DFA	Self-similarity of the voice
	Bandwidth	Frequency range
	HNR	Ratio of signal over noise
	F0	Periodicity alteration
	Spectral features	Spectrum shape information
	LPC(1–3)	Formants and resonances
	STE	Energy variation among frames
	MFCC(1–13)	Subtle changes in the motion of articulators
Transition	PTS	Ability to control vocal fold vibration
	ETS	Ability to control vocal fold vibration
	MFCC(1–13)	Ability to control vocal fold vibration
	BBE(1–25)	Ability to control vocal fold vibration

validation and 30% for testing. Particular attention was devoted to maintain *speaker independence*, hence all words from the same speaker were either in the training or testing set. Additionally, to investigate potential gender-related variations in acoustic features, the dataset was initially divided into male and female speakers, then the pipeline was applied to the two distinct subsets of samples.

The first step involved calculating the absolute value of the Pearson's correlation coefficient (r) between the features and the target variable (rf_0). The aim was to identify features with a strong correlation with the target variable (i.e., rf_0 greater than a threshold denoted as th_1). To determine the appropriate threshold, a tuning procedure was carried out within the training data. This tuning process focused on minimizing classification errors in a 10-fold CV setup using a quadratic Support Vector Machine (SVM), which was employed due to its robust generalization capabilities and its widespread use in classifying samples from PDPs [126, 123, 124, 127]. It is noteworthy to emphasize that, concerning the training-test divisions, the CV procedure was executed based on subject-IDs. This approach ensured that samples from the same subject are not present simultaneously in both the training and validation sets, enhancing the robustness and reliability of the model evaluation. Specifically, th_1 was tuned over a range from 0.3 to 0.6 with increments of 0.1. To

eliminate redundant features, the correlation coefficient was computed between pairs of features (r_{ff}) in order to remove those showing a correlation greater than th_2 times the correlation with the target variable (r_{fo}). The value of th_2 was tuned from 0 to 50% with 5% increments, choosing the value that minimized the classification error rate in a 10-fold CV.

Given that this study employed a speech task made up of 25 isolated words per subject, two feature-fusion approaches were compared to determine the most effective method for extracting information from multiple utterances by the same subject. These fusion strategies were classified as *early fusion* and *late fusion* as introduced in [128]. The *late fusion* approach involved two classification levels. In the first level, 25 models were employed, each taking selected features from individual words as input. The second level consisted of a single model that utilized the outputs from the previous layer, which had been transformed into probability scores using Platt's method [128]. On the other hand, the *early fusion* approach entailed the prior selection of the most informative words, followed by the concatenation of the corresponding features. The identification of the most effective words was determined based on the number of features selected for each word, denoted as f_w. For a word to be considered, f_w had to exceed a predefined threshold, denoted as th_3, whose values was tuned over a range from 1 to 80 with 5-step increments. As in the previous stages, the effectiveness of the two fusion approaches and the different feature sets was assessed by comparing the classification accuracy through a 10-fold CV on the training set, employing a quadratic SVM model.

The pseudocode for the proposed feature selection approach is outlined in Algorithm 1, while in Figure 4.1 is reported a schematic of the two fusion approaches applied.

Classification. Following the identification of the optimal fusion scheme, a comparative analysis of various classifiers was conducted to assess whether different algorithms could potentially yield improved classification accuracy. Specifically, the quadratic SVM was compared to other classifiers, including k-Nearest Neighbors (KNN), Naive Bayes (NB), Decision Tree (DT), a bagged trees ensemble, and a subspace discriminant (SD) ensemble.

To optimize the best-performing model, a grid-search approach was applied. The procedure involved evaluating four different distance metrics (euclidean, city

Algorithm 1 Feature selection pseudo-algorithm from [129]

Input: D: training dataset ▷ N subj., W words per subj., F feat. per word
Output: T: reduced training dataset ▷ N subj., F_1 feat. per subj.

```
for each n ∈ N do
    for each w ∈ W do
        for each f ∈ F do
            r_fo ← crosscorr(f, class);
            if r_fo ≥ t_h1 then
                F_w.add(f);                ▷ Select most significant feat. per word
            end if
        end for
        if F_w.count() ≥ t_h3 then
            W_f.add(w); ▷ Select words with the higher number of feat. selected
        end if
    end for
    Feat ← (W_f(F_w)).merge();             ▷ Merge feat. selected from words selected
    for each f ∈ Feat do
        r_ff ← crosscorr(f_i, f_j);
        if r_ff ≤ t_h2 · r_fo then
            T.add(f);                      ▷ Select feat. with lower inter-feat. corr.
        end if
    end for
end for
```

block, Minkowski, and Chebyshev) and varying the value of k from 2 to N/2, where N represents the number of samples in the training set. In cases where multiple hyperparameters resulted in the same optimal accuracy, preference was given to lower k-values to reduce the computational workload of the model. Given the use of random data splitting procedures, the average accuracy over five iterations was considered as a robust metric for the optimization process.

Time Complexity. To assess the overall performance of the best pipeline, its time complexity was assessed by measuring computational time under varying conditions. The variables used were N (number of subjects), W (number of words per subject), F (initial features per word), and F1 (final selected features).

Within the feature selection (Algorithm 1), selecting significant features and identifying words with the most selected features was estimated to have a time complexity of $O(nfw)$. The process of choosing features with lower inter-feature

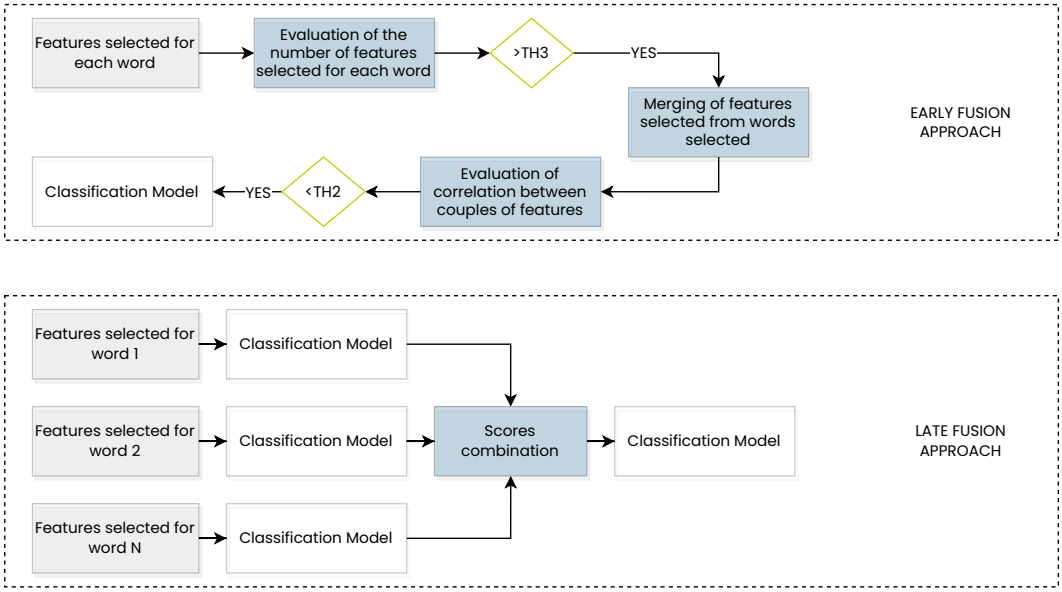


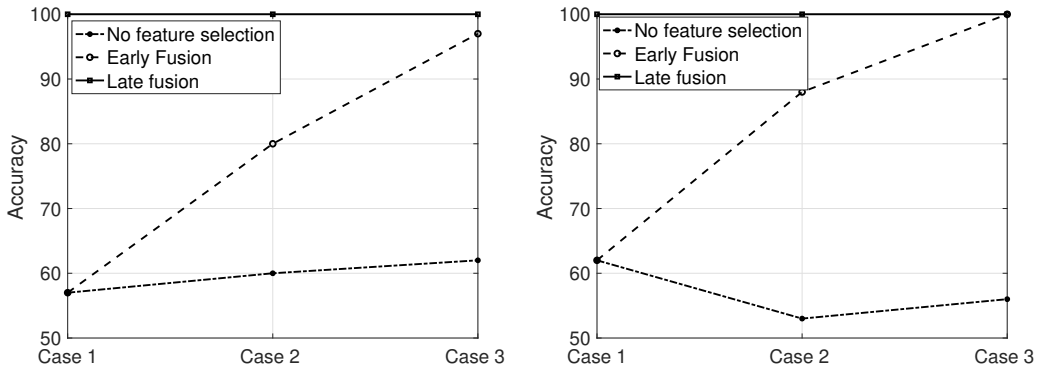
Fig. 4.1 Workflows of early and late fusion approaches

correlation, which includes calculating Pearson correlation coefficients between feature pairs, takes $O(nf_1^2)$ time. In the worst-case scenario where all features and words are selected, and if f_1 equals f_w, the feature selection algorithm time complexity reaches $O(nf^2w^2)$. As for the classification algorithm used, the Matlab implementation of the KNN classifier exhibits a time complexity of $O(log(n))$ [130].

To validate these findings, the overall pipeline was executed multiple times with different inputs, features, and words. For brevity, the analysis is presented for the female dataset, although the same process was applied also to the male's one. All experiments were conducted on a MacBook Pro with a 64-bit OS, a 2.7GHz Intel Core i5 processor, and 8GB of RAM.

Results

Comparison Between Fusion Approaches. Figure 4.2 offers a comparison between the early and late fusion approaches applied to three distinct feature sets. This evaluation measures accuracy within a 10-fold CV using a non-optimized quadratic SVM model. For ease of comparison, the results achieved by combining all features extracted from each word without any feature selection are also provided.



(a) *Male* (b) *Female*

Fig. 4.2 Performance outcomes of early and late fusion schemes, sourced from [129]

To provide an insight into the generalization capability of each best configuration, the models were tested on the 30% of the PC-GITA dataset extracted from the original corpus. While the late fusion results on the test set were not satisfactory, the *Case 3 - early fusion* configuration demonstrated an accuracy of 82% (averaged over 5 iterations for both male and female groups). Regarding the pipeline execution time, which encompasses the time from feature selection to classification, the most effective models resulted in 3.37 s, 4.19 s, and 6.23 s for the three late fusion cases, respectively. In contrast, the only model employing an early fusion approach achieved a considerably shorter time of 0.065 s.

Classification. Table 4.11 displays the results for different classifiers, reporting the average accuracy from 5 iterations using non-optimized algorithms. The optimal threshold values for feature selection were determined to be $th1 = 0.5$, $th2 = 0.1$, and $th3$=10 for males and $th1 = 0.5$, $th2 = 0.1$, and $th3 = 30$ for females. For the two best-performing models (KNN), grid-search optimization led to the selection of cityblock distance with k = 3 for males and k = 6 for females. Moreover, Table 4.12 presents the most significant words and features identified for male and female subgroups.

Given the limited dataset size, the influence of individual speaker characteristics was assessed by running the thus optimized algorithms on random subsets of the original dataset. Results from the 5 iterations are reported in Table 4.13. Additional tests on PC-GITA2 corpus (Section 4.3.3) examined the impact of recording condi-

Table 4.11 Comparative analysis of the six tested models. Results are reported as the average classification accuracy over 10-fold CV from 5 iterations on randomly selected subsets

	Male subset		Female subset	
	10 fold CV	Test set	10 fold CV	Test set
SVM	0.96 ± 0.032	0.74 ± 0.019	0.90 ± 0.025	0.90 ± 0.071
DT	0.95 ± 0.044	0.64 ± 0.017	1 ± 0	0.65 ± 0.020
NB	0.73 ± 0.041	0.50 ± 0.028	92 ± 0.056	0.77 ± 0.022
kNN	0.96 ± 0.025	0.74 ± 0.016	0.99 ± 0.016	0.97 ± 0.034
Bagged Trees	0.92 ± 0.051	0.60 ± 0.02	0.96 ± 0.013	0.56 ± 0
SD	0.94 ± 0.053	0.71 ± 0.016	0.99 ± 0.013	0.96 ± 0.034

Table 4.12 List of the most significant words and features for male and female subsets

	Words	Features	Region
Female subset	Clavo, Crema, Globo, Name	MFCC, BBE, ΔΔMFCC	onset
		Roll-off-point	voiced
		PTS, ETS, MFCC, BBE, ΔΔMFCC	offset
Male subset	Bodega, Braso, Globo, Llueve, Name, Presa, Viaj	MFCC, BBE, ΔΔMFCC	onset
		MFCC, BBE, PTS	off-set

tions on the model performance. Over 5 iterations, the average accuracy was 60% for the male subgroup and 62% for the female subgroup.

To comprehensively evaluate the system performance and make a comparison with similar studies that employed isolated words from the PC-GITA corpus, Table 4.14 is provided. This comparative analysis considers the best validation results from three distinct studies: [131] (utilizing a 10-fold CV approach), [123] (using Leave One Subject Out -LOSO- validation), and [132] (employing a 5-fold CV approach).

Time Complexity. Figure 4.3 present the results of the time complexity analysis. To ensure robust findings, each experiment was conducted five times on randomly selected subsets, and the average time value is reported. In Figure 4.3a, the execution time is depicted in relation to the parameter N. It is worth noting that to provide a realistic analysis despite the limited dataset size, the investigation was extended beyond the original range of 2 to 49. This extension involved additional measurements

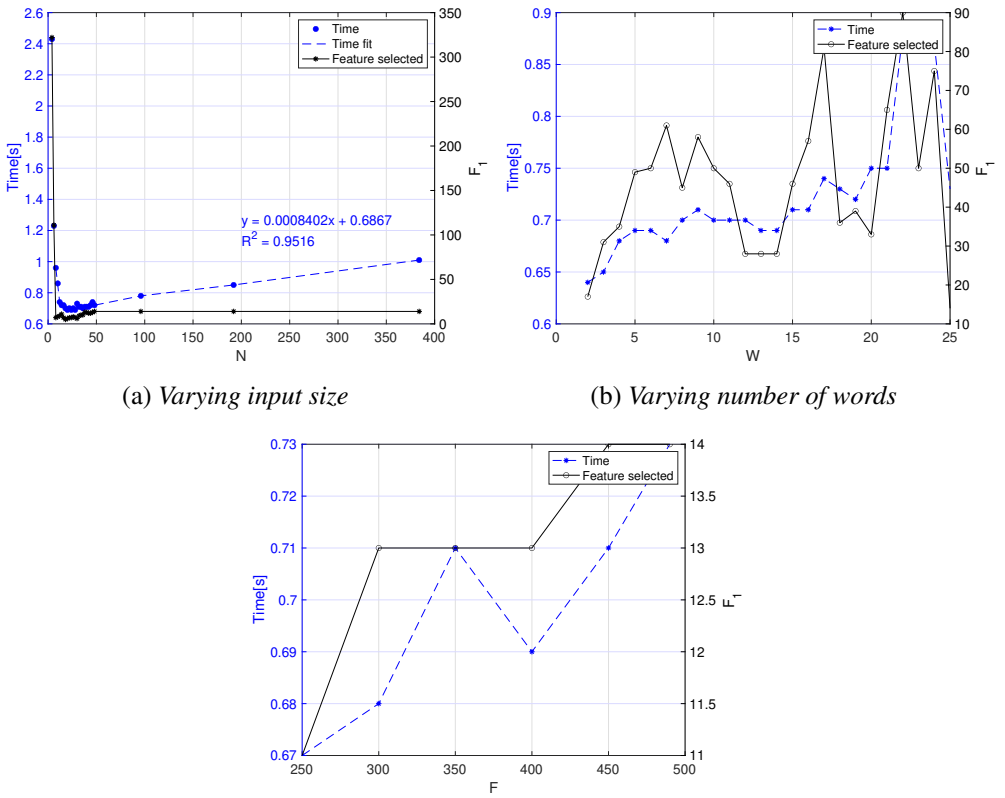
Table 4.13 Comparative performance analysis over five iterations for male and female groups. N Iter.: Number of iteration; Acc: Accuracy; Sens.:Sensibility; Spec.:Specificity

		10-fold CV				Test set			
	N Iter.	Acc.	Sens.	Spec.	AUC	Acc.	Sens.	Spec.	AUC
Male subset	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	1.00	1.00	1.00	1.00	0.94	1.00	0.87	0.94
	3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	4	0.97	1.00	0.94	1.00	1.00	1.00	1.00	1.00
	5	1.00	1.00	1.00	1.00	0.94	0.87	1.00	0.94
	mean	**0.99**	**1.00**	**0.99**	**1.00**	**0.97**	**0.97**	**0.97**	**0.98**
Female subset	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	1.00	1.00	1.00	1.00	0.75	0.63	0.87	0.75
	3	0.97	0.94	1.00	0.97	0.87	0.75	1.00	0.88
	4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	5	1.00	1.00	1.00	1.00	0.94	0.88	1.00	0.94
	mean	**0.99**	**0.99**	**1.00**	**0.99**	**0.91**	**0.85**	**0.97**	**0.91**

Table 4.14 Comparison of results with the best outcomes from similar studies using the PC-GITA corpus for isolated word repetition tasks. n.r. = not reported

Authors	[131]	[123]	[132]	Present study
Year	2015	2020	2020	2020
Model	SVM	SVM	CNN	kNN
Sensibility	0.94	n.r.	n.r.	0.99
Specificity	0.90	n.r.	n.r	0.99
Accuracy	0.92	0.91	0.77	0.99
F1-score	n.r.	0.83	n.r.	0.99

on a simulated, larger dataset created by duplicating the same samples multiple times. Figure 4.3b illustrates the execution time while varying the number of words from 1 to 25. In Figure 4.3c, it is evident a gradual reduction in the number of parameters until the feature selection algorithm remains applicable. It is important to mention that a significant reduction in the number of initial features for each word might lead to an empty set of words that meet the threshold criteria.



(a) *Varying input size*

(b) *Varying number of words*

(c) *Varying number of input features*

Fig. 4.3 Results from the analysis of time complexity, sourced from [129]

Discussion

Comparison Between Fusion Approaches. The analysis of Figure 4.2 reveals a clear improvement in the model performance when expanding the feature set to include more specific features like voiced segments and transition regions, especially during cross-validation. Notably, the late fusion scheme maintains relatively consistent and maximal performance across different feature sets, potentially indicating overfitting. However, it's evident that the most effective system configurations are early fusion with the complete feature set and late fusion regardless of the chosen set. Among these, early fusion setup demonstrated to be also computationally efficient, standing out as the choice to be preferred.

Classification. In Table 4.11, the KNN algorithm demonstrated optimal performance, being also characterized from a smaller standard deviation which indicates

more consistent results across random data splits. The computational time for the final pipeline was calculated, resulting in an average of 0.047 seconds for both male and female subjects, confirming the efficiency of the KNN algorithm. Optimized model parameters included the use of the city block distance and setting k to 6 for males and 3 for females.

Regarding the optimized models, results in Table 4.13 indicate that the model achieved an optimal correct classification rate in both the validation and test sets. However, it is important to note that the choice of different inputs significantly influenced algorithm performance. This effect is particularly pronounced in the male group, where classification accuracy varies from 75% to 100%. In contrast, the classification accuracy in the female group remains stable at 100% in 3 out of the 5 sets analyzed. Given the higher standard deviation and lower performance observed in the male population with most of the tested models, the variation in performance may be attributed to the dataset composition. Additionally, a general decrease in performance in the additional dataset, especially in the male group is observable. While overfitting was not significant in the test set, this reduction can be primarily attributed to the distinct recording conditions in the new dataset. As for the comparison with similar studies, evidence in Table 4.14 suggests that the proposed algorithm performance metrics outperform those of the studies in the comparison. However, it is important to acknowledge that this study does not encompass a large cohort of PD patients.

As for the feature selected, most of them originate from the transition regions (Table 4.12), emphasizing the effectiveness of these areas in the analysis of speech patterns related to PD. PTS and ETS were chosen for both the female and male groups, suggesting their potential as distinctive markers of the pathological condition.

Time Complexity. The results from the time complexity analysis reported in Figure 4.3a indicate that the number of selected features remains relatively consistent, regardless of the value of N. However, when the number of training samples decreases significantly (<6 per subject) a larger number of features is selected due to the inherent limitation of a feature selection procedure based on the correlation coefficient. Limiting the considerations to the region where the model exhibits stability, i.e., with more than 6 subjects per group, the regression line of the curve follows the expected linear trend ($R^2 = 0.9516$).

Regarding Figures 4.3b and 4.3c, they clearly show that computational time generally rises with increasing W ($r = 0.77, P < 0.001$) and F ($r = 0.88, P = 0.020$). However, the exact nature of this relationship may not be easily discernible. This is because the specific execution time values are also influenced by F1, which, in turn, is highly dependent on the selection of words and features used in a given iteration.

4.4.2 Comparative Analysis of Multiple Vocal Tasks

This second study aimed to assess the effectiveness of various vocal tasks and their combinations in capturing vocal changes associated with the presence of PD while minimizing the subject's effort.

The experiment was conducted in two sequential steps. First, a preliminary analysis was performed to determine if different versions of the same vocal tasks (e.g., different vowels for sustained vowel phonation or different phrases for sentence repetition) demonstrated varying effectiveness in discriminating between PD and HCs. If such variations were observed, the aim was to ascertain whether this phenomenon was consistent across multiple corpora, indipendently from the language spoken by the speakers or the data-collection method employed. Subsequently, an investigation was conducted to assess whether different vocal tasks or their combinations could yield improved performance in automating the distinction between PDP and HC. Comprehensive statistical analyses were carried out across multiple datasets to investigate the robustness of the results and study the potential impact of participant demographics and data collection variables on the findings.

Materials

This study utilized four diverse datasets, comprising a total of 279 subjects, with 139 individuals having PD and 140 HCs. The datasets included the IPVS (Section 4.3.1), PC-GITA (Section 4.3.3), ANTHEA-PDSS1 (Section 4.3.2), and a subset of the Suppa dataset, which included 46 PDPs (33 males) and 56 HCs (15 males) (Section 4.3.5).

Data analysis was carried out in Python employing Praat for pre-processing and feature extraction. The Parselmouth library served as an interface to access Praat internal code.

Methods

Pre-processing. To ensure data uniformity, the recordings from the four datasets, which had different sampling rates, were initially down-sampled to 16 kHz. Additionally, signal amplitudes were normalized within the [0, 1] range to minimize the potential impact of speaker-microphone distances on subsequent analyses. To enhance data quality, initial and final silence regions were manually removed, eliminating the need for additional preparatory steps.

For the analysis of the phrase repetition task, the start and end points of voiced regions were detected using Praat software. Subsequently, these regions were merged, and each signal was segmented into 40 ms windows with a 20 ms overlap.

Feature Extraction. A comprehensive set of features was extracted from each vocal samples, encompassing periodicity measures, which included F0, the first three formants along with their bandwidths, as well as noise measures such as HNR, CPP, and GNE. Furthermore, spectral (i.e., flux, skewness, entropy, crest, flatness, slope, roll-off, spread, centroid, kurtosis) and cepstral features (MFCC 1-13), along with their first and second derivatives, were extracted, which have demonstrated effectiveness in the analysis of vocal related to PD [87]. Intensity, DFA, STE, and PLP (1-13) along with their first and second derivatives completed the extracted set of features.

Following feature extraction, these diverse features were organized into a unified vector. For each feature, five key statistics were computed, comprising the mean value, median value, standard deviation, kurtosis, and skewness. Notably, jitter and shimmer variants were evaluated across the entire signal since their definitions inherently involve comparisons among contiguous frames. To ensure feature comparability, a min-max normalization procedure was applied to standardize them within a consistent range.

Feature Selection. To achieve the objective of evaluating the effectiveness of various vocal tasks and their combinations, the /a/ and /e/ vowels from the IPVS and PC-GITA datasets were compared; similarly a set of phonetically balanced phrases from the IPVS, ANTHEA-PDSS, PC-GITA, and Suppa datasets was utilized to identify potential differences between different sentences. It is important to note that

the IPVS and ANTHEA-PDSS1 datasets included the utterance of the same set of sentences; hence, after comparing the two vowels, a merging procedure was applied to combine them into a single corpus. Due to differences in the data collection procedure, a stratified fusion approach was employed to maintain a consistent proportion of both datasets in any subsequent split.

The best-performing vowels and phrases resulting from this analysis were further compared to assess whether a single task or their combination could more effectively distinguish between HCs and PDPs. The combination of these two tasks was achieved through an early fusion of the associated features. Also in this case, the results from different datasets were considered to test the robustness of the findings. It is worth noting that the Suppa dataset exclusively contained the phonation of a single vowel (/e/), automatically making it the most significant for the corpus.

Given the balanced distribution of the datasets used in this study, the effectiveness of each vocal task was compared based on the accuracy achieved by a binary classification model trained with features extracted from the specific task. The pipeline used for this analysis included a feature selection step that utilized the boruta algorithm [133] and a subsequent classification step involving several classic ML models.

Classification. As for the binary classification between PDP and HC, ten different algorithms were compared, including KNN, SVM, Gaussian Process (GP), DT, Random Forest (RF), Artificial Neural Network (ANN), NB, Linear Discriminant Analysis (LDA), AdaBoost (ADA), and XGB. To ensure a fair comparison between vocal tasks and minimize the impact of the classification model choice, the average performance of these algorithms was employed for comparative analysis.

To avoid overfitting and ensure robust model performance, a two-phase approach was employed. In the first phase, feature selection and model training were carried out using 70% of the original dataset, leaving the remaining 30% unaltered for testing purposes. Moreover, to assess the model generalization capability, a 10-fold CV technique was applied during the training phase, while the remaining 30% of subjects were employed to test the performance of the model on previously unseen samples. The vocal task yielding the best performance was assessed through the accuracy metric. F1-score, Sensibility, Specificity, and Area Under the Curve (AUC) were also computed for comparison purposes.

Figure 4.4 illustrates the workflow of the experiment.

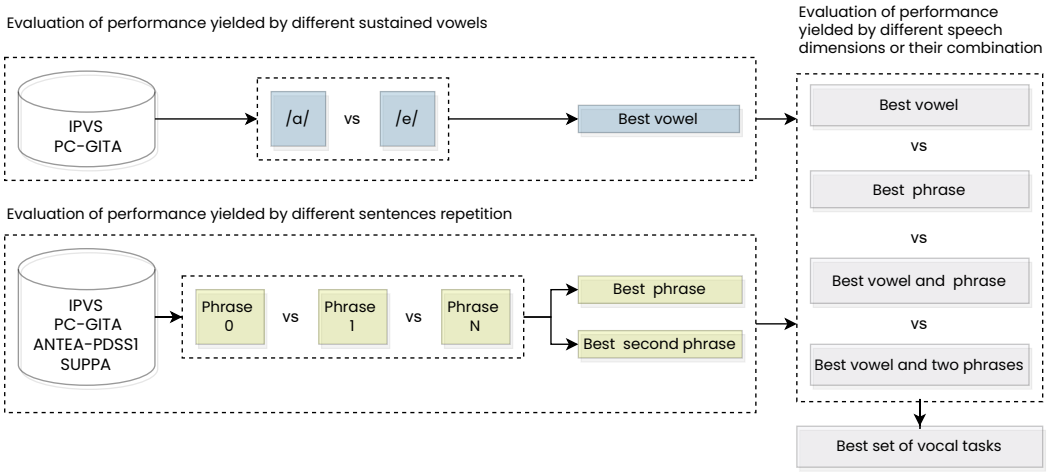


Fig. 4.4 Workflow of the methodology employed for comparing task effectiveness

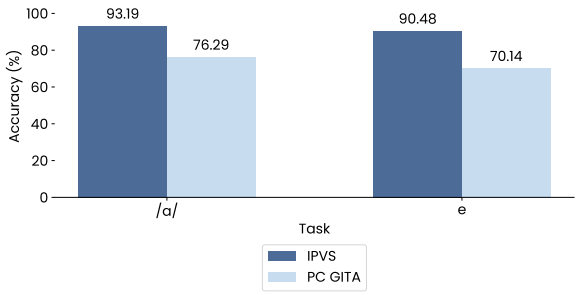
Results

Comparison Between Vocal Tasks. In Figures 4.5a and 4.5b, the results of the initial comparison between different vocal tasks are presented. These figures display the classification accuracy for two separate experiments: one focused on vowel comparisons and the other on phrase comparisons. Moving on to Figure 4.5c, the results of the comparison among the best-performing vocal tasks and their combination are provided.

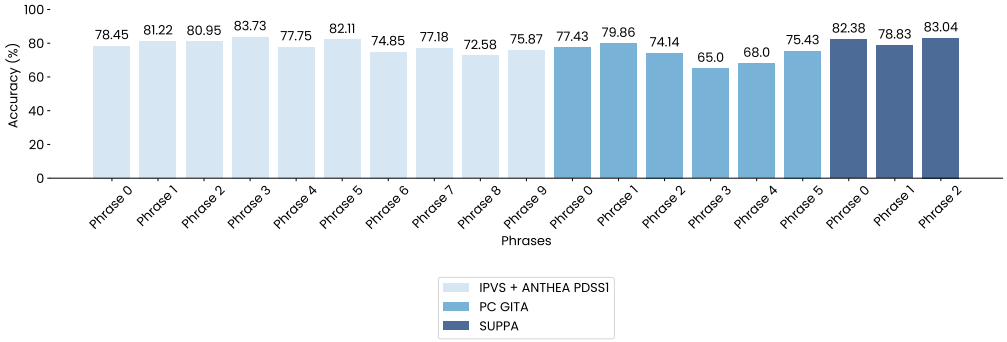
Classification. As for the classification step, the comparison between the different models tested revealed that the top-performing models were KNN and GP, both achieving an average accuracy of 91% on validation sets. These two models underwent the computation of a comprehensive set of metrics on both validation and test sets, in order to evaluate their ability to classify new and previously unseen samples. The detailed results can be found in Table 4.15.

Discussion

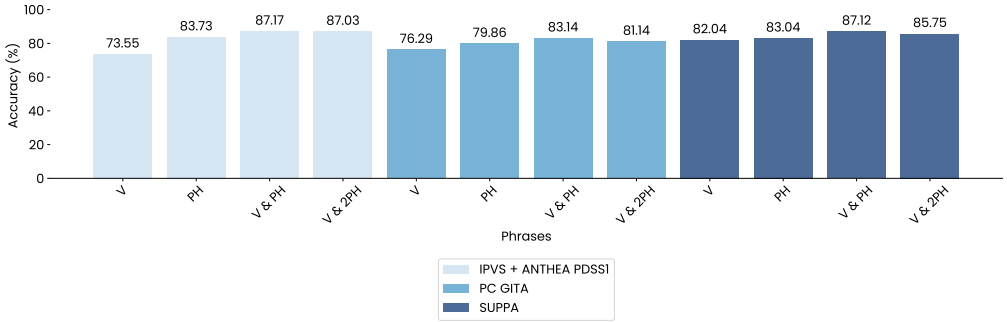
Comparison Between Vocal Tasks. The findings across the three datasets strongly support using a combined approach involving sustained vowel phonation and phonetically balanced phrase repetition to characterize PD-related vocal changes. This approach significantly improved performance, up to 13.6% compared to single vowel



(a) *Vowels*



(b) *Phrases*



(c) *Task Combination. V:Best Vowel; Ph: Best Phrase; 2Ph: Best Two Phrases.*

Fig. 4.5 Results from the comparison of effectiveness across various tasks, adapted from [120]

phonation and up to 5.08% compared to single phrase repetition, as shown in Figure 4.5c. Moreover, despite the complex nature of the articulatory process, the evidence clearly indicated that the use of multiple sentences did not enhance the distinction between HCs and PDPs.

Table 4.15 Classification performance of the top two classifiers

	10 fold CV	Test set				
	Accuracy	Accuracy	Specificity	Sensitivity	AUC	F1-score
			PC-GITA			
KNN	0.90	0.73	0.71	0.77	0.74	0.71
GP	0.90	0.73	0.65	0.85	0.75	0.73
			Suppa			
KNN	0.92	0.81	0.78	0.82	0.81	0.82
GP	0.94	0.74	0.71	0.76	0.74	0.76
			IPVS & ANTHEA-PDSS1			
KNN	0.91	0.88	1.0	0.79	0.89	0.88
GP	0.89	0.88	0.90	0.86	0.88	0.89

Table 4.16 Features selected for each binary classification

	IPVS-ANTHEA-PDSS1
Phrase	Shimmer; DFA; HNR; Spectral features: center of gravity, skewness, kurtosis; 1st Formant; MFCC: 2, 13; ΔΔMFCC: 5; PLP: 0; ΔPLP: 4, 6.
Vowel	ΔΔMFCC: 1, 6, 12; Δ PLP: 0; ΔΔPLP: 0
	PC GITA
Phrase	GNE; Spectral features: roll-off, center of gravity; MFCC: 3; PLP: 3; ΔPLP: 3, 5; ΔΔPLP: 1, 3, 6.
Vowel	PLP: 6, 9.
	Suppa
Phrase	Spectral features: center of gravity, decrease, slope; 1st Formant; 3rd Formant; MFCC: 4, 7; ΔΔMFCC: 3, 5; PLP: 2, 10; Δ PLP: 2.
Vowel	STE; Spectral features: roll-off, flux; PLP: 3; ΔPLP: 9

In specific tasks, vowel /a/ consistently outperformed /e/, as seen in Figure 4.5a, both in the IPVS and PC GITA datasets. Notably, when the ANTHEA-PDSS1 dataset was included in the IPVS corpus, a performance drop associated with single vowel phonation was observable, emphasizing the need for larger datasets to ensure statistically robust results. These findings also underscore the impact of data collection methodology, consistent with prior research in [134] and [14].

Regarding the analysis of the most effective phrases (Fig. 4.5b), *phrase 3: Patrizia ha pranzato a casa da Fabio* performed best in the IPVS and ANTHEA-PDSS1 datasets, while *phrase 8: Marco non è partito* was associated to the worst

performance. This underscores the need for more complex tasks to effectively capture vocal alterations and emphasized the importance of incorporating occlusive and fricative sounds, as in previous studies [14, 135]. Similarly, the comparison between *phrase 1: Omar, que vive cerca, trajo miel* and *phrase 3: Los libros nuevos no caben en la mesa de la oficina* in the PC GITA dataset, confirmed than better performance is achieved with sentences containing complex, articulated sounds and occlusive consonants.

Classification & Feature Selection. The classification performance results, obtained by combining the best-performing vowel and sentence (Table 4.15), revealed KNN and GP algorithms as the top classification models, achieving classification accuracy ranging from 88.7% to 94.5% in a 10-fold CV. Notably, there was no significant degradation in performance when transitioning to test set, indicating strong generalization capabilities. Furthermore, the models exhibited consistent performance across three different datasets, demonstrating their robustness.

Regarding the selected features, the boruta algorithm returned distinct feature sets for the three datasets (Table 5.6), with a notable prevalence of phrase-derived features. A comparison among these three subsets highlights the effectiveness of features such as spectral center of gravity, MFCC, and PLP coefficients.

4.5 Experimental Findings: Acoustic Features Effectiveness

As emerging from the related literature (Section 4.2) several approaches have been proposed to parameterize vocal samples by means of acoustic features related to phonatory, articulatory, and prosodic dimensions. In this context, several recent review studies have undertaken comprehensive analyses to outline the landscape of interpretable features proposed over the years; however, most of them emphasize the absence of a validated acoustic model [82, 83, 86]. Furthermore, while numerous authors concur on the impact of language on vocal analysis models, only a limited number of studies have examined parametrization techniques applied to vocal samples from native Italian speakers, thus raising questions about the generalizability of the proposed models within this specific population.

Within this context, the experiments conducted in this section were primarly devoted to perform an in-depth review of the features proposed thus far for characterizing vocal impairment in PD subjects. The goal was to explore the effectiveness demonstrated in different studies and identify areas of consensus or divergence among them. Additionally, novel acoustic approaches were introduced to capture additional facets of vocal alterations in individuals with PD. These analyses placed particular emphasis on Italian native speakers, seeking to shed light on specific phonetic alterations and validate previous findings within this language.

4.5.1 Review of Acoustic Features in PD Classification

This study aimed to fill the absence of evidence regarding the effectiveness of acoustic features, as emerging from the literature review performed (Section 4.2). Therefore, this section presents the results of a comprehensive review of ML and statistical-based voice analysis models that were used to address vocal alterations associated with PD. The primary objective was to offer valuable insights into the current state of research in vocal analysis related to PD, making it a valuable resource for both researchers and clinicians in this field.

Going into further detail, to establish a baseline understanding, the state-of-the-art was explored by analyzing and discussing a total of 102 research papers selected from the principal electronic databases. A statistical assessment was eventually performed in order to identify the most commonly used features and those deemed most effective from the results reported in the studies. Furthermore, an overview of the algorithms employed was provided, along with information about the public datasets, toolboxes, and general metadata that could potentially serve to enhance the understanding of feature importance and their effectiveness under specific conditions. The results of this work are published in [136].

Materials & Methods

The literature research was conducted in March 2022 employing four different electronic databases, including IEEE Xplore, PubMed, Elsevier, and Web of Science. The following search string was used: *((Parkinson OR Parkinson's disease) AND (speech OR voice OR vocal) AND (feature OR biomarker OR marker))*.

The initial database search produced a total of 1,190 articles. Following the assessment of their relevance through the examination of titles, abstracts, and keywords, the investigation focused on studies that addressed the automatic assessment of vocal alterations in PDPs by analyzing interpretable acoustic features. Notably, research that relied on deep feature-extraction techniques was excluded. Furthermore, to uphold the review robustness, studies employing datasets with a limited sample size (<25 subjects per class), were also omitted. The inclusion criteria encompassed original, peer-reviewed journal articles, and reviews published between January 2017 and March 2022. Only journals published in English within the fields of medicine, biomedical science, and engineering were taken into account, with conference papers, books, book chapters, and letters being excluded. In order to ensure the dataset quality and integrity, duplicate entries were subsequently eliminated. Ultimately, a total of 102 articles were identified as suitable for the research scope, and from each of these, the following information was tabulated: study ID (authorship and year), aim of the work, recording modality (i.e., professional microphone or low-quality equipment), dataset cardinality, participants' demographic, set of features employed, toolboxes used for feature extraction, model employed for classification step.

Results

Table 4.17 reports the tabulated information extracted from the 102 articles investigated.

Table 4.17 Compilation of relevant information from reviewed papers. A1:PD vs HC; A2:Staging; A3:RBD vs PD; A4:Others, HQ: Professional Microphones ; LQ:Smartphones & Laptops n.r.: Not Reported; n.a: Not Applicable; c.r.; Custom Routines; T1: Vowel: T2: DDK; T3; Reading; T4: Monologue

Ref.	Aim	Task	Data	Device	Features	Tools	Model
[137]	A1	T1	(1)[114] (2)[115]	(1)HQ (2)n.r.	DFA, EMD, F0, Form., GNE, GQ, HNR, I, Jitt., MFCC, NHR, PPE, RPDE, Shimmer, TQWT, VFER	Praat, VAT	SVM
[138]	A1	T1	[117]	n.r.	DFA, GNE, HNR, Jitt., MFCC, NHR, PPE, RPDE, Shimm.	n.r.	BT

Table 4.17 continued

[139]	A1, A2	T1, T2, T3, T4	117PD, 41RBD, 98HC	HQ, LQ	DPI, F0, I, Jitt., NHR, NP, NSP, Rhythm, Shimm.	Praat, c.r.	SVM
[140]	A1	T1	[115]	n.r.	DFA, EMD, F0, Form., GNE, GQ, HNR, I, Jitt., MFCC, NHR, PPE, RPDE, Shimm., TQWT, VFER, WT	Praat, VAT	SVM
[141]	A1, A2	T1, T2, T3, T4	149PD, 150RBD, 149HC	HQ	HNR, DDK, DPI, F0, I, NSR, VOT	c.r.	n.a.
[142]	A1	T1	[118]	LQ	A, DFA, Entropy, F0, GQC, GQO, HNR, Jitt., MFCC, OQ, RPDE, Shimm.	c.r.	n.a.
[143]	A1, A2	T1, T2, T3, T4	48PD, 48HC	HQ	CPP, DDK, DPI, F0, HNR, I, MPT, RFA, RLR, VOT	c.r.	n.a.
[144]	A1, A2, A3	T1, T3, T4	90PD, 60RBD, 60HC	HQ	CPP, CPPS	c.r.	n.a.
[145]	A1, A4	T1	80PD, 40HC	HQ	Autocorr., F0, HNR, Jitt., MPT, NHR, Pulse, Shimm., Voicing	Praat	SVM
[146]	A4	T1	51PD, 11HC	HQ	CPP, GFCC, HNR, Jitt., LPC, MS Area, RPDE, Shimm., SRMR	Praat, Darth, c.r.	n.a.
[102]	A1	T1	(1)30PD, 30HC (2)[118]	LQ	CPP, D2, Entropy, GNE, GQ, GQC, GQO, HNR, Hurst, Jitt., LZ-2, MFCC, MFSW, PPE, Shimm., ZCR	c.r.	SVM, ANN
[147]	A1	T4	262PD, 464HC		DFA, F0, HNR, Jitt., MFCC, PPE, RPDE, Shimm.	Praat, c.r.	XGB

Table 4.17 continued

[148]	A1	T1	[118]	LQ	Chroma Feat., Energy, Entropy of En, MFCC, Spect. Feat., ZCR	c.r.	XGB
[149]	A1, A2	T1, T2, T3	100PD, 100HC	HQ	CPP, DDK, DPI, F0, HNR, I, NSR, PSI, RFA, VOT	c.r.	n.a.
[150]	A1	T1	1078PD, 5453HC	LQ	DFA, EMD, F0, GNE, GQ, HNR, Jitt., MFCC, NHR, PPE, RPDE, Shimm., VFER, WT	Darth	SVM
[151]	A1	T1	54PD, HC n.r.	LQ	D2, DFA, F0, HNR, Jitt., NHR, PPE, RPDE, Shimm., Spread	c.r.	KNN
[152]	A1, A2	T1	100PD, 101HC	HQ	DUV, F0, Jitt., MP, NHR, PFR, Shimm., SPI, VTI	c.r.	n.a.
[153]	A1	T3	45PD, 45HC	HQ	Gini index, SHP, Spect. Feat.	n.r.	SVM
[14]	A1	T3	(1)IPVS (2)26PD 18HC	(1)HQ (2)LQ	DFA, DR, ET, I diff., MFCC, RASTA-PLP, Spect. Feat.	n.r.	SVM
[154]	A1, A3	T1	335PD, 112RBD, 92HC	LQ	DFA, EMD, F0, GNE, GQ, HNR, Jitt., MFCC, NHR, PPE, RPDE, Shimm., VFER, WT	Darth	RF
[155]	A1, A3	T1, T2, T3	30PD, 30RBD, 30HC	HQ, LQ	DDK, DPI, F0, HNR, I, Jitt., RFA, RST, Shimm., VOT	Praat, c.r.	n.a.
[156]	A1	T1	[115]	n.r.	DFA, EMD, F0, For, GNE, GQ, HNR, I, Jitt., MFCC, NHR, PPE, RPDE, Shimm., TQWT, VFER, WT	Praat, VAT	SVM
[157]	A1	T1, T2	54PD, HCn.r.	LQ	DDK, Energy, F0, FCR, FTA, Jitt., NTA, PTA, Shimm., TrB, VF, VSA	Bio MetR Tools	n.a.

Table 4.17 continued

[158]	A1, A3, A2	T3	(1)35PD, 32HC (2)50PD, 50HC (3)8PD, 7HC	(1), (2)HQ (3)n.r.	F2i/F2u, FCR, VAI, VSA	c.r.	n.a.
[159]	A1	T1	[115]	n.r.	DFA, EMD, F0, Form., GNE, GQ, HNR, I, Jitt., MFCC, NHR, PPE, RPDE, Shimm., TQWT, VFER, WT	Praat, Darth	NN
[160]	A1	T1	[115]	n.r.	DFA, EMD, F0, Form., GNE, GQ, HNR, I, Jitt., MFCC, NHR, PPE, RPDE, Shimm., TQWT, VFER, WT	Praat, Darth	CNN
[161]	A1	T1	37PD, 35HC	LQ	MFCC	c.r.	Elastic net
[162]	A1	T1	[116]	HQ	Autocorr., F0, HNR, Jitt., NHR, Pulse, Shimm., Voicing	Praat	NN
[163]	A1	T1	(1)110P, 93HC (2)50PD, 50HC	HQ	LFCC, MFCC	c.r.	SVM
[164]	A1	T1, T3	30PD, 15HC	LQ	BBE, MFCC	Neuro speech	Bi-LSTM
[165]	A1	T1	[115]	n.r.	DFA, EMD, F0, Form., GNE, GQ, HNR, I, Jitt., MFCC, NHR, PPE, RPDE, Shimm., TQWT, VFER, WT	Praat, Darth	SVM
[166]	A4	T3	38PD, 37others	n.r.	MFCC, PLP	c.r.	n.a.

Table 4.17 continued

[167]	A2	T1	(1)86 PD (2)[42]	(1)n.r. (2)LQ	CPP, DFA, EMD, En., F0, Form., GNE, GQ, HNR, I, Jitt., MFCC, NHR, PPE, RPDE, Shimm., SoE, TQWT, VFER, WT	Praat, Darth, Voice Sauce	XGB
[168]	A1	T1	(1)[116] (2)31 PD	HQ	Autocorr., F0, HNR, Jitt., NHR, Pulse, Shimm., Voicing	Praat	SVM
[169]	A1, A2	T1	(1)[116] (2)[42]	HQ	Autocorr., DFA, EMD, F0, GNE, GQ, HNR, Jitt., MFCC, NHR, NHR, PPE, Pulse, RPDE, Shimm., VFER, Voicing, WT	Praat, c.r.	Various
[170]	A2	T1	36 PD	LQ, HQ	CPP, D2, MFCC, RPDE	n.r.	n.a.
[171]	A1	T2, T3	34 PD, 25HC	n.r.	DDK, F0, Form., I, Loudness, SPIR	Praat, c.r.	n.a.
[172]	A1	T1	[115]	n.r.	DFA, EMD, F0, Form., GNE, GQ, HNR, I, Jitt., MFCC, NHR, PPE, RPDE, Shimm., TQWT, VFER, WT	Praat, Darth	SVM
[173]	A1	T1	[115]	n.r.	DFA, EMD, F0, Form., GNE, GQ, HNR, I, Jitt., MFCC, NHR, PPE, RPDE, Shimm., TQWT, VFER, WT	Praat, Darth	Various
[174]	A1, A4	T1	(1)[116] (2)[114] (3)90PD	HQ	Autocorr., D2, DFA, F0, HNR, Jitt., NHR, PPE, Pulse, RPDE, Shimm., Spread, Voicing	Praat, c.r.	SVM, RD, XLM
[175]	A1	T1, T3	(1)50PD, 50HC (2)20PD, 20HC	HQ	DFA, Form., GNE, GQC, HNR, IEDCC, IMFCC, Jitt., MFCC, NHR, PPE, RPDE, Shimm., VFER	n.r.	SVM

Table 4.17 continued

[176]	A1	T1	[118]	LQ	BE, Energy, FER, F0, Form., HI, Harmonicity, HNR, Jitt., I, MFCC, Shimm., Spect. Feat., Voicing, ZCR	OS	GB
[177]	A1	T2	50PD 50HC	HQ	EMD	n.r.	SVM
[178]	A1, A2	T1, T4	(1)[114] (2)[42]	HQ	F0, D2, DFA, HNR, Jitt., NHR, PPE, RPDE, Shimm., Spread	Praat, c.r.	GP
[179]	A1	T1	(1)[115] (2)[116] (3)[117]	(1)n.r. (2)HQ (3)n.r.	DFA, EMD, F0, Form., GNE, GQ, HNR, I, Jitt., MFCC, NHR, PPE, RPDE, Shimm., TQWT, VFER, WT	Praat, Darth, n.r.	DT
[180]	A1	T1	[115]	n.r.	DFA, EMD, F0, Form., GNE, GQ, HNR, I, Jitt., MFCC, NHR, PPE, RPDE, Shimm., TQWT, VFER, WT	Praat, Darth	XGB
[181]	A1	T1, T3	(1)50PD, 50HC (2)20 PD, 20HC	HQ	Entropy, GNE, GQ, HNR, Jitt., MFCC, NHR, NMF, Shimm., NMF, VFER	c.r.	SVM
[182]	A1	T3, T4	40PD, 40HC	HQ	BBE, DPI, DUV, DR, DVI, ET, F0, GFCC, Jitt., MFCC, NVS, Posterior probabilities	c.r.	SVM
[183]	A1	T1	[115]	n.r.	DFA, EMD, F0, Form., GNE, GQ, HNR, I, Jitt., MFCC, NHR, PPE, RPDE, Shimm., TQWT, VFER, WT	Praat, Darth	LDA
[184]	A1	T1, T3	(1)50PD, 50HC (2)20 PD, 20HC	HQ	HCC, VMD	c.r.	MLP, RF, SVM

Table 4.17 continued

[185]	A1	T1	(1)160 PD, 100HC (2)[116]	(1)LQ (2)HQ	MFCC	n.r.	SVM
[186]	A4	T1, T4	50PD	HQ	F0, I, Jitt., MPT, NHR, Shimm., Voicing	Praat	n.a.
[187]	A1, A4	T3	80PD, 140 HC, others	HQ	Form., MBE, MFCC	Praat	CNN
[188]	A1	T1, T2, T3, T4	50PD, 50HC	HQ	AQ, BBE, ClQ, Energy, F0, Form., Harmonicity, HRF, Jitt., MFCC, NAQ, OQ, PSP, QOQ, Shimm., SQ	Neuro speech, APARAT	SVM
[189]	A1	T1	(1)28PD,25HC (2)40HC, 40PD.	HQ	EMD	c.r.	SVM, RF
[190]	A1	T1	(1)[115] (2)[116]	(1)n.r. (2)HQ	Autocorr., DFA, EMD, F0, Form., GNE, GQ, HNR,I, Jitt., MFCC, NHR, PPE,Pulse, RPDE, Shimm., TQWT, VFER, Voicing, WT	Praat, Darth, n.r.	LDA
[191]	A1	T1	[117]	n.r.	DFA, GNE, HNR, Jitt., MFCC, NHR, PPE, RPDE, Shimm.	n.r.	GB
[192]	A1	T1	[117]	n.r.	DFA, GNE, HNR, Jitt., MFCC, NHR, PPE, RPDE, Shimm.	n.r.	KNN
[193]	A1, A4	T1	50PD, 100HC	HQ	CHNR, FCR, Form., GNE, HNR, Jitt., MFCC, NNE, Shimm., VSA	n.r.	SVM

Table 4.17 continued

[194]	A1, A4	T1	50PD, 50HC, others	HQ	Entropy, D2, DFA, Entropy, Hurst, LLE, LZ-2, RPDE	c.r.	SVM
[195]	A1, A4	T1	30PD, 20other	LQ	Autocorr., F0, Form., HNR, I, Jitt., MFCC, NHR, Pulse, RASTA-PLP, Shimm., Voicing	Praat	KNN
[196]	A1, A2	T1	(1)[42] (2)[114]	(1)LQ (2)HQ	DFA, HNR, Jitt., NHR, PPE, RPDE, Shimm., Spread	Praat, c.r.	SVM
[197]	A1	T1	[117]	n.r.	DFA, GNE, HNR, Jitt., MFCC, NHR, PPE, RPDE, Shimm.	n.r.	LR, SVM, KNN
[198]	A1, A2	T1	55HC, 320PD	LQ	DFA, F0, GQ, HNR, Jitt., NHR, PPE, RPDE, Shimm.	Darth	SD
[199]	A2	T1	[42]	LQ	DFA, HNR, Jitt., NHR, PPE, RPDE, Shimm., Spread	Praat, c.r.	LR
[200]	A1, A3	T3, T4	80PD 50 RBD.	HQ	AST, DPI, DUF, DUS, DVI, EST, GBIV, LRE, PIR, RLR, RSR, RST	c.r.	n.a.
[201]	A1	T1	94 PD, 8HC	HQ	Form.	n.r.	n.a.
[202]	A1	T1	(1)[114] (2)[116]	HQ	Autocorr., F0, HNR, Jitt., NHR, Pulse, Shimm., Voicing	n.r.	KNN
[203]	A1, A2	T1	40PD, 15HC	HQ	Entropy, WT	c.r.	ELM
[**?**]	A1	T2	50PD, 50HC	HQ	D2, Entropy, Hurst, LLE, PE, RPDE	n.r.	SVM
[94]	A1	T2	27PD, 27HC	HQ	VOT	c.r.	SVM
[204]	A1	T1	[115]	n.r.	TQWT	Praat, Darth	KNN

Table 4.17 continued

[156]	A1	T1	[115]	n.r.	DFA, EMD, F0, Form., GNE, GQ, HNR, I, Jitt., MFCC, NHR, PPE, RPDE, Shimm., TQWT, VFER, WT	Praat, Darth	KNN, SVM, DT
[129]	A1	T3	(1)50PD,HQ 50HC (2)20PD, 20HC		BBE, ET, F0, MFCC, PTS, Spect. Feat., ZCR	n.r.	SVM
[205]	A1	T1	205PD, 74HC	HQ	HNR, NHR	c.r.	n.a.
[206]	A1	n.r.	44PD, HCn.r.	n.r.	F0, D2, DFA, HNR, Jitt., NHR, Spread PPE, PPQ, RPDE, Shimm.	n.r.	Various
[207]	A1, A2	T1	(1)[114] (2)[42] (3)48PD, 20HC (4)4PD	n.r.	F0, D2, DFA, HNR, Jitt., NHR, PPE, RPDE, Shimm.		SVM, LR, MLP
[208]	A1	T1	1483PD, 8300HC	LQ	DFA, EMD, F0, GQ, GNE, HNR, Jitt., MFCC, NHR, PPE, RPDE, Shimm., VFER, WT	c.r.	n.a.
[**?**]	A1	T1, T2, T3	86PD, 50HC	HQ	BBE, DR, DTW, DUV, DVI, FLUF, F0, Form., Jitt., MFCC, NU, NV, RatioDurV/Sig, Shimm., WA	c.r.	Regr.
[209]	A1	T1	[118]	LQ	MFCC	c.r.	SVM
[210]	A1	T1	[118]	LQ	DFA, F0, HNR, Jitt., MFCC, PPE, RPDE, Shimm.	Darth	LR, RF, CNN
[211]	A1, A4	T1	40PD, 40HC, 200other	HQ	AT, FT	c.r.	n.a.

Table 4.17 continued

[212]	A1	T3	50PD 50HC	HQ	Energy, MFCC	n.r.	GMM
[213]	A1	T1	234PD, 50HC	n.r.	Amplitude, AT, F0, FT	Praat	n.a.
[214]	A1	T1	35PD, 45HC	n.r.	Autocorr., HNR, Jitt., MFCC, Shimm.	n.r.	SVM
[215]	A1	T1	45PD, 45HC	HQ	MFCC, PLP	n.r.	ANN
[216]	A1	T1	27PD, 446HC	n.r.	DFA, GQ, HNR, Jitt., MFCC, NHR, PPE, RPDE, Shimm.	Darth, c.r.	SVM, RF
[217]	A1	T1, T3	32PD, 10HC	HQ	CPP	ADSV	n.a.
[218]	A1	T1	30PD, 20HC	HQ	Amplitude, AT, F0, FT	n.r.	n.a.
[219]	A1	T1, T3	30PD, 32HC	HQ	F0, FER	n.r.	n.a.
[220]	A1, A2	T1	320PD, 55HC	n.r.	MFCC, PLP	n.r.	n.a.
[221]	A1, A2	T1	51PD	HQ	BE, FER, F0, FLUF, Form., GNE, HNR, Jitt., NNE, Shimm.	n.r.	XGB
[222]	A1	T1	147PD, 48HC	n.r.	F0, Jitt., PPQ, Shimm., HNR, NHR, RPDE, DFA, D2, PPE, Spread	n.r.	Various
[223]	A1	T1, T2, T3	(1)50PD, 50HC (2)164HC	HQ	LPC, MFCC, RASTA-PLP	n.r.	GMM, GPLDA
[224]	A1	T1	(1)[116] (2)28PD	(1)HQ (2)n.r.	F0, Autocorr., HNR, Jitt., NHR, Pulse, Shimm.	n.r.	SVM
[225]	A1	T1	45PD, 45HC	n.r.	Autocorr., En., HNR, Jitt., NHR, Shimm.	n.r.	Regr., NN
[226]	A1	T1	40PD, 40HC	LQ	GNE, HNR, NHR	Praat, c.r.	n.a.

Table 4.17 continued

[227]	A1, A3	T3	80PD, 50HC	HQ	AST, DPI, DUF, DUS, DVI, EST, GBIV, LRE, PIPR, RLR, RSR, RST	n.r.	SVM
[228]	A1	T1	50PD, 50HC	n.r.	Form.	n.r.	n.a.
[229]	A1	T2	38PD, 38HC	n.r.	DDK, Jitt., Shimm.	n.r.	LR
[230]	A1	T1	147PD, 48HC	HQ	D2, DFA, F0, HNR, Jitt., NHR, PPE, RPDE, Shimm., Spread	n.r.	SVM
[214]	A1	T1	54PD, 45HC	HQ	MFCC	n.r.	RBFN
[231]	A1	T1, T3	115PD, 108HC	HQ	BE, Energy, F0, Harmonicity, HNR, Jitt., I, MFCC, RASTA-PLP, Shimm., Spect. Feat., Voicing, ZCR	OS	SVM

In Figure 4.6, a comparison between the frequency of features (i.e., the number of articles that utilized these features) and their effectiveness (i.e., the number of articles in which these features were reported as effective) is presented. To maintain conciseness, only features that were employed in at least 10% of the papers are included. It is noteworthy that 27 articles did not report this particular information.

Discussion

The primary findings of the literature review revealed that the majority of the analyzed papers focused predominantly on phonatory and articulatory features. Notably, studies by [193] and [232] concurred that the inclusion of the latter, i.e., articulatory features, enhances the system capacity to capture vocal abnormalities in PDPs.

Figure 4.6 compared frequency and effectiveness of the features identified in the review. As expected, the most frequently employed features encompass Jitter, Shimmer, F0, MFCC, HNR, and NHR, which are standard metrics in vocal analysis. Furthermore, alternative processing techniques like DFA and entropy-related features

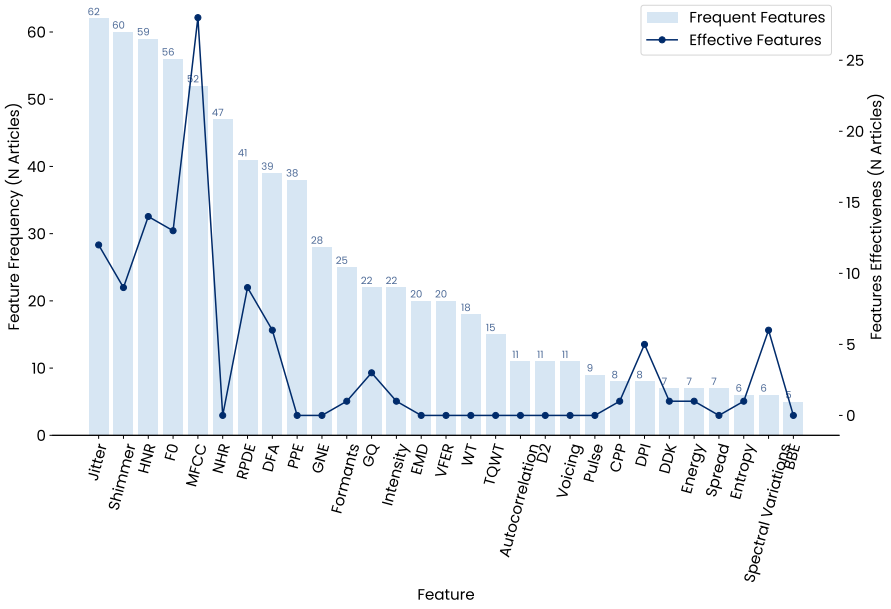


Fig. 4.6 Comparison of features frequency and effectiveness based on literature review

(e.g., RPDE) are prevalent, followed by parametric methods related to phonatory and glottal aspects.

As for feature effectiveness, noise analysis in speech, particularly HNR, demonstrated to be crucial for PD vocal analysis, being closely linked to voice hoarseness, a well-documented trait among PD patients (Section 2.1). In addition to noise-related measures, features related to F0 play a significant role in modeling complex aspects of the phonatory system, especially hoarseness and voice tremor. Furthermore, Jitter and Shimmer are commonly examined, although this review unveiled varying evidence regarding their effectiveness. Indeed, while many authors have demonstrated their efficacy (e.g., [193, 198, 226]) others have found them to be inconsistent indicators [102, 178].

Moreover, non-linear features (e.g., RPDE, DFA, D2) have gained increasing attention in PD vocal parametrization ([102], [14], [194], [196], [**?**], [207]). Nevertheless, while these non-linear features, particularly RPDE, have proven highly effective, two studies [194, 207] concur on the necessity of incorporating them into a wider set of features drawn from diverse domains.

Interestingly, MFCC features emerged as the most effective, distinguishing between HC and PDPs, as well as for UPDRS scoring. Notably, the Cepstral domain has been widely adopted to characterize PD-related vocal alterations, even in the context of complex and noisy audio data. Intriguingly, in studies involving the mPower dataset [118, 119], which comprises data from over 6000 subjects recorded under suboptimal and unsupervised conditions, MFCCs consistently surfaced as a relevant feature group through various feature selection procedures.

Going into further detail among the investigated studies, 57 of them made use of databases recorded with high-quality equipment, often featuring professional condenser microphones equipped with a cardioid polar pattern. In contrast, 25 studies employed databases collected in sub-optimal conditions, such as recordings from smartphones (using omnidirectional electret microphones), laptops, or telephone recordings, taken in both supervised and unsupervised settings. Apart from the evidence reporting MFCC as effective in both type of studies, additional evidence on the influence of the recording modalities were investigated, however, only a small subset of the studies reported clear results on the differences. Among them, Jeancolas et al. [139] observed that there were no significant decrease in performance when transitioning from professional microphones to low-cost equipment. In contrast, [102] reported a marked reduction in performance when utilizing recordings in unsupervised conditions.

Furthermore, studies described in [233, 108, 234] involved comparing features extracted using multiple available toolboxes and different types of recordings, including both smartphone and professional microphone data from the LUHS dataset. Interestingly, despite utilizing the same dataset, these studies do not reach a consensus on the optimal set of parameters to use. However, they all concur on the effectiveness of the YAAFEE toolbox in consistently yielding good classification performance. Additionally, they highlight the necessity of employing distinct sets of features when transitioning from professional microphones to smartphone recordings.

Besides evidence on the impact of recording conditions and routines employed for the extraction, several studies investigated the influence of the participants demographics. Notably, numerous works emphasized the substantial influence of gender on voice production [139, 185, 151, 186**?** , 219]. However, some research works suggested the presence of features that are less susceptible to gender-related effects, such as monopitch [149], Frequency Tremor (FT), and Amplitude Tremor

(AT) [213]. Interestingly, a statistical analysis conducted in [149] revealed that de-novo patients exhibit fewer gender-related differences in PD-related voice characteristics. To integrate this type of meta-information into ML models without the need of extensive stratification, several authors have proposed incorporating non-speech covariates into the feature sets as correction factors. In this context, [102, 235, 167, 236, 237, 192, 197, 148, 238, 207] have all demonstrated the effectiveness of these strategies in mitigating the influence of gender and age, respectively.

It is important to mention that, even though the analysis of UPDRS score, years from diagnosis, and other participant characteristics was carried out for each study, the final results are not reported in this dissertation due to the lack of specific information in the majority of the works. Indeed, since a substantial portion of these studies did not adequately report the required information, statistical analyses were not applied. Nevertheless, the list of available information is provided in detail in [136]

4.5.2 Investigation of Voiced to Unvoiced Transient Regions

This study primarily aimed to assess the effectiveness of an acoustic analysis based on the transient regions (TR) between voiced and unvoiced segments. As discussed in Section 3.1, the lack of coordination of the glottal source, which is typical in PD patients, can lead to difficulties in executing precise and rapid movements. Consequently, a detailed articulatory analysis of this phenomenon may unveil hidden aspects of vocal alterations. Furthermore, given its high specificity, if properly validated, this approach may reveal markers of the alteration that are less susceptible to variations due to the patient's emotional state.

While such a specific parametrization holds potential for articulatory analysis, the pronunciation of specific phonemes varies depending on the language considered. For this reason, this study specifically focused on samples from Italian native speakers, recorded in both optimal and sub-optimal conditions. Subsequently, the work investigated which phonemes in the Italian language are mostly affected by hypokinetic dysarthria and which features are most the suitable for characterizing them. Besides its applicability to PD classification, if properly validated, evidence regarding phonetic misarticulation can offer substantial support to speech therapists

in developing personalized rehabilitation therapy for individual patients, as well as during the follow-up stage.

In Figure 4.7, two pairs of vocal signals are presented, comparing individuals with PD to age and gender-matched HCs. These figures reveal two distinct vocal abnormalities evident in PDPs during the transition between voiceless and voiced speech segments. To delve into further detail, Figures 4.7a and 4.7b showcase a voiceless segment (/s/) occurring after a voiced segment (/e/). It is evident that PDP exhibit substantial difficulty in interrupting vocal cord vibration. This difficulty is manifested in the presence of periodicity, which contrasts with the absence of such periodicity observed in the corresponding HC. Similarly, the second comparison highlights a pronounced diversity between a HC and a corresponding PDP during the pronunciation of the Italian word *sciare* (skiing, /ʃiare/). In particular, a notable distinction is observed during the transition from the voiceless /ʃ/ segment to the subsequent vowel.

These findings underscore significant vocal irregularities in individuals with PD, particularly in their ability to modulate vocal cord vibrations during speech production, which clearly distinguish them from their healthy counterparts. The results from this study are published in [14].

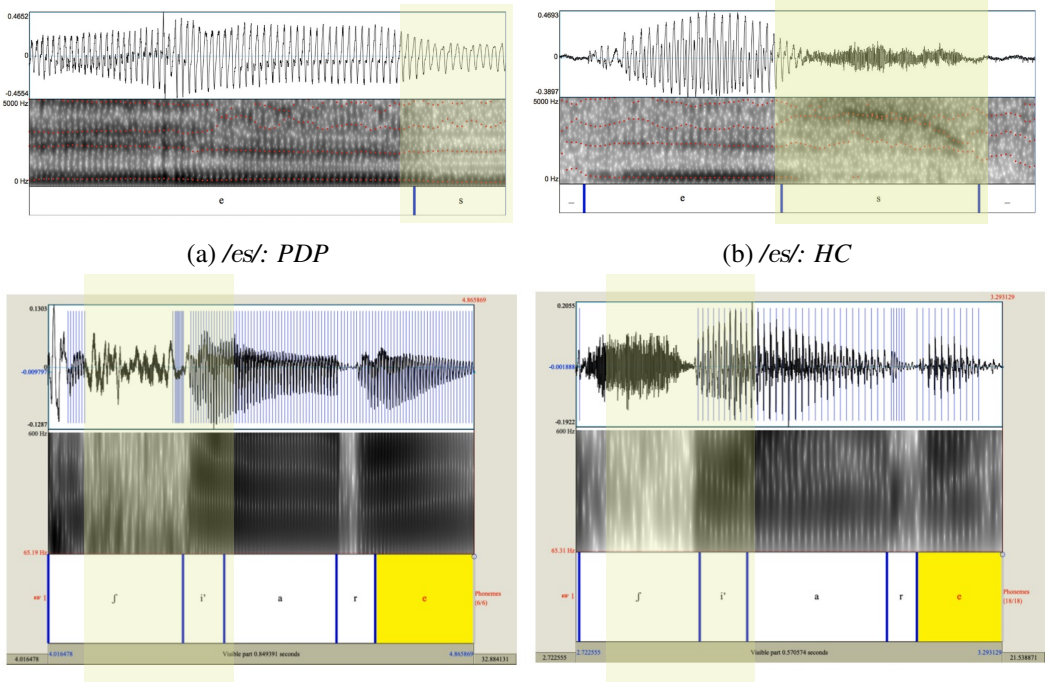
Materials

This study utilized two diverse corpora, namely the IPVS (Section 4.3.1) and the ANTHEA-PDSS1 (Section 4.3.2). Specifically, samples from sentence pronunciation were employed.

Data analysis was performed using Python, where Praat was utilized for pre-processing and feature extraction. The Parselmouth library acted as an interface for accessing Praat internal code.

Methods

Pre-processing. The dataset initially contained recordings with varying sampling rates, which were standardized by down-sampling to 16 kHz to ensure consistent spectral conditions. A de-noising filter with Praat default parameters was then applied to each signal, and amplitudes were normalized to a range between 0 and



(a) */es/: PDP* (b) */es/: HC*

(c) */ʃiare/: PDP* (d) */ʃiare/: HC*

Fig. 4.7 Alterations in transient regions between voiced and unvoiced segments

1 to mitigate any potential influence of speaker-microphone distances on model performance.

The voiced regions in the recordings were manually labeled following their detection using Praat Software. For the analysis of TRs between unvoiced consonants and adjacent voiced segments, manual identification was chosen over automatic segmentation to avoid potential external sources of error. As in Section 4.4.1, TRs were defined as 160 ms long windows centered on the edge of each chunk. The analysis of set of sentences revealed a total of 43 phonetic groups, among which 28 were correctly pronounced by all the individuals. These encompassed various unvoiced consonants in the Italian language, including 13 dental occlusives, 5 velar occlusives, 4 labial occlusives, 2 alveolar sibilants, 1 palatal sibilant, 2 alveolar affricates, and 1 labio- dental fricative.

Feature Extraction. Following the initial pre-processing steps, each TR was further divided into 15 ms frames, with 50% overlap, as outlined in [23]. Subsequently, a comprehensive set of parameters was extracted, incorporating both conventional

phonetic analysis features and innovative parameters capable of capturing variations within the TRs. This set comprised 13 RASTA-PLP coefficients, including their first and second derivatives, which were instrumental in identifying articulatory anomalies associated with specific phonemes. Moreover, the first four spectral moments (mean, standard deviation, skewness, and kurtosis) were employed to model the capacity to rapidly initiate or cease vocal fold vibration. Additionally, 13 MFCC were utilized to detect subtle alterations in articulator motion. The duration ratio (DR) and intensity difference (ID) were used to measure variations between unvoiced consonants and subsequent voiced regions. Lastly, DFA was included to capture increased turbulence due to the lack of control and coordination in vocal fold motion. It is worth nothing that, although DFA is typically applied only to quasi-periodic signals and sentences do not belong to this class, this work focuses on portions of sentences that do not share the same characteristics of the whole signal. In particular, the analysis deals with the TRs, which represent the transition between voiced and unvoiced segments. In addition, the thesis underlying the entire work is the inability of PDPs to promptly start/stop the movements of the vocal folds, implying a perceivable distortion of unvoiced consonants. For this reason, it was dediced to specifically introduce this feature to model the altered periodicity of the vocal signal in PDPs. To complement this feature set, EST feature, as detailed in Section 4.4.1, was introduced to provide a more comprehensive characterization of spectral differences within the TRs.

Feature Selection and Statistical Analysis. After feature extraction, the boruta algorithm [133] was applied to each phonetic group identified during the pre-processing stage. Subsequently, the Pearson correlation coefficient was calculated between the selected features and the class of membership to assess the discriminatory potential of the acoustic parameters. Furthermore, a thorough analysis of the most frequently selected phonetic regions was conducted with the objective of identifying any meaningful pattern among the phonemes that exhibited the highest capacity of describing vocal alterations in PD.

Classification. Leveraging the diversity of phonemes pronounced by the same subjects, the classifier input was created by performing an early fusion of the features extracted from all examined segments into a single vector (Section 4.4.1). Subsequently, seven classifiers were compared, including NB, KNN, SVM, RF, as well as ensemble methods such as GB, BAG, and ADA models. The classification accuracy

was used to compare the classifier performance. Given the inherent randomness introduced by the validation process random splitting procedure, each experiment was repeated 20 times on 20 randomly selected subsets. The average accuracy served as the primary metric for classifier comparison. Following the selection of the optimal classifier and fine-tuning of its hyperparameters, the model stability was further assessed by evaluating accuracy, F1 score, precision, and recall as averages over 20 iterations.

It is worth noting that a dual experiment was conducted to assess the impact of data collection modalities on the results. In the first iteration, the entire pipeline, which encompassed feature extraction, statistical analysis, and classification, was applied to a dataset consisting solely of samples from the IPVS dataset, which were recorded under optimal conditions. In the second iteration, the analysis was extended to include samples from the ANTHEA-PDSS1 dataset, recorded under suboptimal conditions. To ensure the robust generalization of the model in both experiments, the databases were split into two subsets: one for training and validation (80%) and another for testing (20%). It is important to emphasize that for both the IPVS and ANTHEA-PDSS1 datasets, a balanced splitting strategy was applied, ensuring an equal representation of the two datasets in both the training and testing subsets.

Results

Feature Effectiveness and Phonetic Groups Examination. Figure 4.8 provides an overview of the count of phonemes that exhibited statistically significant features, as identified by their selection through the boruta algorithm and their correlation with the class ($p < 0.001$). Within the IPVS corpus, 28 phonetic segments presented at least two features that correlate with the membership class ($0.51 < |r| < 0.86$). Notably, the DFA coefficient extracted from the transition between the occlusive sound /p/ and the vowel /e/, as well as the fifth MFCC from the region between the sibilant /ʃ/ and the vowel /i/, exhibited the highest correlation ($r = 0.85, p < 0.001$). As for the combined corpora, a reduction in overall performance is observable, with correlation coefficients ranging from 0.37 to 0.62 (absolute values). Nevertheless, several features still maintain significant correlations with the membership class. Among these, the DFA derived from the transition between the occlusive consonant /t/ and the vowel /a/ yielded the highest correlation coefficient of 0.62 ($p < 0.001$).

Figure 4.9 depicts the count of features chosen for each phoneme class, presented in absolute values (on the left axis) and as a percentage (on the right axis) relative to the total number of features specific to the given segment type.

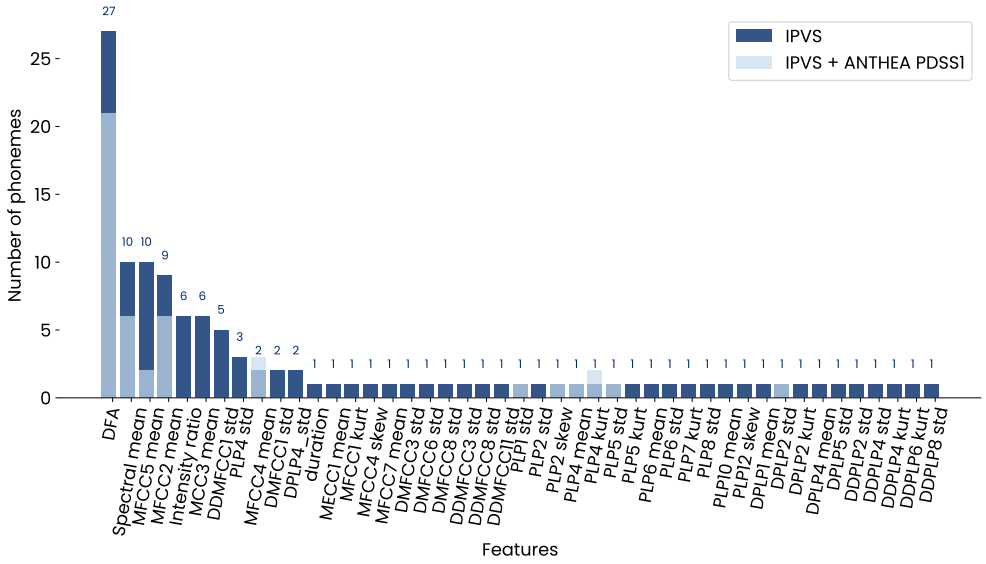


Fig. 4.8 Analysis of feature effectiveness within different phonemes. The reported features were selected via the boruta algorithms and exhibited a significant correlation with the membership class ($p < 0.001$)

Classification. Table 4.18 presents a comparison of classification accuracy of the seven models tested. During the hyperparameter optimization phase, the C parameter was fine-tuned within the range of [10, 100, 1000], while the gamma parameter underwent variations at 0.1, 0.001, and 0.0001. Additionally, the performance of SVMs were analyzed using linear, polynomial, and RBF kernels. Finally, the most effective configuration was identified as SVM with C = 10, gamma = 0.001, and the RBF kernel. The performance results of this optimized SVM on both the validation and test sets can be found in Table 4.19. Furthermore, details regarding the phonetic groups employed and the selected feature types are provided for reference.

In the second experiment, which involved merging the two distinct corpora, Table 4.20 compares the classification accuracy of the seven tested models. These values result from 10-fold CV, averaging over 20 iterations. Following classifier selection, grid-search hyperparameter optimization was performed within the training set,

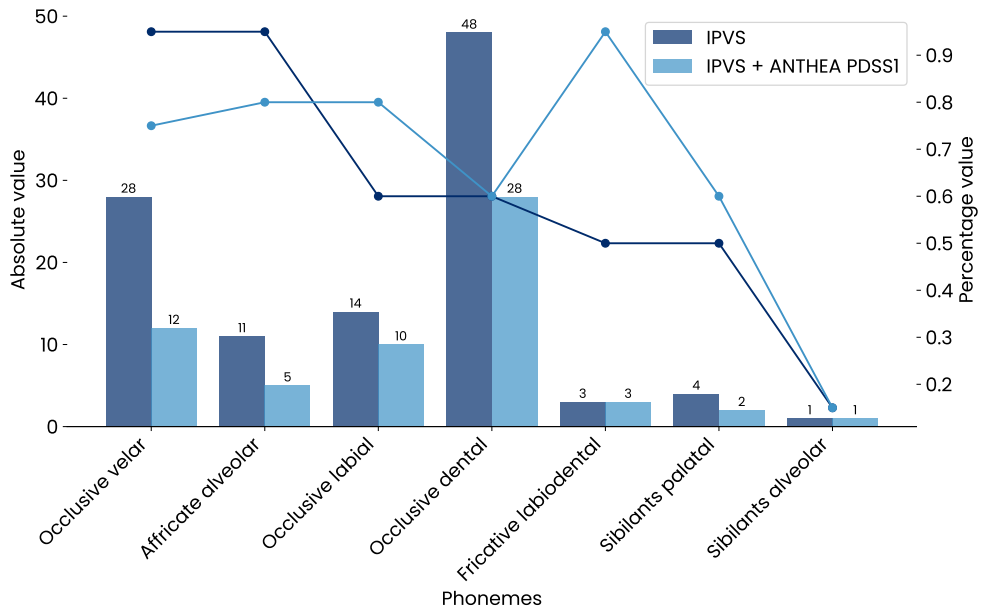


Fig. 4.9 Analysis of phonemes effectiveness within various features. The figures reports the number of features for each phoneme that were selected via boruta algoritm and exhibited significant correlation with the membership class ($p < 0.001$)

Table 4.18 Comparison of classification accuracy among seven classifiers using a 10-fold cross-validation on the IPVS corpus

Model	Accuracy	Precision	Recall	Specificity	F1-score	AUC
NB	0.92±0.03	0.95±0.02	0.94±0.03	0.91±0.04	0.93±0.03	0.98±0.02
KNN	0.96±0.02	0.95±0.03	1.00±0.00	0.92±0.06	0.97±0.02	0.99±0.00
SVM	0.98±0.01	0.98±0.02	1.00±0.00	0.97±0.02	0.98±0.01	1.00±0.00
ADA	0.88±0.05	0.90±0.05	0.92±0.04	0.83±0.08	0.89±0.04	0.92±0.04
GB	0.87±0.03	0.89±0.03	0.93±0.04	0.82±0.07	0.88±0.03	0.90±0.05
BAG	0.97±0.02	0.96±0.03	1.00±0.00	0.93±0.05	0.98±0.02	0.99±0.01
RF	0.96±0.02	0.96±0.02	0.98±0.02	0.93±0.03	0.96±0.02	0.99±0.05

revealing the most effective configuration as SVM with C = 100, gamma = 0.001, and an RBF kernel. The results of the optimized model are provided in Table 4.21.

Table 4.19 Performance details of the optimized SVM model on the IPVS corpus. Results are expressed as an average over 20 iterations.

Metric	Validation set	Test set	Phonetic groups	Type of features
Accuracy	0.98 ± 0.01	0.97 ± 0.06	dental occlusive, labial occlusive, palatal sibilant, velar occlusive	DFA, spectral mean, MFCC2, MFCC5
Precision	0.98 ± 0.02	0.96 ± 0.07		
Recall	0.99 ± 0.02	1.00 ± 0.00		
Specificity	0.97 ± 0.02	0.93 ± 0.14		
F1 score	0.98 ± 0.01	0.98 ± 0.04		
AUC	1.00 ± 0.00	0.96 ± 0.01		

Table 4.20 Comparison of classification accuracy among seven classifiers using a 10-fold cross-validation in the IPVS and ANTEA PDSS1 corpora

Model	Accuracy	Precision	Recall	Specificity	F1-score	AUC
NB	0.83±0.02	0.89±0.03	0.82±0.04	0.85±0.04	0.84±0.02	0.89±0.03
KNN	0.84±0.03	0.87±0.04	0.85±0.03	0.82±0.04	0.85±0.03	0.92±0.02
SVM	0.86±0.03	0.90±0.03	0.87±0.03	0.86±0.04	0.87±0.02	0.94±0.02
ADA	0.84±0.04	0.88±0.04	0.85±0.03	0.83±0.05	0.85±0.03	0.91±0.04
GB	0.76±0.04	0.82±0.04	0.77±0.04	0.75±0.06	0.77±0.04	0.83±0.03
BAG	0.84±0.03	0.88±0.03	0.85±0.03	0.83±0.04	0.85±0.02	0.93±0.02
RF	0.83±0.03	0.87±0.03	0.84±0.03	0.82±0.03	0.84±0.03	0.93±0.02

Table 4.21 Performance details of the optimized SVM model in the IPVS and ANTEA PDSS1 corpora. Results are expressed as an average over 20 iterations

Metric	Validation set	Test set	Phonetic groups	Type of features
Accuracy	0.88 ± 0.03	0.90 ± 0.07	labiodental fricative; dental, labial, and velar occlusives; affricate and sibilants alveolars; palatal sibilants	DFA, spectral mean, MFCC4, ΔMFCC1, ΔΔMFCC3, PLP3, PLP5, PLP11 ΔPLP3, ΔΔPLP1
Precision	0.93 ± 0.03	0.95 ± 0.05		
Recall	0.87 ± 0.03	0.88 ± 0.10		
F1-score	0.89 ± 0.03	0.91 ± 0.06		
Specificity	0.89 ± 0.04	0.93 ± 0.07		
AUC	0.94 ± 0.02	0.91 ± 0.06		

Discussion

Feature Effectiveness and Phonetic Groups Examination. Figure 4.8 emphasizes the potential of TRs between unvoiced consonants and adjacent sound segments.

Notably, 28 phonetic segments displayed significant correlations with the membership class, ranging from 0.52 to 0.85 (absolute values). The DFA coefficient from the transition between the occlusive sound /p/ and the vowel /e/ and the fifth MFCC from the TR between the sibilant sound /ʃ/ and the vowel /i/ demonstrated high effectiveness. When merging the ANTHEA-PDSS1 corpus, a slight performance reduction was observed. Nevertheless, several features remained significantly correlated with the class. Notably, the DFA derived from the transition between the occlusive consonant /t/ and the vowel /a/ exhibited the highest correlation with the class.

As for the specific types of features chosen, besides DFA, the analysis demonstrated the importance of MFCC2, MFCC3, MFCC5, intensity ratio, and spectral mean, which displayed a strong correlation with the class across several selected phonetic groups, particularly when considering signals recorded in optimal conditions. The inclusion of the second corpus lead to a reduction in the number of significant features. As evident from Figure 4.8, the RASTA-PLP coefficients are the most frequently selected parameters, thus suggesting an enhanced ability to discern differences between PDPs and HCs, even when dealing with recordings of sub-optimal quality.

Figure 4.9 displays the results of the analysis evaluating the effectiveness of different phonetic groups. Consistent with prior research [24], the most significant pronunciation impairments are associated with occlusive consonants, which may be due to the intricate articulatory movements necessary to produce such sounds. Indeed, unlike other consonants that do not involve a complete closure of the vocal tract, occlusive consonants are generated when airflow from the lungs encounters an obstruction resulting from a sudden change in the position of the articulatory organs. The complexity of executing precise and rapid movements is challenging for individuals with PDP when articulating these sounds.

As for the place of articulation, velar occlusive consonants appeared to present the highest discriminative power among native Italian speakers, possibly due to the required withdrawing of the tongue towards the soft palate. Upon merging the two corpora, features related to occlusive consonants remains frequently selected. However, an important role is also played by parameters associated with fricative labiodental sounds, which, in the current database, are represented by the syllable /fa/.

Classification. As evident from Table 4.18, the SVM classifier yielded the highest performance, achieving an accuracy of 98%. Notably, this classifier not only provided the highest accuracy but also displayed the lowest standard deviation, indicating a high level of model stability.

Following the hyperparameter optimization, the model was tested on an independent test set. As can be derived from Table 4.19, the model performance remains consistent when transitioning from the validation data to entirely new samples. This suggests that overfitting is absent, and the selected model exhibits strong generalization capabilities. As for the phonetic groups, the results align with those reported in Figure 4.8 on the single phonemes and underscore the significance of occlusive consonants and palatal sibilants.

Although the lack of entirely comparable work in the state of art, the comparison with the most similar study employing TRs [88] revealed the high potential of the proposed algorithm. Indeed, in the mentioned work the authors achieved $94\% \pm 1$ accuracy (AUC = 0.99, Sens = 0.9, Spec = 1) in a 11-fold CV and $82\% \pm 13$ (AUC = 0.95, Sens = 1, Spec = 0.57) in the cross corpora experiments employing a GMM-UBM classifier, PLP as features and the DDK speech task. Moreover, the accuracy reported when addressing the same task considered in this work (i.e. text dependent utterance) is $89\% \pm 7$ (AUC = 0.93, Sens = 0.91, Spec = 0.91) in a 11-fold CV.

Regarding phonetic groups, the results emphasize occlusive and sibilant consonants. While many features overlap between Experiments 1 and 2, the latter includes a broader feature set, adapting well to unsupervised recordings. Also in this second case, the performance remains stable when moving from validation to the test set, although the standard deviation slightly increases in the latter case. These findings indicate that the inclusion of non-supervised recordings leads to good performance. However, the models yielding the best performance also exhibited substantial differences both in terms of the selected hyperparameters and most effective features. This indicates that, although the classification task can be performed regardless of recording conditions, it is imperative to develope tailored models capable of handling the existing differences between acquisition modalities.

4.5.3 Investigation of Time Evolution of Speech Attractors

Both normophonic and non-normophonic speakers often experience non-linear phenomena during voice production, primarily resulting from factors like pressure flow in the glottis, stress-strain properties of vocal fold tissues, and vocal fold collision (Section 3.2.6). These are further complicated by compensatory movements often observed in PDP, which are intended to mitigate their motor dysfunctions.

Building on these assumptions, several studies in the related literature have employed non-linear metrics to quantify these impairments (Section 4.2). However, none has specifically delved into the temporal evolution of vocal trajectories within the reconstructed phase space as a method for identifying PD hallmarks. This section introduces a model based on three-dimensional geometry and its time-dependent changes with the aim of extracting information related to the speaker's health status.

As in the previous study (Section 4.5.2), the analysis is conducted on Italian native speakers, with the goal of augmenting the body of evidence related to this specific language. Furthermore, a dataset including both early-drug naive individuals (referred to as *de-novo*) and those in mid-advanced stages of PD is employed to explore significant correlations between the extracted features and the stage of the patient's disease. The results from this study are published in [239].

Materials

A subset of the Suppa corpus including 100 PD patients (54 mid-advanced and 46 de-novo) and 113 age- and gender-matched HCs was employed for this study.

Python libraries were used for data analysis. More specifically, the Topological Signal Processing library (Teaspoon) was employed to determine the most suitable parameters for the voice embedding procedure. Additionally, the αshape and Trimesh libraries were introduced for the calculation and parameterization of α-geometries derived from the reconstructed vocal attractors.

Methods

Embedding Approach. As previously described in Section 3.2.6, nonlinear aspects within vocal samples can be investigated through a representation of vocal signals

in state space. In this study, considerable attention was dedicated to preserving the ability to visualize the reconstructed signal, hence the embedding dimension, denoted as θ, was empirically set to 3. This choice allowed to maintain a manageable level of system complexity while exploring the nature and quality of the information extracted from a 3-dimensional representation of the vocal signal. It is worth noting that, in order to minimize the impact of noise on the reconstructed trajectories, a 50-sample moving average was systematically applied to each point within the attractor set.

Subsequently, the reconstructed ensemble of trajectories was characterized through the utilization of α-shapes corresponding to each set. Indeed, these representations have found versatile applications across diverse domains, including the approximation of bounding polytopes around a set of data points [240, 241]. According to the initial definition introduced by Edelsbrunner et al. in [32], the α-shape of a given set of points, represented as S, can be conceptualized as a graph composed of linear segments. Within this graph, the vertices correspond to the α-extreme points, while the edges connect these vertices to their respective α-neighbors. An α-extreme point within the set S is distinguished by the existence of a closed disk having a radius of $1/\alpha$, which encompasses all points within S. Similarly, two α-extreme points are regarded as α-neighbors if there exists a closed disk with a radius of $1/\alpha$ that includes both points on its boundary while encompassing all other points within S.

Building upon these principles, an α-shape can be constructed in such a way that its boundaries include all the points within the reconstructed attractors. This approach effectively defines the smallest volume within the phase space that contains the set of trajectories. It is noteworthy that the geometric representation of α-shapes typically consists of an assembly of triangles. The overall morphology of these shapes can described by the collection of their edges and vertices. This representation, involving triangular meshes, can facilitate the examination of vocal signals within the phase space by employing efficient and lightweight algorithms. In the context of this study, the α-shape solid for each of the reconstructed attractors was generated by empirically setting the value of α to 30.

Feature Extraction. As a consequence of the underlying computational attractors theory, it becomes evident that in presence of more regular voice production systems,

there is a greater tendency for trajectories to overlap and converge towardss a predetermined pattern. Consequently, the volume occupied by the points within the reconstructed phase space retains pivotal information concerning the original signal [24, 136]. Furthermore, to further focus on the temporal evolution, measurements of volume in adjacent vocal frames were incorporated into the present analysis. After removing initial and final periods of silence, two 1-second frames were selected from the original signal. The first frame, referred to as $Volume_{01}$, encompasses the initial transient regions, while the latter captures a more stable phase labeled as $Volume_{12}$. To ensure the robustness of the analysis and avoid potential bias from the final decay of phonation, only recordings longer than 3 seconds were retained.

It is noteworthy that the first window encompassed the attack phase of the voice signal, a period that often exhibits more chaotic behavior even in normo-phonic speakers. Subsequently, as the initial transient phase ends, the signal typically shows more predictable behavior, and the reconstructed trajectories evolve towards a predetermined pattern. In order to capture these subtle changes and explore their correlation with PD-related alterations, two metrics were employed: the volume variation between adjacent windows (ΔVolume) and the distance between two consecutive α-shape triangular meshes. This latter metric contains information regarding changes in the overall geometry. This distance was computed by aligning the two consecutive triangular meshes using the principal axes of inertia as an initial reference and subsequently measuring the average square distance per point included on the surface of the 3-D object.

Finally, a binary feature, denoted as *Watertight* (WT), was introduced to describe whether the 3-D geometry is represented by a closed surface devoid of holes. Indeed, such holes may be associated with more chaotic structures or lower recurrence periods, with attractor points accumulating at the center.

Feature Analysis. The effectiveness of the employed features underwent evaluation in two sequential steps. Initially, with regard to volume and distance measurements, feature distributions and trends were examined using violin plots. Subsequently, a Kruskal-Wallis statistical test was employed to determine whether the features could differentiate between the various classes. For a more deep understanding of the physical significance of each feature, additional statistical tests were conducted between paired groups: (i) comparison between HCs and PDPs (both in

early and mid-advanced stages) to assess the effectiveness of the approach in modeling vocal alterations; (ii) comparison between HC and early-stage PDPs to assess the model capability to identify early markers for neurodegeneration; (iii) comparison between early-stage and mid-advanced PD patients to evaluate the capability to differentiate between different stages of the disease.

As for WT, given its categorical and binary nature, the number of occurrences of watertight solids for each window and each class was evaluated. Thereafter, the cardinalities for each class were compared to investigate the presence of recurring 3-D geometries. It is worth noting that prior to conducting the feature importance analysis, an outlier removal step was performed, retaining only those instances falling within the 20^{th} and 80^{th} percentiles of the data distribution.

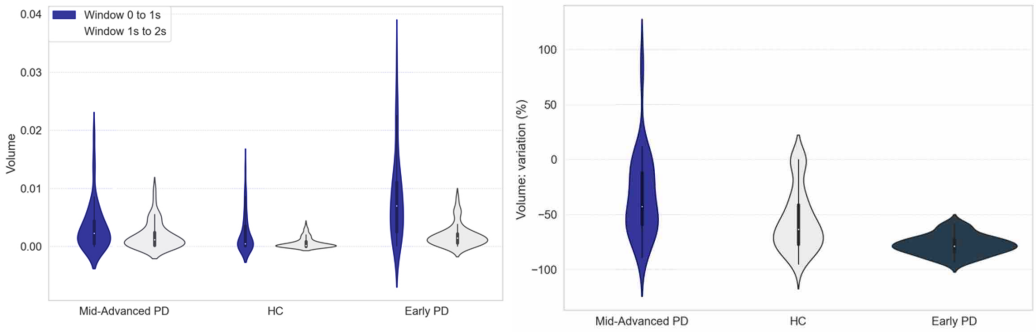
Results

In Figure 4.10, the violin plots illustrate the features examined in this study. Figure 4.10a specifically presents the distinction between $Volume_{01}$ and $Volume_{12}$, with the comparison presented for each class. Figure 4.10b illustrates the percentage variation in α-Shape volumes, while Figure 4.10c reports the distances between α-Shape geometries in the first and the second window. As for the count of α-Shape geometries presenting a watertight structure, the results indicated 59%, 55%, and 88% watertights solids for Mid-Advanced PD, HC, and Early PD, respectively. To maintain conciseness, results are reported solely for the second investigated window, as no significant variations related to this feature were noted when changing the analyzed time window.

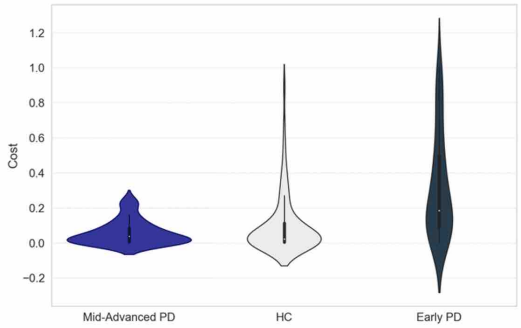
In Table 4.22 the results of the Kruskal Wallis test conducted to assess the presence of statistically significant differences among HCs, early-stage PDs, and mid-advanced PDs are reported.

Discussion

The results of this study confirmed the effectiveness of an approach based on reconstructed vocal attractor analysis to investigate vocal alterations related to PD. As illustrated in Figure 4.10, the analysis of feature distributions highlights significant increases in attractor volumes among PD patients. This finding aligns with



(a) $Volume_{0-1} and Volume_{1-2}$ (b) ΔVolume



(c) Distance between α-Shapes in the two temporal windows analyzed

Fig. 4.10 Distribution of features derived from the reconstructed attractors

previous research and emphasizes that a disrupted voice production process leads to the formation of highly chaotic attractors, which do not tend to converge towardss predictable patterns.

Moreover, although preliminary, the evidence showed a correlation between altered temporal evolution of attractors and PD arising. It is important to note that this phenomenon appears to impact individuals in the early stages of the disease differently of those in the mid-advanced stages. Indeed, in the case of the former group, higher volumes were observed in the first window, followed by a significant reduction in the adjacent segment. In contrast, mid-advanced patients presented lower volumes in the initial windows that remained relatively stable throughout the entire signal. As for HC, the analyses conducted revealed, as expected, a minimal volume variation, probably due to the rapid accumulation of points along the same trajectories, with no significant alterations in the overall 3D geometry. Intriguingly, early-stage patients took an intermediate stance between HC and Advanced PD, displaying a higher capability to rapidly overcome the effects of the attack phase,

Table 4.22 Statistical results obtained from the Kruskal Wallis test

Kruskal-Wallis test results	Feature name	p-value
HC vs Early PD vs Mid-Advanced PD	$Volume_{01}$	<0.001
	$Volume_{12}$	<0.001
	#Volume	<0.001
	Cost	<0.001
HC vs PD (Early PD and Mid-Advanced PD)	$Volume_{01}$	<0.001
	$Volume_{12}$	<0.001
	ΔVolume	0.63
	Cost	0.0018
HC vs Early PD	$Volume_{01}$	<0.001
	$Volume_{12}$	<0.001
	ΔVolume	<0.001
	Cost	<0.001
HC vs Mid-Advanced PD	$Volume_{01}$	0.064
	$Volume_{12}$	0.0012
	ΔVolume	<0.001
	Cost	0.86
Early PD vs Mid-Advanced PD	$Volume_{01}$	<0.001
	$Volume_{12}$	0.29
	ΔVolume	<0.001
	Cost	<0.001

much like HC. This dynamic contrasts with advanced patients, who exhibited a nearly stable, chaotic behavior without the capacity of rapid adjustment. In contrast, for individuals with PD, both subclasses displayed increased volumes in each analyzed time window. However, early-stage PD subjects tended to rapidly exhaust the effects of the attack phase, akin to HC subjects. This attack phase is not observed in advanced PD patients, who instead exhibited nearly stable chaotic behavior throughout the duration of the signal.

Regarding the 3-D geometry of the α-shape, a significant trend is evident among early PD patients in producing non-watertight solids. From a physical perspective, this outcome implies that the set of points within the attractors linked to the vocal of early drug-naive PD patients describes broader trajectories.

The results of the Kruskal-Wallis test, as summarized in Table 4.22, underscored substantial distinctions among HCs, early-stage PD, and mid-advanced PD groups for both punctual and differential volume characteristics, as well as the cost metric. Punctual volume metrics generally demonstrated high effectiveness, with particular efficacy observed when considering the initial attack phase for distinguishing between various disease stages. Conversely, differential volume metrics consistently exhibited statistically significant results ($p < 0.001$), especially in comparisons between HC and mid-advanced PD individuals. As for the cost metric, the statistical analysis revealed significant differences among the compared populations ($p < 0.001$), except in the case of HC and mid-advanced PD ($p = 0.86$).

4.6 Experimental Findings: Influence of External Factors

As evident from the analysis of the literature related to general vocal analysis (Section 4.2), as well in the specific field of PD (Section 4.5.1), the process of vocal parametrization requires to consider the complex mechanics underlying signal production. Indeed, this complexity introduces several speaker-specific factors (such as gender, age, or concurrent medical conditions) that affect the produced signal. Additionally, as previously discussed (Section 4.5.2), while simple phonation can be considered almost independent of the subject's language, the influence of the speaker's nationality becomes relevant when more complex analyses are conducted. These factors collectively lead to limited generalizability of results obtained from a specific dataset. Moreover, while the majority of studies tend to adopt a consistent data collection method, typically involving the placement of a microphone a few centimeters from the speaker's mouth in a quiet environment, it's crucial to acknowledge the potential substantial variability across cases due to additional external factors. This variability needs careful consideration to ensure the generalizability of results.

In this context, the experiments included in this section aim to evaluate the influence of external factors on the extracted acoustic parameters to validate previously obtained results and provide evidence for future studies. Specifically, the conducted analyses have primarily focused on assessing the robustness of acoustic features with respect to the subjects' demographic characteristics and the general recording

conditions. In addition, preliminary analyses aimed at evaluating the influence of medication (L-dopa) on the produced vocal signal have also been conducted. The results from this section are published in [120]

4.6.1 Assessment of Acoustic Features Robustness

This section provides a detailed description of the experiments undertaken to assess the impact of external factors, such as gender, language, and recording modality, on acoustic features, and to evaluate their reliability in distinguishing between HC and PD samples.

The process was carried out in two distinct stages. First, a statistical analysis was executed to identify features that exhibited consistent behavior across diverse datasets. These datasets included participants from various nationalities, each with distinct characteristics, and were recorded using a diverse range of equipment. Subsequently, a separate investigation was conducted to examine the feasibility of binary classification between HC and PD using a heterogeneous dataset.

Materials

Five diverse corpora, encompassing a total of 279 subjects (139 with PD and 140 HCs) from three different nationalities, were utilized in this study. These corpora included the IPVS (Section 4.3.1), PC-GITA (Section 4.3.3), ANTHEA-PDSS1 and ANTHEA-PDSS2 (Section 4.3.2), and the Hlavnicka corpus (Section 4.3.4). To ensure the model robustness and control for potential sources of complexity, vocal samples associated to the sustained phonation of the vowel /a/ were selected from all the included corpora.

Data analysis was carried out in Python employing Praat for pre-processing and feature extraction. The Parselmouth library served as an interface to access Praat internal code.

Methods

Pre-processing. To maintain data consistency across the corpora, which originally had different sampling rates, a down-sampling process was first applied to standardize

them to 16 kHz. Moreover, signal amplitudes were normalized within the [0, 1] range to reduce the potential impact of variations in speaker-microphone distances on subsequent analyses. To further improve data quality, initial and final silence regions were manually removed, eliminating the necessity for additional preparatory procedures.

Feature Extraction. A comprehensive set of vocal features was extracted from each vocal sample, encompassing both periodicity measures (such as F0, the first three formants, and their respective bandwidths) and noise-related measures (including HNR, CPP, and GNE). Furthermore, spectral characteristics, comprising features like flux, skewness, entropy, crest, flatness, slope, roll-off, spread, centroid, and kurtosis, were computed. Cepstral features consisted of MFCC from 1 to 13, including their first and second derivatives. Intensity, DFA, STE, PLP from 1 to 13, along with their derivatives, completed the feature set. Thereafter, for each feature, five essential statistics were calculated, which included the mean, median, standard deviation, kurtosis, and skewness. It is worth noting that jitter and shimmer variants were computed across the entire signal since their definitions inherently involve comparisons among contiguous frames. To ensure feature consistency, a min-max normalization procedure was applied to standardize them within a uniform range.

Statistical Analysis. The U Mann-Whitney test was initially utilized to identify features that exhibited a statistically significant distinction ($p < 0.05$) between individuals with PD and HC in a minimum of three datasets. Indeed, if this evidence occurs, it is possible to assume that the acoustic parameter presents robustness versus different datasets, the subject's characteristics, and the recording modalities. Thereafter, a unified dataset was constructed by merging all the corpora employed, and the same test was reiterated to investigate the impact of dataset heterogeneity on the identified features.

Subsequent to this, a Kruskal-Wallis test was carried out to ascertain the presence of statistically significant variations ($p < 0.05$) in the distribution of features attributed to external factors such as gender, language, and the modality of data collection. To mitigate potential bias arising from differences between the HC and PD groups, the test was independently applied to each subgroup. A feature was considered significantly different only if the null hypothesis was rejected in both populations.

Following the outcomes of these statistical analyses, a robust and effective subset of features underwent subsequent steps in feature selection and classification.

Feature Selection and Classification. After conducting feature extraction, the boruta algorithm was implemented on the unified dataset. The selected features were then fed into ten different classifiers, including KNN, SVM, GP, DT, RF, ANN, NB, LDA, ADA, and XGB, with the aim of performing an analysis minimally influenced by the model characteristics. To mitigate the potential risk of overfitting and ensure the robustness of the models, a two-phase methodology was adopted. In the initial phase, feature selection and model training were performed using 70% of the original dataset while the remaining 30% of subjects were exclusively employed for testing, with no further optimization or training applied to them. Moreover, a 10-fold CV technique was employed during the training phase to assess the generalization performance.

Results

Table 4.23 summarizes the statistical test results, focusing on features that showed both statistical significance and robustness against language, gender, and dataset variations. As for the binary classification, Table 4.24 presents the classification accuracy of the best models using two feature sets. The first set comprises features selected by the boruta algorithm in at least three corpora or in the unified dataset and minimally affected by external factors. The second feature set included external covariates (i.e., gender, language, data collection modality) integrated before feature selection, aiming to enhance generalization as noted in [87].

Table 4.23 Detailed results of the statistical analysis aimed to assess feature robustness

Present statistically significant differences between PD and HC in at least three dataset (p<0.05)	Present statistically significant differences between PD and HC in the unified dataset (p<0.05)	
	Robust to language, dataset type and gender (p>0.05)	Robust to language and dataset type (p>0.05)
Jitter (Lab); HNR (mean, std, skewness); STE (skewness, kurtosis); CPP (skewness); Spectral features: center of gravity (mean, std), roll-off (kurtosis, skewness), flux (kurtosis, skewness); $1^{std}Formant$ (std); $1^{std}Formant-bandwidth$ (std); MFCC: 2 (mean, median), 3 (mean); ΔMFCC: 11 (skewness); PLP: 2 (mean); Δ PLP: 1(mean, std); ΔΔ PLP: 1 (std), 2(std)	MFCC: 4 (mean, median); ΔMFCC: 13 (median); ΔΔMFCC; 7 (median), 13 (skewness, median).	Jitter (Lab); MFCC: 5 (skewness); $1^{std}Formant-bandwidth$ (median); PLP: 6 (median)

Table 4.24 Results of the classification step performed on the unified dataset

	10 fold CV	Test set				
	Accuracy	Accuracy	Specificity	Sensitivity	AUC	F1-score
XGB	71.8	64.9	70.2	60	65.13	64
	Jitter (Lab); Spectral center of gravity (mean); 1^{st} Formant (std); $\Delta\Delta$ MFCC: 7 (median); PLP: 6 (median)					
RF	70.2	70.1	75.8	65.9	70.8	71.6
	Spectral center of gravity (mean, std); 1^{st} Formant (std);					

Discussion

Statistical Analysis. Table 4.23 and Table 4.24 provide the results of the feature statistical analysis and subsequent classification procedure. According to the findings, MFCCs proved to be effective in distinguishing between individuals with PDP and HC, even when dealing with diverse datasets. Notably, they exhibited no statistically significant associations with language, gender, or dataset characteristics. Similarly, while the F0 itself lacks robustness against external factors, the associated Jitter features, which capture differential amplitude measures, seems to assume a pivotal role in discriminating between the two groups, being minimally affected by language and dataset variations.

Feature Selection and Classification. In the classification phase, XGB and RF demonstrated superior performance compared to the other classification models considered. Moreover, incorporating external factors like language and gender before the feature selection process, as previously noted in [87], led to improved generalization capabilities probably due to an overall model that better account for population-specific characteristics. Importantly, the transition from the validation to the test set did not result in a significant drop in performance.

The performance achieved confirm the feasibility of training a classification algorithm on a heterogeneous dataset. This evidence can be of crucial importance for future studies. Indeed, it is widely acknowledged that the size of the database is a primary challenge in developing automatic tools for assessing vocal pathology, as it can lead to overfitting of feature selection and classification outcomes to the specific population under investigation, as noted by Gomez-Vilda [242]. Moreover, although

highly homogeneous datasets may yield better results, replicating identical conditions can be challenging, limiting practical applicability in real-world scenarios. In this context, the conducted statistical tests, while not exhaustive, offer important insights into the impact of external factors on acoustic features, aiding in the identification of aspects necessitating stratification and those can be managed through algorithmic solutions (e.g., introducing covariates before feature selection).

Although there are no studies in the related literature encompassing an identical approach, the obtained classification results were compared with those reported in [96]. To the best of our knowledge, this represents the sole comparable study conducted to date. In their research, the authors achieved a 75% classification accuracy in a 10-fold CV performed on a heterogeneous dataset consisting of 241 PDPs and 265 HC individuals. However, additional analyses on a separate test set were not conducted.

4.6.2 Evaluation of Medication and Disease Progression Impact

Establishing consistent thresholds for categorizing PD stages in clinical contexts is a continuous challenge, often relying on partially exhaustive indicators such as UPDRS. Furthermore, the impact of medication intake on vocal samples remains poorly understood (as discussed in Section 4.5.1). Within this context, this section outlines the experiments conducted to explore the feasibility of automatically identifying various stages of PD progression and assessing medication status from vocal samples.

In this study, several binary classifications involving different disease stages and medication conditions were performed. Additionally, a dedicated post-hoc analysis of the acoustic features affected and the models achieving the best performance was carried out to uncover patterns and similarities among various tasks. It is worth mentioning that this study was conducted under the supervision of researchers affiliated with the University of Rome Tor Vergata, Rome, Italy, and is published in [243].

Materials

In this study a super-set of the Suppa corpus (Section 4.3.5) was employed, encompassing a total of 72 Early PDP (66.67% Male, 64.85 ± 8.36 years) and 88 Advanced

PDP (68.18% Male, 70.75 ± 8.87). Among these latter, 52 were recorded both in ON (i.e., within 1–2 h of the last administration.) and OFF (i.e., at least 12 h after the last medication intake) L-dopa status, whereas samples from the remaining subgroup were collected outside of the medication effect. Early PD subjects did not receive any medication due to their recent diagnosis. Additionally, 266 age- and gender-matched HCs were included as the normo-speaker counterpart.

Regarding the data collection procedure, vocal samples were recorded using either a Y6S Honor smartphone (manufactured by Huawei, Guangdong, China) or the professional equipment described in Section 4.3.5. Smartphone recordings were acquired through a dedicated application that ensured no compression or filtering and maintained the same sampling frequency as the professional microphones. For the purpose of this study, vocal samples associated to the sustained phonation of the vowel /e/ were used.

Signal processing, data analysis, and model training were conducted using Python 3.8, MATLAB R2022b (MathWorks, Natick, MA, USA), and Praat.

Methods

Pre-processing. Given the non-homogeneous data collection procedure performed through smartphones or professional microphones, a preliminary pre-processing was applied aiming at minimizing the differences between the two modalities.

Specifically, a noise reduction was applied through an algorithm based on spectral subtraction, which individually adapted to the noise profile of each audio recording. Moreover, to address the frequency response, a pre-emphasis process was employed to mimic the declared response of the Shure WH20 microphone, whereas the response of an omnidirectional MEMS microphone can be approximated as flat [244]. Additionally, a further low-pass filtering at 12 KHz was used, considering that smartphone responses decline in that frequency range, and the quantity of relevant information in voice signals is minimal. For this purpose, a 30-tap FIR filter implemented in MATLAB was utilized.

Feature Extraction. A total of 453 vocal features were selected for assessing voice disorders associated with PD. Among them, 339 features were derived applying the Voice Analysis Toolbox (Section 3.3), including acoustic parameters related to

F0, Jitter, Shimmer, HNR, MFCC, as well as various non-linear features like pitch period entropy and glottal-to-noise excitation. Additionally, 18 features related to low-frequency vocal tremor were extracted using a Praat script originally proposed in [245]. The final set of 96 features pertaining to vocal formants and their energy was extracted through Parselmouth coupled with custom routines. Five vocal formants were extracted from each vocal sample and subsequently the Teager-Kaiser energy operator (TKEO) was applied to estimate their instantaneous energy. From each formant and its energy, eight numerical parameters were eventually derived, including mean, standard deviation, range, percentile, and slope.

Feature Selection and Classification. In order to obtain robust results, minimally influenced by the characteristics of the pipeline employed, three distinct feature selection methods were implemented and compared, namely the Information Gain (IG), the Correlation Feature Selection (CFS), and the Minimum Redundancy Maximum Relevancy (mRMR). As for the latter, a variant of the classic algorithm, originally introduced by Tsanas et al. [246] and using the Spearman coefficient (mRMRS), was used

Within this task, three different classification models were employed, namely KNN, NB, and SVM. These algorithms were chosen for their simplicity and computational efficiency, enhancing the obtainment of interpretable and robust results in the presence of datasets with limited numerosity. Also, previous results from similar experiments and literature reviews generally demonstrated their satisfactory performance achieving resilience to overfitting (Section 4.4.1, Section 4.4.2, Section 4.5.1)

A 10-fold CV was then used to compare the performance using statistical metrics including accuracy, sensitivity, specificity, F1-score and AUC. It is important to highlight that the dataset utilized in the current experiment featured only one sample per subject. This inherently ensures speaker independence throughout the training-validation-test splitting procedures. A Bayesian optimization procedure aimed at minimizing the miss-classification error was eventually applied to the model to identify the best hyper parameters for each classifier.

Given the main objective of this study (i.e., exploring the effect of medication intake and disease progression on vocal samples), four different comparisons were

investigated: (i) HC vs Mid-Advanced PD, (ii) Early PD vs HC, (iii) Mid-Advanced PD vs Early PD, (iv) Mid-Advanced PD ON vs Mid-Advanced PD OFF.

Results

Table 4.25 displays the results derived from the Bayesian hyperparameter optimization process applied to the three distinct feature selection methods, namely CFS, IG, and mRMRS. For each binary classification, as well as each feature selection algorithm, the best combination of the number of features and classification model, resulting in the highest classification accuracy, is reported.

Table 4.25 Comparison across three feature selection algorithms employed. For each method, the number of features and the model that enhances the best performance is reported. The results are expressed in terms of 10-fold cross-validation accuracy

	Feature Selection	N Features	Model	Accuracy
Advanced PD	CFS	12	KNN	0.80±0.01
vs	IG	100	SVM	0.74±0.04
HC	mRMRS	50	SVM	0.77±0.01
Early PD	CFS	17	NB	0.82±0.01
vs	IG	30	SVM	0.78±0.16
HC	mRMRS	70	SVM	0.83±0.02
Advanced PD	CFS	17	KNN	0.85±0.02
vs	IG	30	NB	0.79±0.02
Early PD	mRMRS	10	NB	0.78±0.01
Advanced PD-ON	CFS	10	KNN	0.79±0.01
vs	IG	10	NB	0.66±0.03
Advanced PD-OFF	mRMRS	10	NB	0.69±0.02

To provide a more comprehensive evaluation of the efficacy of each feature selection algorithm, Table 4.26 presents the classification accuracy for each method averaged over the three classifiers. The internal hyperparameters were configured as outlined in Table 4.25.

In Table 4.27 the feature selected from the three algorithm employed for each binary and multiclass classification carried out in the current study are reported. For the sake of brevity, only the 5 top-ranked parameters for each feature selection method are included.

Table 4.26 Classification accuracy with respect to each feature selection algorithm employed

	CFS	IG	mRMRS
1. Adv. PD vs HC	0.78±0.09	0.73±0.04	0.75±0.05
2. Early PD vs HC	0.80±0.05	0.74±0.02	0.78±0.04
3. Adv. PD vs Early PD	0.84±0.01	0.75±0.02	0.75±0.02
4. Adv. PD-ON vs -OFF	0.72±0.05	0.56±0.1	0.63±0.06
Average	0.78±0.05	0.70±0.09	0.73±0.07

Table 4.27 Identification of the top five features resulting from the feature selection procedures

1. Advanced PD vs HC		
CFS	mRMRS	IG
std 8Δ Δ	std 8ΔΔ	std 10ΔΔ
std 11Δ	det TKEO mean 1	mean 5ΔΔ
std MFCC 1st	mean ΔΔ LogEn	std 8Δ Δ
VFER SNR TKEO	Shim F0 abs dif	std 8Δ
std MFCC 10th	GNE std	mean 6Δ
2. Early PD vs HC		
CFS	mRMRS	IG
std 4Δ	app En log 2	ATrPS
app LT En log 9	mean MFCC 4th	Ed2 1
IMF NSR En	det LT TKEO mean 3	app En log 6
FTrCIP	det En Sh. 1	app LT En Sh. 1
std 1ΔΔ	std MFCC 3rd	det LT En Sh. 1
3. Advanced PD vs Early PD		
CFS	mRMRS	IG
std 10Δ	std 10Δ	std 8Δ
std 10ΔΔ	mean 7ΔΔ	std 10Δ
std MFCC 10th	GNE std	std 10ΔΔ
GNE std	Shim F0 PQ3 Sch.	app LT TKEO mean 3
std MFCC 7th	F0 slopeLinFit	std 9Δ

Table 4.27 continued

4. Advanced PD-ON vs Advanced PD-OFF

CFS	mRMRS	IG
Jitt F0 PQ5 Baken	mean MFCC 6	app LT TKEO std 6
F1 TKEO mean	F0 slopeLinFit	AMoN
F5 rangePerc	F5 TKEO perc95	std 11Δ
det LT En Sh. 2	std 2Δ	F4 perc5
mean MFCC 6	F1 perc5	Jitt F0 PQ11

Discussion

Classification. The results derived from this study demonstrated the efficacy of shallow ML methods in discriminating between the vocal samples of individuals with PD and HC. Notably, this discrimination remains effective even in the early stages of the disease. Furthermore, the analysis reveals the ability to distinguish the voices of mid-advanced stage patients before and after therapy administration.

As for the feature selection algorithms, the choice of different methods have been demonstrated to substantially impact on classification accuracy, as highlighted in Table 4.25. This is particularly evident in the classification between Advanced PDPs in ON and OFF states, with a difference of almost 10% between the best and worst-performing algorithms. Moreover, when assessing the average performance across different classifiers (as presented in Table 4.26), the CFS method consistently outperforms IG, mRMRS ranking second.

Among the classification algorithms employed, KNN tends out to be the most effective option for the given tasks, despite its simplicity. Diving deeper into the performance achieved, the binary classification between Advanced PD and HCs consistently yielded optimal results with an average accuracy of 80% in a 10-fold CV setting. Similar performance was observed in binary classifications between different disease stages (e.g., Early vs. Advanced) or early diagnosis (Early vs. HC). Interestingly, the automatic evaluation of medication intake also delivered satisfactory results (accuracy 79%), indicating the feasibility of automatically recognizing medication status, albeit with increased complexity.

Feature Analysis. A post-hoc analysis of the feature selected by the different algorithms was performed to investigate which parameters demonstrate high effectiveness in both disease assessment and staging. Furthermore, considering the divergent findings within the existing literature that lack consensus regarding the impact of medication intake on vocal samples, special attention was dedicated to studying evidence from the comparison between Advanced PD OFF and Advanced PD ON. The objective was to validate and reinforce prior evidence on this specific aspect (Section 4.5.1).

The investigation performed revealed that the top-ranked features generated by CFS and mRMRS demonstrate considerable overlap, highlighting the robustness of the results. Notably, the features identified by CFS/mRMRS predominantly pertain to perceptual characteristics such as F0, shimmer, formants (e.g., F1, F5), and glottal model-based macroscopic indicators (e.g., VFER). This evidence confirms the relevance of pitch-related and prosodic features in the detection and staging of PD in voice. In contrast, IG frequently identifies distinct, less perceptually interpretable features.

Additionally, MFCCS and their derivatives were frequently selected, especially for mid-advanced PD patients, suggesting their utility in characterizing disease progression. Additionally, a significant number of features related to F0, shimmer, and jitter, commonly employed in voice analysis, were identified when comparing vocal samples from ON and OFF L-Dopa PD patients, suggesting a significant improvement in the periodicity of the produced signal after the medication intake, as suggested from earlier findings [99, 247].

4.6.3 Analysis of the Role of Recording Devices

The studies presented so far in this thesis have primarily focused on data recorded using professional microphones, occasionally introducing supplementary samples recorded under sub-optimal conditions (Sections 4.4.1, 4.4.2, 4.5.2, 4.6.1, 4.6.2). These suboptimal conditions may have involved lower-quality recording equipment, altered environmental noise levels, or a lack of supervision during the recording process, or a combination of these factors. However, given that this project forthcoming aim is to develop a light-weight and easy-to-use tool for remote monitoring,

where ideal conditions are often challenging to replicate, an additional analysis was conducted to specifically delve into the influence of recording conditions.

Within this context, the present study aims to study this aspect by conducting a detailed comparison among samples simultaneously recorded with two devices and evaluating the influence of recording modalities on the extracted acoustic parameters.

Materials

In this study, two different datasets encompassing recordings collected simultaneously with different devices were employed, namely the LUHS corpus (Section 4.3.6) and the ANTHEA-PDSS2 corpus (Section 4.3.2). Given the objectives of this study and to minimize additional sources of complexity arising from speakers' diverse nationalities, recordings of to sustained vowel phonation were exclusively employed.

Data analysis was carried out in Python; OpenSmile was used during the feature extraction procedure.

Methods

Pre-processing and Feature extraction. As described in Section 4.3.6, the LUHS corpus is distributed with pre-extracted features, hence no preprocessing or feature extraction procedures were applied to it.

As for the ANTHEA-PDSS2 corpus, a preliminary preprocessing step was implemented to remove initial and final silent regions. Afterward, the OpenSmile toolbox was employed to extract the CompaE2016 set of features. This step was taken to enable a direct comparison between the two corpora. Notably, the selected set of features is also part of the LUHS toolbox collection, representing a superset of many other included sets. By utilizing a predefined routine, it was possible to ensure that the extraction process remained consistent between the two corpora, thus avoiding discrepancies arising from the extraction procedure. Z-score standardization was then applied to reduce all the features to the same range.

Feature Selection and Classification. Given the primary objective of this study, which is to assess the impact of different recording modalities on the model ability

to differentiate between HCs and PDPs, a three-step pipeline was implemented for each subset of the LUHS dataset. This approach was adopted to mitigate potential biases stemming from variations in feature extraction toolboxes, feature extraction algorithms, and classification models.

In order to enable a fair comparison between the two recording modalities, each feature set in the LUHS dataset underwent three distinct feature selection methods (ANOVA, boruta, and correlation-Based), and the resulting subsets were input into four different classification models (SVM, KNN, GNB, GB). Moreover, an extra optimization step was introduced to fine-tune the hyperparameters associated with feature selection. This adjustment aimed to enhance the performance of the most effective feature selection-classifier combination.

The correlation-based approach is a customized algorithm, adapted from that presented in Section 4.4.1, which aims to select the most relevant features exhibiting a strong correlation with the class variable while ensuring non-redundancy (low cross-correlation). The internal algorithm involved calculating the Pearson coefficient (r) for each feature with respect to the class variable (r_{f0}) and considering its absolute value. Consequently, only features with the highest significance (i.e., $r > 0.3$, $p < 0.05$) were retained. Subsequently, intra-feature correlations (r_{ff}) were calculated, and for feature pairs where the inter-correlation exceeded the intra-correlation (i.e., $r_{ff} > r_{f0}$), the feature with lower correlation with the class variable was removed [129]. During the optimization process, the correlation threshold with the class variable was tuned from 0.3 to 0.4 with steps of 0.01. As for ANOVA, the initial value of k (i.e., the number of features returned by the model) was initially set to 10 and then fine-tuned from 5 to 50 with a step of 5. For boruta, the percentage of false positive values was tuned from 90 to 100 with steps of 1.

For each classifier, all accuracy values obtained through k-fold CV were recorded, and their mean value was calculated. The best-performing classifier, i.e., that with the highest mean value among the three considered feature selection methods, was selected based on the highest mean accuracy. Subsequently, the three most effective feature selection-classifier pairs were identified, and that with the highest validation accuracy was retained for the final optimization.

To ensure unbiased results, the original dataset was initially divided into an 80% portion for training and validation and a 20% portion for test set, without further optimization. Additionally, a 10-fold CV procedure was employed during the

pipeline selection and optimization, using the accuracy metric. Figure 4.11 provides a block diagram of the process applied to each feature set within the LUHS corpus, for both high- and low-quality recording equipment (i.e., 17 subsets x 2 recording modalities = 34 subsets).

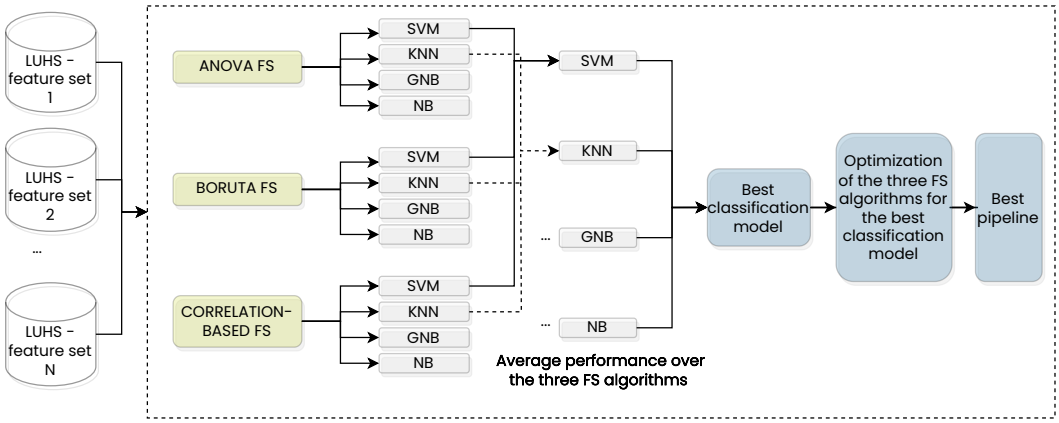


Fig. 4.11 Overview of the pipeline selection process applied to the LUHS dataset

Cross-device Validation. To delve further into the influence of recording conditions on classification performance, the best-performing pipeline from each subset of the LUHS corpus was subject to an additional cross-device validation.

In this analysis, the model trained on data recorded with the microphone was tested on data recorded with the smartphone (and vice versa). To ensure impartial results, the dataset was initially divided into training and test sets. Subject IDs were randomly assigned to either the training or test group, and the respective recordings were used either for model training or testing. An illustrative diagram of this process is provided in Figure 4.12.

Statistical Analysis. To investigate the impact of recording modality on the extracted features, the Wilcoxon signed-rank test was applied to each set of features within the LUHS dataset. This test provided a p-value for each feature, which was then sorted in ascending and descending order. Among them, the top 20 features that exhibited statistically significant differences ($p < 0.05$) between high-quality and low-quality acquisitions and the top 20 features that did not show statistically significant distinctions ($p > 0.05$) were selected.

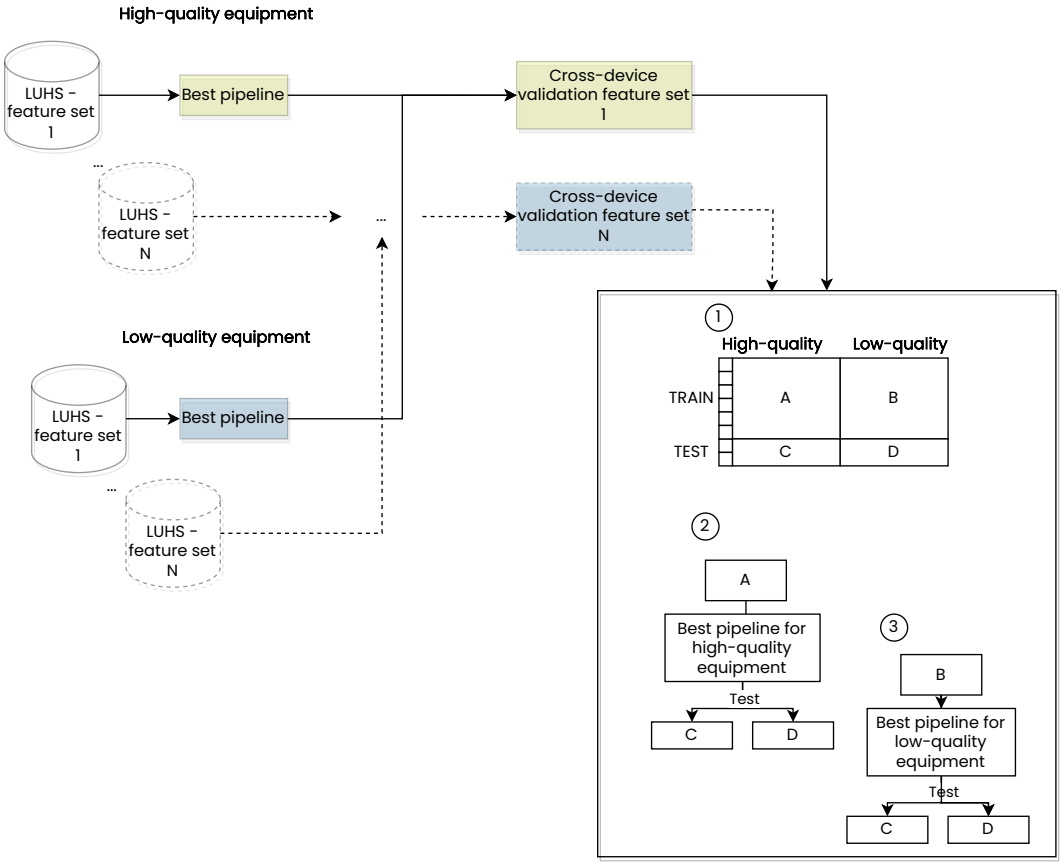


Fig. 4.12 Overview of the cross-validation process applied to the LUHS corpus

Subsequently, and in-depth analyses was conducted to identify which feature families demonstrated statistically similar values between the two recording modes thus proving invariant with respect to the recording technique. In more detail, after identifying the features with the maximum and minimum differences for each dataset within the LUHS corpus, a cross-toolbox analysis was carried out. This analysis assumed that, despite variations, the LUHS datasets evaluated similar sets of features. It is important to note that, due to discrepancies among various toolboxes, the analysis was simplified by comparing *feature families* that were created by grouping different features referring to the same aspect. For instance, diverse toolboxes compute distinct statistical parameters derived from the F0. Performing a one-to-one comparison of these features would be impractical. Nevertheless, grouping these statistical parameters into domains allows for a more feasible high-level study, enabling the assessment of similarities across different studies within a specific domain.

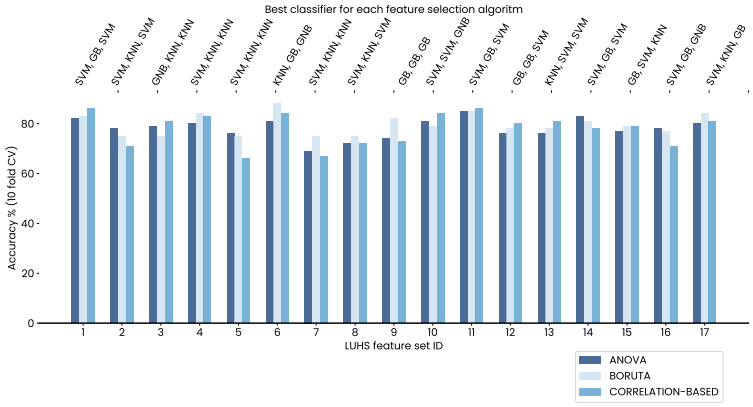
Within this approach, the term *F0 related* encompasses all statistical parameters associated to the fundamental frequency, while *Jitter* and *Shimmer* group together various Jitter and Shimmer variants (e.g., APQ5, AP3, and so on). In the case of different spectral statistics and cepstral parameters, including MFCCs and CPP, they are collectively referred to as *Spectral* and *Cepstral*, respectively. The *Noise* category includes all parameters related to noise characterization in voice, as discussed in Section 3.2.2, such as HNR and GNE, among others. *Energy in specific bands* encompasses features that are directly linked to energy in given spectral bands (as regions in the acoustic signal spectrum measures from those bands). *Probability of voiced* includes all features related to the alternation between voiced and unvoiced segments. *Loudness* encompasses all measures related to the sound intensity of the acoustic signal, while *Envelope Descriptors* includes all descriptors of the signal envelope. Finally, *Area of moments* groups together measures related to the distribution of energy in the acoustic signal, although only one toolbox employed this type of feature.

Despite its limitations, this analysis provided insight into how individual features are influenced by the recording technique, regardless of the specific algorithm used for their calculation. Indeed, different toolboxes often employ diverse algorithms to evaluate the same feature.

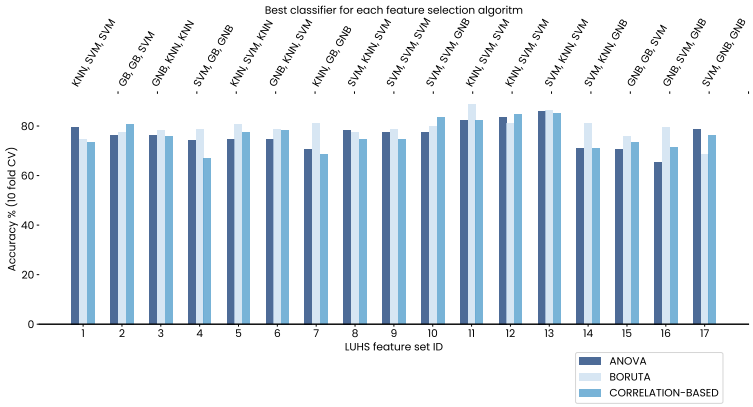
Building on the results obtained from the LUHS dataset, a second corpus was used to assess the generalizability of the findings on previously unseen samples recorded in a different setting. Indeed, since the LUHS corpus came with pre-extracted features that did not allow signal analysis and manipulation, a second dataset, namely the ANTHEA-PDSS2, was recorded simultaneously using both professional equipment and smartphones. All the samples in this second corpus underwent pre-processing and feature extraction employing the ComPare set from OpenSmile. From this set, the 20 features with the most significant differences in the LUHS corpus, denoted by low p-values, were selected and investigated. Subsequently, considering that the presence of background noise is one of the primary challenges in recordings made with omnidirectional microphones, typically found in smartphones, the *spectral subtraction* technique was applied to each signal. The features were then extracted once again, and their distributions were recalculated to investigate the impact of a targeted preprocessing technique aimed at mitigating one of the primary sources of difference between the two recording modes.

Results

Classification. In Figure 4.13 the results of the comparison between the classifiers and feature selection algorithms are reported for the High-quality and Low-quality sub-dataset included in the LUHS corpus. Performance are expressed in terms of 10-fold CV Accuracy. For the sake of brevity, only the best performing classifiers are displayed. In Table 4.28 are reported the performance of the optimized models for both high- and low-quality equipment.



(a) High-quality equipment



(b) Low-quality equipment

Fig. 4.13 Results of the comparison between classifiers and feature selection algorithms for sub-datasets in the LUHS corpus. Performance are expressed in terms of 10-fold CV Accuracy

Table 4.28 Performance details of the optimized models for high- and low-quality equipment. Acc: Accuracy

LUHS set ID	High-quality equipment				Low- quality equipment			
	Feature selection	Classifier	Val. Acc.	Test Acc.	Feature selection	Classifier	Val. Acc.	Test Acc.
1	Correlation (r:0.32)	SVM	0.86	0.85	ANOVA (k=35)	KNN	0.82	0.55
2	ANOVA (k=25)	SVM	0.84	0.8	Correlation (r:0.3)	SVM	0.81	0.7
3	Correlation (r:0.3)	SVM	0.81	0.75	boruta (p=97)	KNN	0.81	0.85
4	boruta (p=97)	KNN	0..85	0.6	boruta (p=96)	GB	0.8	0.8
5	ANOVA (k=25)	SVM	0.75	0.65	boruta (p=98)	SVM	0.85	0.7
6	boruta (p=99)	GB	0.84	0.80	boruta (p=97)	KNN	0.78	0.55
7	boruta (p=99)	KNN	0.75	0.55	boruta (p=96)	GB	0.84	0.7
8	boruta (p=93)	KNN	0.78	0.7	ANOVA (k=10)	SVM	0.77	0.5
9	boruta (p=99)	GB	0.84	0.65	boruta (p=93)	SVM	0.85	0.7
10	Correlation (r:0.31)	GNB	0.84	0.7	Correlation (r:0.3)	GNB	0.85	0.6
11	Correlation (r:0.3)	SVM	0.86	0.8	boruta (p=100)	SVM	0.89	0.80
12	Correlation (r:0.3)	SVM	0.8	0.75	Correlation (r:0.3)	SVM	0.85	0.85
13	Correlation (r:0.31)	SVM	0.82	0.6	boruta (p=98)	KNN	0.88	0.85
14	ANOVA (k=20)	SVM	0.84	0.65	boruta (p=100)	SVM	0.79	0.85
15	boruta (p=31)	KNN	0.8	0.8	boruta (p=98)	GB	0.8	0.8
16	ANOVA (k=20)	SVM	0.80	0.75	boruta (p=98)	SVM	0.82	0.55
17	boruta (p=99)	KNN	0.86	0.6	ANOVA (k=10)	SVM	0.82	0.6

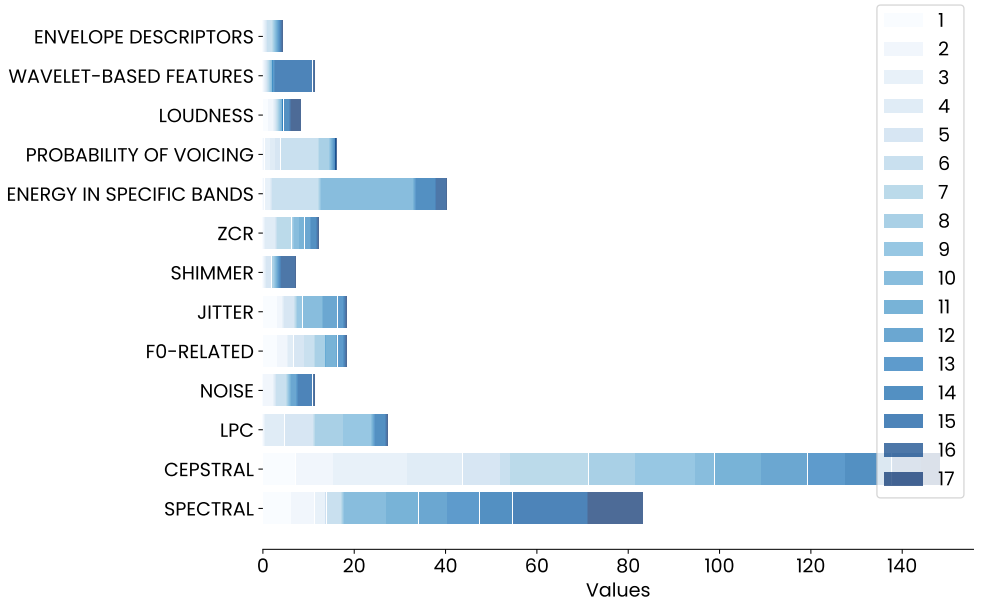
Cross-device Validation. In Table 4.29 the results from the cross-device experiment conducted are reported. To maintain conciseness, these experiments were only carried out for the LUHS subset that yielded the best performance. Specifically, the first set of experiments pertains to high-quality equipment, while the eleventh set pertains to low-quality equipment, as shown in Table 4.28.

Table 4.29 Results obtained from the cross-device experiment

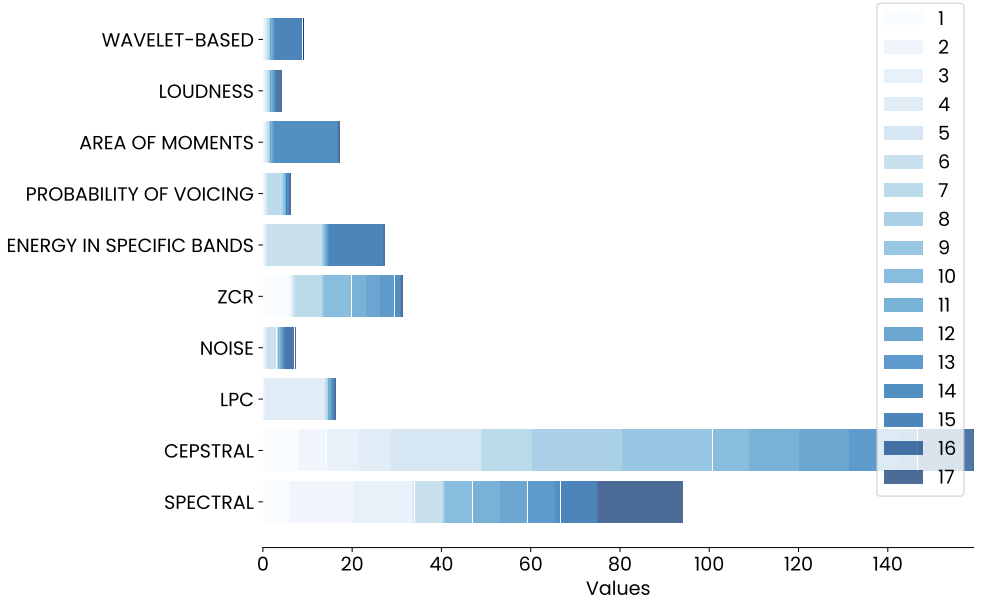
	High-quality equipment	Low-quality equipment
LUHS subset leading to highest performance	1	11
Classifier	SVM	SVM
Feature selection	Corr (r=0.32)	boruta (perc=100)
Train accuracy	0.86	0.96
Test accuracy	0.85	0.80
Cross-device test accuracy	0.75	0.70

Statistical Analysis. In Figure 4.14 the results of the Wilcoxon test are reported for those features yielding the highest and the smallest difference between the recording conditions. Furthermore, to conduct a more thorough investigation into the cepstral class that is frequently selected in both categories, an in-depth analysis was performed to determine whether differences also exist in the specific type of MFCC coefficients selected. According to the findings, the lower coefficients, particularly MFCC 1, is frequently associated to the most significant differences between the two recording modalities.

Results on the ANTHEA-PDSS2 Dataset. In Figure 4.16 the results of the validation on the ANTHEA-PDSS2 corpus are reported. In particular, the figure reports the distribution of the top 20 features exhibiting the most significant differences between recording modalities. Boxplots allow the comparison between samples recorded under high-quality conditions, low-quality conditions, and low-quality conditions following a denoising procedure. For each feature the corresponding OpenSmile notation is reported.

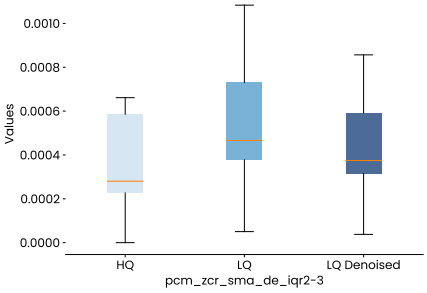


(a) High-quality equipment

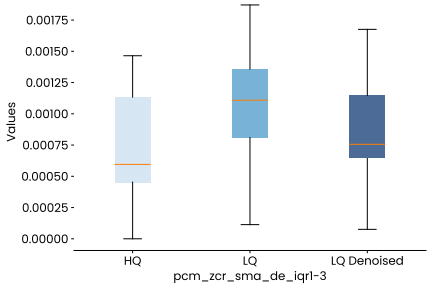


(b) Low-quality equipment

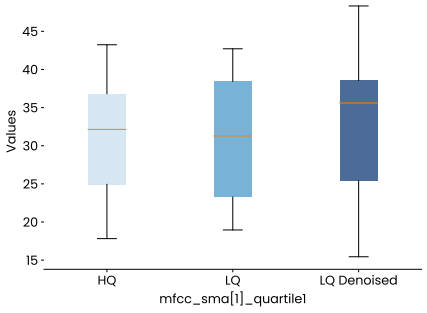
Fig. 4.14 Results of the Wilcoxon test for features showing the smallest differences between recording devices


0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
Values
HQ
LQ
LQ Denoised
pcm_zcr_sma_de_iqr2-3

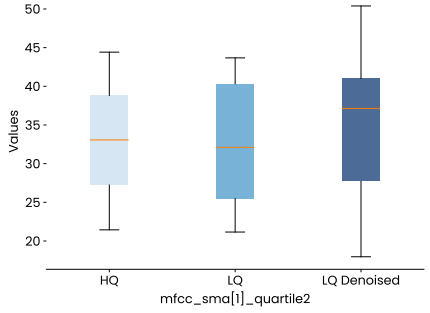
(a)


0.00175
0.00150
0.00125
0.00100
0.00075
0.00050
0.00025
0.00000
Values
HQ
LQ
LQ Denoised
pcm_zcr_sma_de_iqr1-3

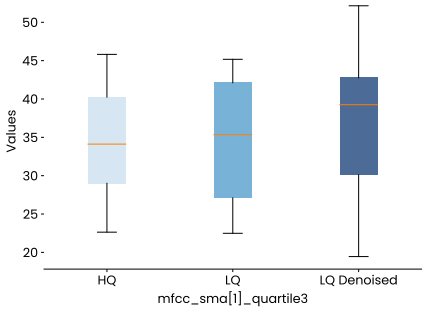
(b)


45
40
35
30
25
20
15
Values
HQ
LQ
LQ Denoised
mfcc_sma[1]_quartile1

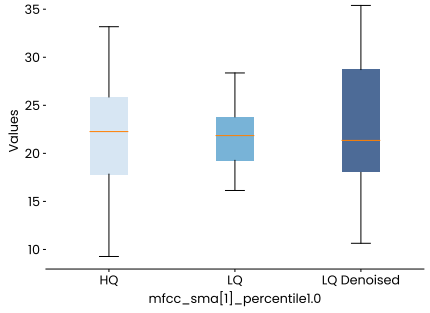
(c)


50
45
40
35
30
25
20
Values
HQ
LQ
LQ Denoised
mfcc_sma[1]_quartile2

(d)

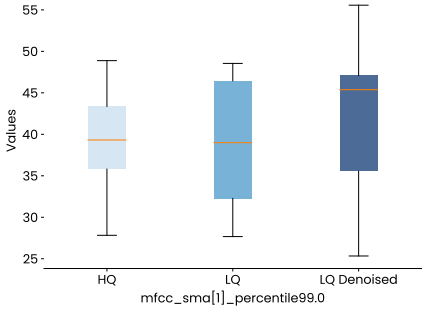

50
45
40
35
30
25
20
Values
HQ
LQ
LQ Denoised
mfcc_sma[1]_quartile3

(e)

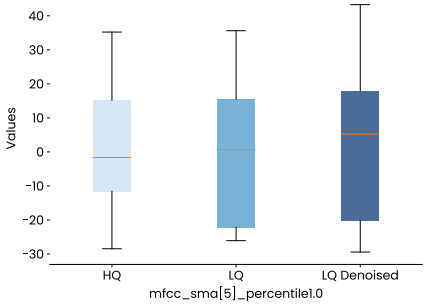

35
30
25
20
15
10
Values
HQ
LQ
LQ Denoised
mfcc_sma[1]_percentile1.0

(f)

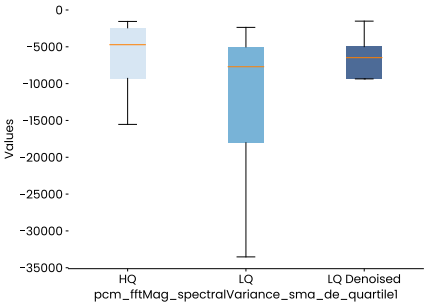
Fig. 4.15 Continue



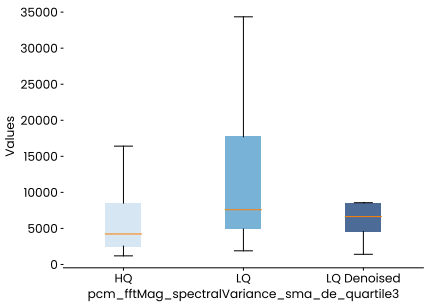
(g)



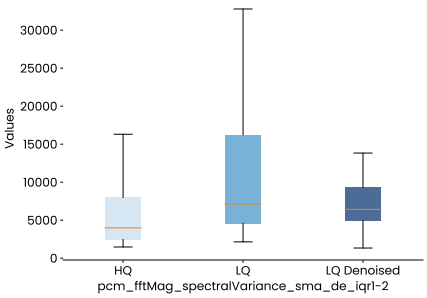
(h)



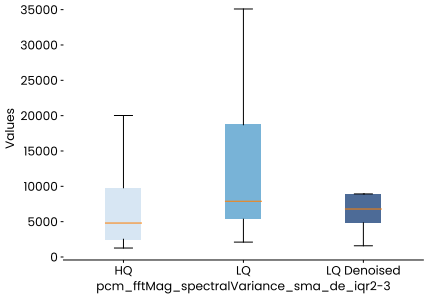
(i)



(j)

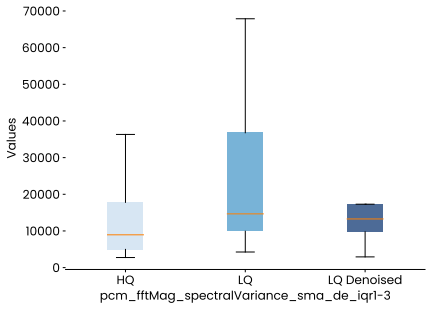


(k)

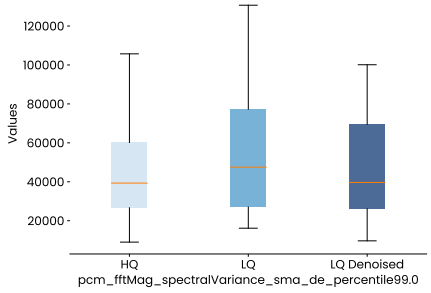


(l)

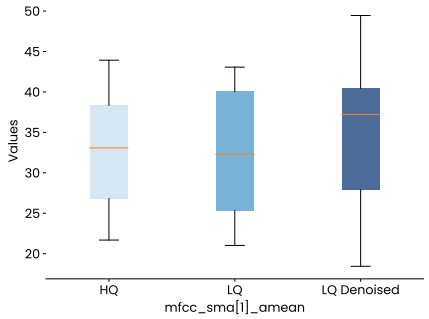
Fig. 4.15 Continue



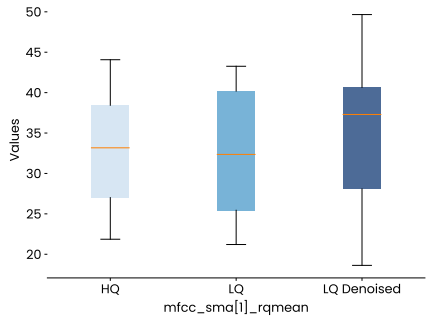
(m)



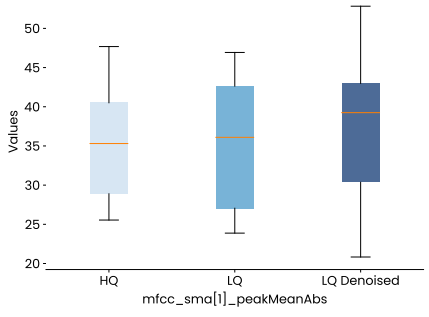
(n)



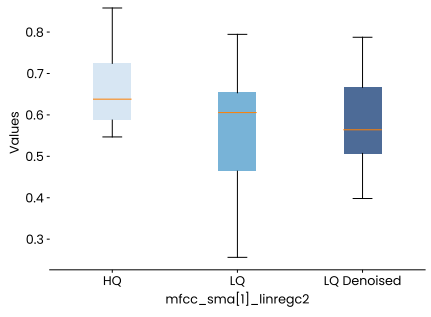
(o)



(p)

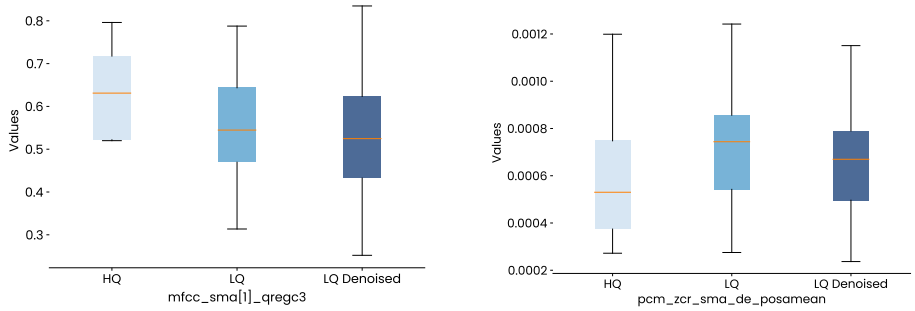


(q)



(r)

Fig. 4.15 Continue



(s) (t)

Fig. 4.16 Distribution of the top 20 features exhibiting the most significant differences between recording modalities. This representation illustrates the data for samples recorded under high-quality conditions (HQ), low-quality conditions (LQ), and low-quality conditions following a denoising procedure (LQ Denoised)

Discussion

Classification. The classification step, performed on the 17 different feature sets included in the LUHS corpus, demonstrated optimal performance for both high- and low-quality recordings with no significant degradation between the two modalities (Figures 4.13). Regarding the pipeline used, no consistent trend emerged from the comparison of results across different feature sets. In fact, all feature selection and classification algorithms proved effective for some feature sets, although SVM and KNN were more frequently selected as the best performing classifiers.

Following the optimization of hyperparameters for feature selection, the best-performing pipelines were tested on an independent set composed of previously unseen samples. The results, shown in Table 4.28, confirmed similar performance between the different recording techniques. Notably, boruta and Correlation-based feature selection algorithms were among the most effective ones . In terms of classification models, SVM and KNN consistently demonstrated their efficacy in distinguishing between HCs and PDPs, even when dealing with a limited dataset.

One noticeable difference emerged among various models, with some of them exhibiting a strong generalization capability, while others yielding variable performance between the validation and test sets. It is important to note that all sets

underwent the same model selection and optimization procedures, suggesting that these outcomes may be attributed to inherent characteristics of the original dataset. Furthermore, overlaying these findings with the comparison of recording modalities, it is possible to observe that only five sets (2, 3, 11, 12, 15) consistently demonstrated generalization capabilities regardless of the data acquisition equipment used, thus suggesting the robustness of the features included versus the data recording modality.

Cross-device Validation. The results of the cross-device validation support and reinforce the findings from previous analyses: it is feasible to classify between controls and patients with PD using voice signals, recorded under either optimal or sub-optimal conditions, as long as a specific pipeline is trained on similar signals that closely resemble those the model will be applied to.

As demonstrated in Table 4.29, even though the two training sets are composed of the same subjects, randomly selected and differing only in the data collection technique, transitioning from one recording modality to another yields approximately 10% reduction in classification accuracy. This reduction occurs whether the model is trained on higher-quality data and tested on lower-quality data or vice versa. Considering the limited size of the dataset, it is reasonable to assume that a substantial increase in sample size could reduce this effect.

Statistical Analysis. The comparison between features with the smallest differences and those with the highest differences (Figure 4.14) based on the data collection modality revealed that statistical parameters extracted from F0 and Jitter-related features are the most robust, regardless of the data collection modality. These features were exclusively included within the group with the highest p-values and were selected in 8 and 6 feature sets, respectively. Therefore, their robustness is likely independent of the algorithm used for their computation. As for Shimmer-related features, they also are included within the features with the lowest differences. However, there is less agreement among different datasets, indicating reduced robustness in the results. Envelope descriptors, Wavelet-based features, and area of moments were included only in a single feature set each, thus further validation of the results is necessary.

Regarding spectral and cepstral features, a deeper analysis was deemed necessary due to the large number of samples in both identified classes. The results indicated

a remarkable difference in the occurrences of the statistical descriptors Variance and Flatness. Variance exhibits a high dependence on the acquisition mode, while Flatness exhibits low dependence. Similarly, an in-depth analysis of various MFCC coefficients revealed that lower coefficients, particularly the first one, are more prominently selected among the features that exhibit a higher dependency on the classes. This phenomenon may be attributed to the increased presence of background noise captured in recordings made with omnidirectional microphones, as is the case with smartphones.

Results on the ANTHEA-PDSS2 Dataset. A further experiment was conducted to assess the influence of applying a denoising filter on the extracted features. Indeed, the results of binary classification and cross-device validation indicate that vocal signals, whether acquired with professional or low-cost equipment, contain crucial information about the patient's health status. Furthermore the proposed automatic classification models exhibited similar performance in both cases, suggesting that the key constraint is related to parameter extraction and the development of models tailored to the data collection mode. To validate this evidence without direct access to the raw audio signals in the LUHS corpus, it was decided to use a second dataset containing samples collected in similar conditions, to evaluate the influence of some pre-processing steps aimed to reduce the difference between the recording modalities.

With this in mind, the 20 features that exhibited the highest differences between the two recording modalities were computed from the new set of samples recorded with professional microphones and smartphones. Additionally, features were also extracted from smartphone-recorded samples after applying a pre-processing step involving a denoising filter based on spectral subtraction. The results, as shown in Figure 4.16, revealed that the application of the denoising step effectively reduced the differences between the data collection modalities for Zero Crossing Rate (ZCR) and Spectral features. However, the preprocessing step did not yield significant improvement for more complex parameters such as MFCC.

4.7 Overall Conclusions and Future Works

In this dissertation, several experiments were conducted to explore various facets of voice analysis in the context of PD. The primary objectives of these experiments were

twofold: first, to demonstrate the effectiveness of automatic models in recognizing vocal impairments associated with PD; second, to explore specific aspects of the data collection and analysis procedures related to this study. The findings revealed several key insights.

It has been shown that different vocal tasks exhibited variable performance, and as expected, an effective and concise protocol for distinguishing between control subjects and individuals with PD should encompass the use of a vowel, with the vowel /a/ yielding superior results, along with the inclusion of a sentence. Interestingly, when comparing various phonetically balanced sentences, no substantial differences was revealed. Nevertheless, the use of occlusive sounds proved particularly effective. In the context of the Italian language, occlusives and fricatives demonstrated a superior capacity to capture the characteristic impairment associated to PD. In scenarios involving multiple tasks, an early fusion of parameters within the feature space should be preferred over a majority voting approach (late fusion). This evidence was supported by improved performance, generalization capacity, and computational efficiency.

Regarding acoustic parameters, MFCC coefficients seem to be the preferred choice as they are included in most available toolboxes, have proven highly effective in numerous studies, appear robust even under suboptimal recording conditions, and can capture vocal alterations even in heterogeneous datasets with varying subject demographics (e.g., language and gender). In addition to MFCC, commonly used acoustic parameters include F0, Shimmer, Jitter, descriptive spectrum parameters (e.g., center of gravity, flatness), HNR, PLP, DFA, and RPDE. Among these, F0, Jitter, and Shimmer, as expected, exhibit a strong gender dependency that should be considered in analysis. In general, it is a good practice to introduce cofactors such as gender and age before feature selection to assist algorithms in selecting the right feature subset, resulting in improved system performance. When it comes to complexity parameters, the analyses conducted and the literature review demonstrated their effectiveness, even though it is advisable to use them in conjunction with other features. Parameters such as F0, Shimmer, and formants are also effective in capturing vocal alterations following medication intake. These alterations primarily manifest as changes in the periodic aspect of the signal.

Additionally, the inclusion of specific parameters designed to assess fine motor impairment through transition zone analysis showcased significant potential in

distinguishing between HCs and individuals with PD. Another promising approach involved the analysis of the temporal evolution of vocal signals reconstructed in phase space, offering insights into differences between control subjects and those with PD, as well as distinctions among patients in the early and advanced stages of the disease.

In terms of the toolboxes employed, Praat emerged as the preferred choice, primarily due to its widespread usage, which enables direct comparisons with other studies. Praat also provided validated acoustic parameters, supported study reproducibility, included routines for analyzing various tasks, and could be seamlessly integrated into Python using the Parselmouth library. However, it is essential to note that the included parameter set was not exhaustive, necessitating to be complemented with routines from other libraries.

An experiment conducted on a heterogeneous dataset, albeit limited to the use of the vowel /a/ , successfully demonstrated strong classification ability even when data were recorded with different devices and included subjects with diverse demographic characteristics. This finding holds promise, especially in the context of biomedical signals, where data availability may be limited, potentially facilitating the development of cross-lingual models.

The evidence of analyses conducted on the data collection modality, collectively suggests that it is possible to classify individuals with PD from HCs using recordings made in unsupervised environments with low-quality microphones, as those embedded in smartphones. Nevertheless, it was observed that recording in unsupervised settings significantly impacted performance. In the case of recordings made with smartphones in supervised environments, existing literature and conducted experiments converged on the evidence that classification was attainable if a tailored pipeline was constructed, utilizing features and algorithms suitable for recordings different from those captured in ideal conditions. It was emphasized that, in general, training models with data that reflected the intended use conditions was the best practice to be applied. However, in situations where this was not feasible, the application of pre-processing techniques aimed at minimizing differences between recording modalities could lead to improvements in certain parameters, while others, such as MFCC, may exhibit no improvement or even impair, warranting careful attention.

Despite the promising results reported in this dissertation, one major limitation was the frequent reliance on small sample sizes. Future research endeavors should

prioritize the expansion of sample sizes to ensure more robust analyses. Furthermore, future investigations should explore the simultaneous influence of multiple pathologies on vocal signals, examine the feasibility of personalized patient rehabilitation based on preliminary vocal analysis to identify the most impactful aspects, extend the analysis of medication effects to develop support for physicians in administering and dosing therapy, and assess the viability of a differential analysis.

While certain challenges and limitations were identified, the promising results suggest that with the expansion of datasets, refinement of techniques, and sustained research efforts, voice analysis can be a valuable asset in the management of this complex neurological disorders.

Chapter 5

Application II: GERD and Obesity

5.1 Obesity

5.1.1 Incidence and Prevalence

Obesity is a complex and multifactorial chronic condition characterized by the excessive accumulation of adipose tissue [248]. Over the last few decades, there has been a consistent global increase in the incidence and prevalence of obesity, with estimates indicating that it has more than doubled since 1980 [249]. This rise can be attributed to various factors, including shifts in dietary habits, sedentary lifestyles, and genetic predisposition.

According to the World Health Organization (WHO), as of 2022, the global impact of obesity is significant, affecting over 640 million adults and 110 million children and adolescents worldwide [248]. However, it is important to note that the incidence and prevalence of obesity vary by region and country. High-income nations, such as the United States, tend to report higher obesity rates compared to low-income countries. Recent WHO studies have highlighted that North America, Europe, and Oceania have a higher prevalence of obesity [248].

Furthermore, disparities in educational attainment have proven to be a determining factor, with higher rates of obesity observed among individuals with lower levels of education [248].

5.1.2 Pathophysiology

At cellular level, obesity is characterized by the hypertrophy and hyperplasia of adipocytes. Increased caloric intake results in an accumulation of excess triglycerides within adipocytes, setting off a chain of metabolic alterations and changes in the secretion of adipokines—bioactive substances released by adipose tissue.

Going into more detail, the dysfunction of adipose tissue in obesity primarily entails chronic low-grade inflammation and disruptions in adipokine production. Adipokines play a pivotal role in the regulation of appetite, energy metabolism, insulin sensitivity, and inflammation. Dysregulation in the production and release of adipokines contributes to metabolic irregularities and low-grade systemic inflammation [250]. Consequently, adipocytes release elevated levels of pro-inflammatory cytokines, resulting in tissue inflammation and insulin resistance, often prevalent in individuals with obesity [250]. Furthermore, the malfunctioning of adipose tissue disrupts the balance of other hormones involved in appetite regulation, leading to an altered perception of hunger and satiety. This disruption in appetite control contributes to excessive calorie consumption and further weight gain.

Due to this series of underlying pathological changes, obesity extends its impact beyond adipose tissue and affects multiple organs and systems. It must be regarded as a risk factor for various diseases and conditions including among others, Type 2 diabetes mellitus, cardiovascular diseases, metabolic syndrome, GERD, chronic kidney disease, hypertension, cancer, obstructive sleep apnea, and depression [249, 248]

5.1.3 Etiology

The pathogenesis of obesity is a complex interplay of genetic, environmental, and behavioral factors leading to an imbalance between energy intake and expenditure.

As for *Genetic factors*, recent evidence suggests that the heritability of Body Mass Index (BMI) ranges from 40% to 70%, underscoring the pivotal role of genetic factors in the development of obesity [250]. Genomic studies have identified deficiencies in leptin and melanocortin-4 receptors as the most common genetic causes of obesity. These genes are predominantly expressed in the hypothalamus and play a crucial role in the neural circuits that regulate energy homeostasis [250].

Table 5.1 Classifications of adults based on Body Mass Index

Classification	BMI	Risk of comorbities
Underweight	<18.5	Increased risk of other clinical problems)
Normal range	18.50 - 24.99	Average
Overweight	$\geq$ 25.00	
Preobese	25.00-29.99	Increased
Obese class I	30.00-34.99	Moderate
Obese class II	35.00-39.99	Severe
Obese class III	$\geq$ 40.00	Very severe

Moreover, *Environmental factors* demonstrated to play a pivotal role. Indeed, ongoing research indicates a multitude of factors that have contributed to a positive energy balance and weight gain, particularly in high-income countries [248]. These include increased availability and consumption of high-calorie foods, along with a decline in occupational physical activity. Furthermore, the substitution of leisure-time physical activities with sedentary pursuits such as television watching and computer games has further exacerbated this trend. Finally, the growing use of drugs that can cause weight gain as a collateral effect and inadequate sleep has also been demonstrated to have a strong influence on the increased obesity incidence [250].

5.1.4 Diagnostic Criteria and Complicating Factors

The diagnostic criteria for obesity primarily rely on the assessment of BMI and abdominal circumference. Despite the influence of sex, age, and race [251–253], this measurement provides objective indicators of adiposity and aids in classifying individuals into different weight categories. The classification of different subgroups, in accordance with the WHO [254], is shown in Table 5.1.

It is important to ackowledge that BMI functions as a screening tool and does not directly measure body fat percentage or consider variations in body composition. Therefore, it is subject to limitations, particularly in individuals with high muscle mass or variations in bone density. In these cases, clinical judgment and additional evaluations may be necessary to complement the BMI assessment and provide a comprehensive diagnosis [254].

5.1.5 Treatment

The treatment of obesity involves a comprehensive approach designed to achieve weight loss, improve overall health, and mitigate obesity-related complications. This approach encompasses lifestyle modifications, dietary interventions, increased physical activity, behavioral therapy, pharmacotherapy, and, in severe cases, bariatric surgery.

- *Lifestyle modification.* The primary method for managing weight is to reduce caloric balance through dietary adjustments and increased physical activity. Behavioral counseling, along with the use of applications delivered via smartphone or computer, is often employed to provide daily support to patients and address emotional and behavioral factors contributing to overeating and sedentary habits [250]. Additionally, collaborative counseling techniques are used to enhance adherence to lifestyle modifications and prevent weight regain.

- *Pharmacotheraphy.* Pharmacological options may be considered for individuals with a BMI ≥ 30 or a $27 \leq$ BMI ≤ 29 with at least one weight-related coexisting condition [250]. These individuals may not have achieved adequate weight loss through lifestyle interventions alone. Medications may include drugs that reduce appetite and nutrient absorption, such as orlistat, lorcaserin, and liraglutide, or those that modify neurochemical pathways involved in appetite regulation, like phentermine-topiramate and naltrexone-bupropion [250]. However, the use of pharmacotherapy is generally limited to severe cases due to associated risk factors and common weight regain after the termination of drug treatment [250].

- *Bariatric surgery.* Bariatric surgery may be considered for individuals with a BMI ≥ 40 or ≥ 35 with serious obesity-related comorbidities who have not achieved adequate weight loss through lifestyle interventions alone [255]. Currently, three main types of procedures are utilized, including adjustable gastric banding, gastric sleeve, and gastric bypass [250]. Adjustable gastric banding is considered the less invasive procedure and involves placing a silicone band around the gastric fundus to create a pouch. The band can be inflated with a saline solution to induce early satiety, contributing to moderate weight reduction. It can be adjusted by adding or removing saline solution to achieve tailored control. In sleeve gastrectomy, a large portion of the stomach

is surgically removed, reducing food intake. Lastly, gastric bypass (Roux-en-Y gastric bypass) restricts food consumption by creating a small pouch at the top of the stomach. The small intestine is then rearranged to connect to the newly created pouch, bypassing a portion of the stomach and the upper part of the intestine. In this case, too, food intake is reduced, leading to weight loss and improvements in metabolic parameters [256].

5.2 Gastroesophageal Reflux Disorder

5.2.1 Incidence and Prevalence

Gastroesophageal reflux disorder (GERD) is a chronic condition characterized by recurrent and troublesome heartburn, marked by a burning sensation in the chest or throat, and regurgitation. It is estimated to affect approximately 13% of the global population [257]. A considerable geographic variation is however appreciable, with higher rates, around 20%, in high-income countries like those in North America and Europe [258, 259]. Interestingly, the prevalence of the disease has increased by approximately 50% since the 1990s but has since stabilized [257].

The prevalence of GERD is influenced by various factors, including age, gender, lifestyle habits, genetic predisposition, and the presence of comorbidities such as obesity and hiatal hernia. GERD is more common in older adults, and there may be variations in symptom presentation and severity between males and females [258].

5.2.2 Pathophysiology

GERD is a pathological condition characterized by a dysfunction at the level of the esophagogastric junction barrier, resulting in increased regurgitation of acidic gastric contents into the esophagus [258]. The pathophysiology of GERD is multifactorial and involves a combination of mechanisms, including the reduced tone of the lower esophageal sphincter (LES), the presence of a hiatal hernia, esophageal motility, and delayed gastric emptying [259].

The LES is a circular band of muscle located at the junction between the esophagus and the stomach, primarily responsible for preventing the reflux of stomach

contents into the esophagus. In individuals with GERD, the LES may exhibit decreased resting tone or inappropriate relaxation, reducing its effectiveness. Factors contributing to LES dysfunction include genetic factors, hormonal influences, impaired neural control, and the effects of substances such as alcohol, caffeine, and tobacco. The presence of a hiatal hernia, where a portion of the stomach protrudes through the diaphragm into the chest cavity, can further weaken the LES and impair its proper function, making individuals more susceptible to GERD [259].

Reduced esophageal peristalsis can contribute to impaired esophageal clearance by prolonging exposure to corrosive acid substances, increasing the likelihood of tissue damage. Additionally, although the underlying mechanism is not entirely clear, delayed gastric emptying can increase the volume and pressure of gastric contents, promoting reflux into the esophagus [259].

5.2.3 Etiology

The primary cause of GERD remains unknown, but several risk factors have been identified in its pathogenesis. Increasing BMI in obese individuals is associated with a higher risk of developing GERD. This is primarily due to the reduction of lower esophageal sphincter pressure, a higher incidence of hiatal hernia, and an increase in intra-gastric pressure resulting from fat accumulation [260, 258, 257]. Similarly, tobacco use and alcohol consumption have also been shown to exacerbate GERD symptoms, although the extent of their effects is still a subject of debate [257, 258].

While the etiology of GERD appears to be primarily influenced by environmental factors, genomic studies have revealed the possibility of heritability. However, no individual mutation has been found to be significantly associated with the development of GERD, suggesting a polygenic scenario [257, 258].

5.2.4 Symptoms

The primary symptoms of GERD include heartburn and acid regurgitation, which significantly impact ADLs [258, 257]. Most GERD patients report a burning sensation or discomfort in the chest, behind the breastbone, or in the throat. This heartburn typically occurs after meals, especially when lying down or bending over. Moreover, night-time episodes can also lead to sleep difficulties [258]. Acid regurgitation is a

frequent symptom and can occur either along with or independently of heartburn. Chest pain, which can be similar to cardiac pain, is also a common symptom that may occur alone or in conjunction with heartburn and regurgitation [261].

The clinical spectrum of GERD patients can include less common symptoms, such as dysphagia (difficulty swallowing), chronic cough, asthma, chronic laryngitis, hoarseness, and teeth erosion due to frequent exposure to acidic and irritating liquids [258].

5.2.5 Diagnostic Criteria and Complicating Factors

The diagnosis of GERD is a comprehensive process that involves a systematic evaluation, including clinical assessment, an analysis of the cardinal symptoms, a review of the patient's medical history, and specific diagnostic tests.

The initial step always begins with a thorough analysis of the patient's medical history to assess the frequency and duration of common symptoms, such as heartburn and regurgitation. However, it is important to note that GERD symptoms are non-specific and can overlap with those of other disorders, making the diagnosis process more complex. In cases where the clinical presentation strongly indicates the presence of GERD, guidelines recommend a short-term trial of Proton Pump Inhibitors (PPI) or other acid-suppressing medications to assess the response of symptoms. If there is no improvement in symptoms following this therapy but the diagnosis still appears likely, more precise diagnostic tests, including endoscopy, esophageal manometry, and pH monitoring, can be utilized [258].

5.2.6 Treatment

The treatment of GERD involves a multifaceted approach with the goal of alleviating symptoms, healing esophageal inflammation, preventing complications, and enhancing the patient's quality of life.

Lifestyle interventions, accompanied by dedicated counseling, have proven to be effective in managing GERD [258, 261]. Specifically, weight loss and smoking cessation are crucial in reducing GERD symptoms. For individuals with nocturnal GERD, recommendations include elevating the head of the bed and avoiding late-night meals. In terms of pharmacological treatment, PPIs such as omeprazole,

lansoprazole, and pantoprazole are commonly prescribed to reduce gastric acid production by inhibiting hydrogen-potassium ATPase in the parietal cells of the stomach [258]. However, recent evidence, while not definitive, has suggested possible risks of adverse effects associated with prolonged use of PPI therapy. Therefore, patients who do not experience adequate relief after 4 to 8 weeks of treatment should undergo further evaluation for a more accurate differential diagnosis and a more tailored treatment plan [258].

In cases where medications and lifestyle modifications are ineffective or not well-tolerated, surgical intervention may be considered. Potential options for GERD patients include laparoscopic fundoplication or bariatric surgery, particularly when obesity is the primary cause of reflux disease [258]. These procedures involve wrapping the upper portion of the stomach around the LES to strengthen the barrier and prevent reflux. However, the invasiveness of fundoplication as a treatment option requires careful consideration in accordance with established guidelines. It is typically reserved to selected patients who have undergone comprehensive and objective assessments, especially if they are young and in good health. Recently, there have been emerging endoscopic and less invasive surgical techniques that show promise in reducing the reliance on long-term PPI and fundoplication. Nonetheless, the long-term safety and efficacy of these approaches still require scientific validation [258]

5.3 Effects of Obesity and GERD on Voice Production

GERD and obesity are complex and prevalent medical conditions that can have a significant impact on an individual's health and quality of life. GERD is characterized by the backflow of stomach acid and irritants into the throat, often resulting in inflammation of the vocal cords, which can manifest as hoarseness, voice changes, or chronic laryngitis. In contrast, obesity, defined by excessive accumulation of adipose tissue, is associated with various health challenges, including alterations in respiratory patterns due to increased body weight, which can influence the process of vocal production.

The assessment and diagnosis of GERD traditionally involve invasive techniques, often relying on the subject's response to pharmacological therapy, which may have numerous side effects. Moreover, the co-occurrence of GERD and obesity is not

well-understood, particularly regarding its impact on vocal production. This lack of knowledge underscores the need for detailed investigations into how the concurrent presence of these two conditions affects the overall health status and, in particular, the vocal signals.

Within this context, this section provides an in-depth analysis of the influence of GERD and obesity on vocal signals. Specifically, it explores whether the coexistence of these conditions alters the produced vocal signal and, if so, the nature of these alterations. Through the use of automatic models based on vocal analysis, this study aims to shed light on the vocal characteristics associated with GERD and obesity, ultimately contributing to a better understanding of these conditions and their effects on voice. The results from this study are published in [262].

5.3.1 Related studies

A recent body of literature have initiated investigations into the relationship between body weight and voice, particularly focusing on vocal alterations in obese patients (OP) [20] due to the breathing compliance reduction, airflow resistance increase, and respiratory muscle disorders that typically arise as body weight increases [263]. However, these studies have provided only limited insights into the impact of obesity on vocal production [264–266]

A study conducted by Souza et al. [20] delved into the effects of obesity on voice by analyzing data from 84 female participants. The study identified a negative correlation between BMI and key vocal parameters, including the F0 and MPT. Nevertheless, this analysis utilized a restricted set of vocal features and exclusively included female subjects. Similarly, Fonseca et al. [21] examined the impact of obesity on voice in a study involving 114 obese patients (52 before and 62 after bariatric surgery) and 20 HCs. The study, however, suffered from limitations, including the use of a narrow set of vocal features and an imbalanced distribution of subjects among classes, potentially introducing bias.

Regarding GERD, Milani et al. [267] explored the feasibility of classifying individuals into two groups: HC and patients with GERD (PR). Their approach involved feeding a classifier with MFCCs derived from voice samples. The dataset consisted of 30 voice samples from patients, equally distributed among hyperkinetic, hypokinetic dysphonia, and left cricothyroid muscle diseases, along with 10 HC

samples. While the dataset was substantial, the study solely reported classification accuracy without presenting additional performance metrics. Nevertheless, the reported average classification accuracy of 0.88 across the four classes suggests the potential for assessing GERD through voice recordings.

5.3.2 Materials

The study involved the recruitment of a total of 92 participants, who were classified into four groups: HC, PR,OP, and obese patients with concomitant GERD (OPR). These participants, all aged *geq* 18 years, met the criteria for inclusion, which excluded any phonatory apparatus injuries such as vocal cord paralysis. The assessment of obesity and GERD was performed by clinicians using BMI measurements and symptom evaluation questionnaires, respectively.

The vocal samples were recorded under controlled conditions in a spacious, soundproof environment to minimize external interference. Each participant was individually seated, instructed to maintain a posture with their back and arms supported by the backrest and armrests of the chair, and asked to speak at a comfortable volume. Microphones were positioned at a consistent distance of 5 cm from the participant's mouth to ensure standardized recording conditions. Vocal samples were captured using a dynamic headset microphone (WH20) manufactured by Shure (USA), featuring a male 3-pin XLR connector. The recordings were made using a H5 voice recorder produced by Zoom (Tokyo, Japan), capturing the audio in a high-quality uncompressed format (.wav) at 24-bit resolution and a 44.1 kHz sampling rate.

The experimental protocol was tailored for the Italian language and included two specific vocal tasks. The first task involved sustained phonation of the vowel /a/, while the second task required participants to repetitively articulate a designated sentence, ˈa kaβˈal donˈato nˈon sˈi ɣwˈarða ˈin bˈokka (i.e., Don't look a gift horse in the mouth) (S1). This sentence was chosen to provide insights into the prosodic and articulatory capabilities of the participants.

Of particular significance was the selection of sentence S1, which features occlusive sounds. As previously discussed (Section 4.5.2), unlike other consonants, occlusives are characterized by a sudden obstruction of airflow by the articulators. Hence, they are highly dependent on precise airflow regulation and involve intricate articulatory movements. For this reason, S1 was considered suitable for capturing

vocal alterations in the various patient groups given that individuals with a higher body fat percentage may experience respiratory system changes [263]. Moreover, patients with GERD may encounter challenges related to laryngeal control and structural changes in laryngeal tissue [268].

Data analysis was carried out in Python, with the primary tools being Praat, which facilitated the pre-processing of audio recordings and the extraction of relevant vocal features. This research adhered to the principles outlined in the Declaration of Helsinki and was granted approval by the Ethics Committee of the *Policlinico di Tor Vergata* under approval number 7/20. All participants provided written informed consent for their involvement in the study, and all demographic and clinical information was documented anonymously.

Table 5.2 provides a summary of the key characteristics of each group, offering a clear overview of the study's participant demographics and distribution.

Table 5.2 Demographic details of participants in the study. Results are reported in terms of mean and standard deviation. GERD: Gastro-Esophageal Reflux Disorder; M: Male; F:Female

	Healthy Control	GERD	Obesity	GERD & Obesity
Age	43 ± 15.8	45 ± 19.1	59 ± 9.2	55 ± 10.9
Gender	8M, 19F	6M, 15F	1M, 17F	11M, 15F
BMI	22.71 ± 2.0	22.6 ± 1.7	34 ± 4.6	37 ± 8.9

5.3.3 Methods

Pre-Processing

The initial preprocessing steps involved the normalization of recorded audio signals to fall within the range of [-1, 1]. This normalization procedure was applied to mitigate potential variations in the recorded signals caused by differences in the speaker-microphone distance, thus ensuring a standardized input for subsequent analysis.

Following normalization, the segments of initial and final silence were removed from the audio recordings. Utilizing Praat software, the onset and offset points of

voiced regions were detected within the audio signals. This step was essential for isolating the regions of interest, where vocalization occurred.

It is noteworthy that, due to the use of professional recording equipment with high signal quality, no denoising filters were applied to preserve all relevant information within the recorded audio, as the application of denoising filters could potentially result in the loss of significant acoustic details critical for subsequent analysis.

Feature Extraction

Given the distinct methodological approach taken in this study and the absence of a predefined feature set with established high relevance to the specific application, an empirical feature extraction process was undertaken, encompassing approximately 500 features for the vowel /a/ and 600 for the sentence S1, subsequently evaluated for their efficacy.

Going into more detail, following the identification and merging of voiced regions, each audio signal was segmented into 40 ms windows with a 50% overlap, facilitating the extraction of features from each of these temporal segments. These extracted features were then grouped together into a unified vector, upon which five key statistical measures were computed, namely, the mean value, median value, standard deviation, kurtosis, and skewness. This set of *low-level* features was augmented with *high-level features*, computed over the entire duration of the signal, which included measures such as Jitter, Shimmer, among the others.

To ensure uniform scaling of features and mitigate the influence of potential outliers on model performance, a Z-score normalization was applied to the entire feature set.

An overview of the selected features and their relevant information is provided in Table 5.3.

Feature Selection

The feature selection procedure employed in this study was adapted from a similar investigation focusing on individuals with PDP, previously discussed in Section 4.4.1. This methodology leveraged a correlation-based approach, aimed at identifying

Table 5.3 Features employed in the study together with the region they were applied to. V: Vowel / a/, Vs: Voiced regions of S1 sentence

Feature Name	Applied to
F0	V, Vs
1^{st}, 2^{nd}, 3^{rd} Formants	V, Vs
13 MFCC + Δ, ΔΔ	V, Vs
Jitter (Lab, RAP, DDP, PPQ5)	V
Shimmer (dB, APQ3, APQ5, APQ11)	V
Intensity	V, Vs
STE	V
Noise: GNE, HNR, CPP	V, Vs
Spectral features: mean, std, skew., kurt., roll-off, slope	Vw, Vs
13 RASTA-PLP, + Δ, ΔΔ	V, Vs
24 BBE	V, Vs
DFA	V

the most influential features, characterized by high feature-target correlation, while simultaneously ensuring their non-redundancy.

The procedure unfolded in two primary stages. Firstly, an evaluation of the absolute value of Pearson's correlation coefficient (r) between the features and the target variable (r_{ft}) was carried out, and only the most substantial correlations were preserved ($r > th_1$, $P < th_2$). Subsequently, intra-feature correlations (r_{ff}) were computed, and in cases where inter-correlation surpassed intra-correlation (i.e., $r_{ff} > r_{ft}$), the feature demonstrating weaker correlation with the target variable was eliminated.

To determine the optimal threshold values tailored to the specific application at hand, initial values of $th_1 = 0.3$ and $th_2 = 0.05$ were set. These threshold values were then iteratively fine-tuned, ranging from 0.3 to 0.7 for th_1 and from 0.03 to 0.07 for th_2, with incremental steps of 0.1 and 0.01, respectively. The selection of optimal threshold values was based on the evauation of the highest accuracy achieved within a 5-fold CV.

Classification

Given the aim of this study (i.e., assess the influence of GERD and obesity of vocal samples as well as investigate how the concurrent presence of these two conditions

affects the signals generated) four different binary classifications were performed, namely (i) HC vs PR; (ii) OP vs OPR; (iii) HC vs OP; (iv) PR vs OPR. Subjects without GERD or obesity were categorized into the HC group, while individuals with either reflux or obesity were placed in the PR and OP groups, respectively. The OPR group consisted of patients with both reflux and obesity.

To prevent potential issues with model generalization, the dataset was randomly divided into two subsets: 80% for training and validation, and the remaining 20% for testing. Feature selection, model selection, and optimization were exclusively performed on the training and validation set, while the testing set did not undergo further optimization procedures.

Features extracted from / a/ and S1 were combined into a single vector through an *early fusion* approach (Section 4.4.1) and used as input for the classifier. Four different models were assessed: NB, KNN, RF, and SVM. The model exhibiting the highest performance was selected for further refinement, which involved hyper-parameter optimization using a Grid Search approach. The best accuracy achieved in a 5-fold CV was used as the primary metric for model comparisons.

To evaluate the stability of the final model, given the random data splitting procedure, performance metrics such as accuracy, F1-score, precision, sensitivity, specificity, and AUC were calculated as averages over five iterations.

5.3.4 Results

Prior to feature selection, a comprehensive exploration was conducted to assess the importance of the features. This evaluation was based on Pearson's correlation coefficients between each feature and the respective class. Table 5.5 presents the three most correlated features along with their associated p-values for each class under investigation. Subsequently, the feature selection process was executed. The outcomes of this procedure are detailed in Table 5.6.

The reduced set of features was utilized as input for four distinct classifiers, and the classifier achieving the highest classification accuracy was chosen. The comparison of these classifiers are reported in Table 5.4.

The best models eventually underwent a grid-search optimization process. The parameters considered for optimization included C values (1, 10, 100, 1000), gamma

Table 5.4 Classification accuracy of the four models tested

	HC vs PR	OP vs OPR	HC vs OP	PR vs OPR
Model	SVM	NB	NB	SVM
Accuracy	0.95	0.86	0.84	0.94

Table 5.5 Importance of the top-three selected features from each binary classification performed. S: Sentence, V: Vowel

	Feature	Task	r	P-value
H vs PR	ΔPLP 5 skewness	S	-0.42	0.003
	3 BBE skewness	S	0.40	0.005
	13 MFCC kurtosis	V	0.40	0.005
	6 PLP median	V	-0.50	0.001
	11 MFCC kurtosis	V	-0.43	0.004
H vs OP	BBE8 median	S	-0.48	<0.001
	GNE median	S	-0.47	<0.001
	7 BBE mean	S	-0.46	0.001
PR vs OPR	12 ΔΔPLP skewness	S	-0.57	<0.001
	1 ΔΔMFCC std	S	-0.52	0.001
	1 BBE std	S	-0.50	0.001

values (0.1, 0.01, 0.001, 0.0001), and kernel options (linear, polynomial, Radial Basis Function (RBF)) for SVM, as well as smoothing values (ranging from 1 to $1e^{-09}$ with steps of $1e^{-01}$) for NB. The best hyper-parameters identified were C = 1, gamma = 0.01, kernel = RBF for SVM, and smoothing = 1.0 for NB.

The performance of the optimized models is reported in Table 5.8.

5.3.5 Discussion

The results obtained in this study confirmed the influence of GERD and obesity on vocal production as well as the potential of utilizing ML analysis of vocal tests as a non-invasive, cost-effective, and efficient tool for the detection of GERD.

The feature relevance analysis, as presented in Table 5.5, underscores the significance of features such as MFCCs, PLP, and BBE. These features are commonly employed to model the vocal tract during articulation, capturing the resonance prop-

Table 5.6 Features selected from sentence and vowel repetition tasks for each binary classification performed

	From Vowel	From Phrase
HC vs PR	CPP; MFCC: 8,13 MFCC Δ: 3,6,10; MFCCΔΔ: 3,7; PLP: 11; PLPΔΔ	1^{st} Formant; HNR; BBE: 1,3,4; MFCC: 3; MFCCΔ: 2,12; MFCCΔΔ: 1,12; PLP: 1,11,12; PLPΔ: 1,5,9,12; PLPΔΔ: 2, 11
OP vs OPR	2^{nd} Formant; Spectral Slope; MFCC: 11; MFCCΔ: 6,12,13; MFCCΔΔ: 11-13; PLP: 6; PLPΔΔ: 9	GNE; Spectral mean; BBE: 1-5,22;MFCC: 8; MFCCΔΔ: 2; PLP: 2,5-7,10; PLPΔ: 2,5;PLPΔΔ: 5,12
HC vs OP	MFCCΔ: 4,7	GNE; BBE: 4,5,7-9,12; MFCCΔ: 3,8; PLP: 5,7,12; PLPΔ: 3; PLPΔΔ: 7,8,10
PR vs OPR	3^{rd} Formants ; Spectral features: flux, kurtosis, skewness, roll-off pt; HNR; MFCC: 9,11,13; MFCCΔ: 12; MFCCΔΔ: 1,12,13	Spectral features: mean, skewness; BBE: 1,8,9,12,13; MFCC: 3; MFCCΔ: 1; PLP: 0,2,4; PLPΔ: 0-2,11; PLPΔΔ: 0,2

Table 5.7 Performance of the optimized models on the validation set.

	HC vs PR	OP vs OPR	HC vs OP	PR vs OPR
Accuracy	0.8 ± 0.06	0.78 ± 0.06	0.82 ± 0.04	0.87 ± 0.04
Precision	0.96 ± 0.03	0.71 ± 0.06	0.87 ± 0.06	0.91 ± 0.06
F1	0.79 ± 0.092	0.77 ± 0.05	0.80 ± 0.05	0.84 ± 0.06
Sensibility	0.72 ± 0.12	0.89 ± 0.04	0.78 ± 0.04	0.81 ± 0.08
Specificity	0.97 ± 0.02	0.72 ± 0.10	0.86 ± 0.06	0.93 ± 0.04
AUC	0.94 ± 0.03	0.87 ± 0.05	0.92 ± 0.03	0.92 ± 0.03

Table 5.8 Performance of the optimized models on the test set.

	HC vs PR	OP vs OPR	HC vs OP	PR vs OPR
Accuracy	0.84 ± 0.08	0.71 ± 0.11	0.80 ± 0.15	0.75 ± 0.11
Precision	0.85 ± 0.12	0.68 ± 0.18	0.81 ± 0.16	0.81 ± 0.16
F1	0.78 ± 0.11	0.72 ± 0.09	0.76 ± 0.14	0.73 ± 0.12
Sensibility	0.75 ± 0.16	0.80 ± 0.10	0.72 ± 0.20	0.70 ± 0.19
Specificity	0.90 ± 0.08	0.64 ± 0.23	0.87 ± 0.12	0.80 ± 0.19
AUC	0.83 ± 0.09	0.72 ± 0.10	0.79 ± 0.15	0.75 ± 0.11

erties of the supralaryngeal vocal tract and detecting kinematic changes in the vocal apparatus, as noted in relevant literature [269]. The importance of MFCC for GERD patients was also highlighted in a prior study [267]. The inclusion of noise-related features supports the notion of altered noise levels in pathological voices, consistent with findings in the literature [270, 268].

As for the vocal tasks employed, the evidence indicated S1 demonstrating higher efficacy compared to / a/. This observation is further supported by the results of the feature selection process, as detailed in Table 5.6. Indeed, features derived from S1 were more frequently selected.

In terms of classification outcomes, the model demonstrates robust performance, with no significant decline observed when transitioning from the validation to the test set, as presented in Table 5.8. This absence of over-fitting and the consistent performance across iterations suggest excellent generalization capability.

An analysis of the four experiments conducted reveals that the model ability to detect GERD outperforms its ability to detect obesity, indicating a potential interplay between the two conditions. Notably, GERD detection in the presence of obesity yields lower accuracy compared to cases without obesity, both in the validation and testing phases. Similarly, the detection of obesity in the presence of GERD shows slightly better performance during validation but exhibits a substantial reduction when applied to the testing set, implying a reduced overall classification capability. This suggests that the presence of GERD may induce more pronounced variations than obesity alone, and the concurrent presence of the two conditions may lead to distinct vocal changes compared to their individual presence.

5.3.6 Conclusion and Future Works

This study explored the feasibility of classifying patients with obesity and GERD through an analysis of vocal tests. Despite the simplicity of the workflow, ML models demonstrated efficiency, achieving accuracy ranging from 0.78 to 0.87 in CV and from 0.71 to 0.84 in testing. The higher performance observed in GERD detection may be attributed to its more pronounced impact on the vocal apparatus. Furthermore, a mutual influence between GERD and obesity was evident, with the presence of obesity associated with reduced model performance.

Future research will address the current limitations stemming from dataset size and composition. These limitations mainly include an unbalanced distribution of gender and age, recognized for their influence on voice characteristics, as noted in Section 4.7. Additionally, the assessment of obesity was based on BMI, without considering the actual distribution of body fat, which will be taken into account in future investigations.

Moreover, although the analyses conducted thus far have been focused on GERD and obesity, the findings are potentially applicable to more impactful health conditions, such as PD. In this context, future studies will extend the investigation to simultaneously evaluate two co-existing conditions within this field, aiming to validate how the concurrent presence of two diseases influences vocal analyses.

Chapter 6

Application IV: Sleep Quality

6.1 Sleep Quality

6.1.1 Statistics

Recent estimates reveal that sleep disorders impact a significant portion of the world population, with approximately 50 to 70 million people affected in the United States alone [271]. Among these, insomnia and sleep apnea are among the most common alterations, with an increasing trend of incidence. [271]. Moreover, poor sleep is intricately linked to a higher incidence of chronic health conditions, including cardiovascular diseases, diabetes, obesity, and mental health issues [271].

According to the 2011-2014 report from the National Center for Health Statistics, 31.7% of US adults do not meet the National Sleep Foundation recommendation for at least 7 hours of sleep per night [271]. This emerging trend is due to an array of sociodemographic, social integration, and health behavior factors that play a role in sleep quality and duration.

Among them, gender disparities are evident, with women more prone to sleep problems and often experiencing shorter sleep durations compared to men. Additionally, non-urban residence fosters improved sleep quality, potentially due to reduced urban noise and pollution levels. The socioeconomic status exhibits mixed associations, with some studies correlating low income and education with sleep issues, while others fail to establish such a connection.

Moreover, health behaviors such as insufficient physical exercise, smoking, and excessive alcohol consumption can diminish sleep quality, though these associations may vary across different populations [272].

6.1.2 Pathophysiology

Sleep is a dynamic and transitory state of altered consciousness that encompasses a multifaceted array of functions. While it certainly plays a crucial role in promoting restorative processes, it extends its significance beyond it. One of its pivotal contributions lies in the active participation of the brain glymphatic system, which is responsible for the efficient clearance of metabolic waste products from the central nervous system (CNS) [273, 274].

The glymphatic system constitutes a complex and finely regulated neural waste clearance mechanism which predominantly operates during non-rapid eye moment (NREM) sleep, pronounced predominance during slow-wave sleep (SWS). In more detail, during wakefulness, the interstitial space within the brain is limited, hindered by the swelling of glial cells that support neuronal function. However, as an individual transitions into NREM sleep, the glial cells, particularly astrocytes, undergo a process of cell volume regulation. This causes a noticeable shrinkage of these cells, effectively expanding the interstitial space. This facilitate the bulk flow of cerebrospinal fluid (CSF) through the brain parenchyma. The CSF, laden with essential nutrients and freshly oxygenated resources, enters the brain tissue through the perivascular channels surrounding arteries. As it permeates the parenchyma, it carries with it the accumulated metabolic waste products and cellular debris, including soluble proteins such as beta-amyloids , which are known biomarkers in neurodegenerative disorders like PD [273, 274].

Poor sleep quality may led to both short-term and long-term consequences on physical, mental, and emotional well-being. As for the former, increased fatigue, mood disturbances, and cognitive or physical impariment are among the most frequent. As for long-term effects, persistent poor sleep quality is associated with chronic health condition, cognitive decline, as well as menatl health disorders and cognitive decline.

6.1.3 Assessment Criteria and Complicating Factors

The evaluation of sleep quality encompasses a range of multidimensional criteria. Key parameters include sleep latency, quantifying the time required to initiate sleep, and sleep duration, reflecting the total duration of sleep achieved during the sleep period. Additionally, sleep architecture, characterized by sleep stage distributions, rapid eye movement (REM) sleep patterns, and non-REM sleep stages, plays a crucial role in evaluating sleep quality. Moreover, the presence of repetitive patterns of sleep disruption over weeks or months completes the diagnostic scenario in assessing sleep quality [275].

Polysomnography (PSG) [276], a common method for assessing sleep, is an invasive diagnostic test involving the recording of biosignals during sleep using numerous electrodes. This comprehensive procedure includes electroencephalography (EEG) for monitoring brain activity, electrooculography (EOG) to track eye movement, and electromyography (EMG) for assessing muscle tone. Additional measurements include respiratory parameters, such as airflow, thoracic and abdominal effort, and blood oxygen levels, recorded through nasal pressure sensors, respiratory inductance plethysmography, and pulse oximetry, respectively. Electrocardiography (ECG) provides information about cardiac activity, while leg movement sensors detect any periodic leg movements during sleep. However, despite its comprehensiveness, PSG has several drawbacks, including its intrusive and costly nature, which often necessitates an overnight stay in a sleep laboratory. The presence of numerous sensors and electrodes may disrupt the natural sleep environment, potentially influencing sleep quality. Moreover, the complexity and cost of polysomnography restrict its widespread use and accessibility, thereby making alternative methods increasingly attractive for assessing sleep quality.

To mitigate the limitations associated with these techniques, subjective assessments through standardized questionnaires are frequently employed. The Pittsburgh Sleep Quality Index (PSQI) [277] is one of the most renowned tools for capturing individuals' subjective perceptions of their sleep quality. However, evaluating sleep quality through questionnaires may introduce biases based on individuals' subjective perceptions, hence additional objective tests are often necessary for a comprehensive assessment.

6.2 Automated Vocal Analysis for Sleep Quality Assessment

The study of sleep quality is an exceptionally vital field within medicine, considering the prevalence of conditions arising from poor sleep quality. Furthermore, numerous studies have already demonstrated how altered sleep patterns during various sleep phases can represent the precursors to various neurodegenerative diseases, such as PD (Section 4.1). However, the analysis of sleep quality presently relies on highly invasive and complex methodologies, like PSG, involving the collection of biosignals during sleep through a significant number of electrodes.

Within this context, this study aims to explore the possibility of utilizing vocal signals to support sleep monitoring. This investigation is grounded in the concept that poor sleep quality is often associated with an overall fatigue that manifests as perceivable alterations in vocal production. In parallel, by examining the correlation between sleep quality indices and vocal signals, the second objective of this study is to assess whether and to what extent sleep rhythm disruptions may influence the produced vocal signal. Beyond sleep analysis itself, insights in this regard could be highly relevant in the periodic monitoring of various conditions through vocal sample collection. As previously discussed (Section 4.1), in cases where vocal sample collections occur sporadically, it is crucial to understand which cofactors affect the produced signal and which alterations are characteristic of the examined condition. In this context, comprehending the relationship between sleep quality and vocal production may prove to be pivotal.

6.2.1 Related Literature

In recent years, an increasing body of literature have initiated investigated the relationship between sleep quality and voice. Within this context, the study by Icht et al. [278] presents compelling evidence of a degradation in voice quality under stressful conditions, notably in the context of sleep deprivation. This investigation involved the analysis of vocal samples from 47 healthy participants, which were recorded using professional-grade equipment while the subjects performed various vocal tasks. Notably, the results of this study unveiled a significant decline in HNR

values following 24 hours of sleep deprivation, with the most pronounced effects observed within the female subgroup.

In a related research endeavor undertaken by Kim et al. [279], the focus shifted towards the potential of voice analysis in predicting Sleep Quality (SQ). In this study, 203 healthy native English speakers were enlisted to complete a series of questionnaires through mobile devices and provide voice samples encompassing free speech, sentence articulation, and text reading. The evaluation of regression performance, conducted through a 5-fold CV framework, centered on the Concordance Correlation Coefficient (CCC) between the actual and predicted SQ scores. The study reported a CCC value of 0.5 for the SQ index, indicative of promising outcomes.

Furthermore, in a study by Botelho et al. [280], an investigation into the detection of Obstructive Sleep Apnea (OSA) was conducted, drawing upon speech samples from 45 Portuguese subjects, comprising 25 individuals with OSA and 20 control subjects. The participants engaged in both free monologue and text reading tasks. Vocal features, including F0, HNR, and cepstral coefficients, were computed and subsequently input into an ensemble model employing SVM, LDA, KNN. The results of this study demonstrated a notable performance, yielding 88% True Positive Rate (TPR) and 80% True Negative Rate (TNR) in the context of OSA detection.

6.2.2 Materials

Data collection for this study involved Italian speakers who accessed a user-friendly web application (WA) accessible on both desktop and mobile web browsers. The WA was meticulously designed to guide users through a voice test and two sleep-related questionnaires. The voice test required participants to read a phonemically balanced text, as proposed in [104] and previously reported in Section 4.3.1. This text was selected for its ability to capture various aspects of altered voices, given its length, intricate phonetics, and the need for expressive variation during reading.

Following the vocal task, participants were required to complete a questionnaire assessing their sleep quality, in addition to a survey which inquired about their daily habits and sleep-wake cycle. Furthermore, a section was included in which participants provided their age, gender, and level of education.

In total, 135 anonymous volunteers took part in the study, of which 55 were male. Among these, 70 subjects (37 males) completed all the tasks, including the

recording and the sleep questionnaires. For the subsequent analysis, this subgroup was the primary focus. Given the impact of gender on many of the extracted vocal features, the dataset was divided into two groups (males and females), and the same analysis workflow was applied to each cluster. Data collection adhered to the principles outlined in the Declaration of Helsinki and received approval from the Ethics Committee of the A.O.U Città della Salute e della Scienza di Torino (number 00384/2020). Informed consent was obtained for the observational study, with demographic and clinical data collected anonymously.

Table 6.1 provides an overview of the information gathered. Regarding age, the values listed exclude 10 subjects (8 females) who did not input their age on the online form. Given that this subset represents 14% of the entire group (and 23% of the female subjects), for subsequent analysis the missing values were imputed with the median age of the entire group, categorized by gender.

Data analysis and classification were conducted using Python, with Praat predominantly utilized for pre-processing and feature extraction.

Table 6.1 Demographics of the included subjects

	Age	Education Level	Remote Working	sPSQI	SLEEPS score
Female	41.4 ± 18.1	Middle: 3 (9%) Secondary: 7 (21%) Degree: 23 (70%)	9 (27%)	6.38 ± 1.51	2.29 ± 1.04
Male	36.9 ± 14.5	Middle: 2 (5%) Secondary: 10 (27%) Degree: 25 (68%))	12 (33%)	6.05 ± 1.88	2.39 ± 0.92

6.2.3 Methods

Vocal Analysis

Signal pre-processing. To enhance the data quality and consistency, several pre-processing procedures were applied. Firstly, the recordings were down-sampled to 16 kHz, facilitating uniformity across the dataset. Additionally, a de-noising filter with default Praat hyperparameters was applied to each signal, effectively reducing background noise interference. To ensure the model robustness and mitigate any potential impact of variations in speaker-microphone distances, the signal amplitudes were normalized within the range [0, 1]. Initial and final silence regions were manually removed, obviating the need for further preparatory steps. Finally, the Praat software was employed to identify the start and end points of voiced regions within the vocal signals.

Feature extraction. Given that a well-defined set of features tailored explicitly for this application was unavailable, it was opted to extract a comprehensive array of 96 vocal features, aiming to explore their efficacy. Two distinct groups of features were derived, with the first group focusing on timing measures extracted from the entire vocal signal. These features aimed to discern alterations in the rhythmic characteristics of speech and encompassed the NP, the DPI, and the RST metrics. The second group of features focused on periodicity measures, and included F0 and the first three formants, noise measures, such as HNR, CPP, GNE, as well as amplitude-related features encompassing Intensity. Furthermore, spectral and cepstral features, such as 12 MFCC and 13 PLP coefficients, were extracted. These features had proven their effectiveness in similar studies, as documented in [279, 280]

Each signal was segmented into 25 ms windows with a 10 ms overlap after the identification and merging of voiced regions. Features were then extracted from each segment, unified into a feature vector, and subjected to the computation of five essential statistics, namely the mean value, median value, standard deviation, kurtosis, and skewness. Furthermore, Z-score normalization was applied to the entire feature set, thereby ensuring consistent scaling across the features.

Sleep Analysis

Sleep Quality Assessment. To evaluate the overall sleep quality of participants, the shortened version of the Pittsburgh Sleep Quality Index (sPSQI) was selected, being a well-validated 13-item survey adapted from the original PSQI. The sPSQI provides a global score that distinguishes between individuals classified as having either *good* or *poor* sleep quality. Unlike the standard PSQI, the sPSQI solely relies on self-reported responses from participants. While PSG is considered the gold standard for diagnosing sleep disorders, the PSQI is a reliable and commonly used tool in both research and clinical practice.

The sPSQI evaluates sleep quality based on five components: Sleep Latency, Duration, Efficiency, Disturbances, and Daytime Dysfunction. Among the 70 subjects in the dataset, the sPSQI scores ranged from 0 to 10, with 10 indicating the most severe sleep disturbances. The average sPSQI score was 6.21 ± 1.72, with a median score of 6.0. Previous research has suggested a cutoff value of 4 for identifying poor sleepers using the shortened PSQI. However, based on the dataset range and distribution, a cutoff value of 5.0 was established, which classified 26 participants as good sleepers and 44 as poor sleepers.

Sleep Features. A supplementary sleep survey, denoted as SLEEPS, was administered through the WA. This survey comprised 21 self-reported items designed to offer insights into the sleep-wake schedules and habits of the participants, thereby investigating potential correlations with their sleep quality.

One item in the SLEEPS survey assessed the SLEEPQ, a parameter often evaluated using sleep diaries in actigraphy studies and compared to the actual value. Participants rated their sleep quality on a 5-point scale ranging from *Excellent* to *Very Poor*. All items in the survey were scored on a Likert scale from 0 to 4, with 4 indicating the most negative response. This scoring method aligns with established protocols as outlined in [277] and is consistent with the design of the sPSQI. Binary or numerical responses were appropriately adjusted to this scale. The set of questions within the SLEEPS questionnaire is reported in Table 6.2.

Table 6.2 Items and scores of the SLEEPS questionnaire

Item	Score
I. General Data	
Covid-19 diagnosis	Scale
OSA diagnosis	Scale
Insomnia	Scale
University	Y/N
Work	Y/N
II. Work and Study Habits	
Remote Working/Learning	Y/N
Hours of Screen Time	Scale
End of use Time of Electronic Devices	Numeric
Blue Light Filter	Y/N
III. Leisure Time Habits	
Time spent away from home during workdays	Numeric
Time spent outside over the weekend	Numeric
Excercise hours (outdoors, per week)	Numeric
Excercise hours (indoors, per week)	Numeric
Time spent working on a hobby (per week)	Numeric
IV. Sleep Habits	
Nocturnal awakenings	Scale
Getting up at night	Scale
Morning drowsiness	Scale
Morning fatigue	Scale
Fatigue, poor concentration and impaired performance	Scale
Difficulty falling asleep	Scale
Perceived sleep quality	Scale

Feature selection

In order to implement an adequate feature integration, an early fusion procedure was performed within the vocal and sleep parameters. Subsequently, a feature selection workflow, adapted from a similar study previously described in Section 4.4.1, was applied.

The feature selection process employed a correlation-based approach, aimed at identifying the most significant features with respect to their association with

the target variable (i.e., high feature-target correlation) while minimizing feature redundancy (i.e., low inter-feature correlation). The Pearson correlation coefficient r_{fo} between individual features and the target variable was assessed. The absolute value of this coefficient for each feature was calculated, and features demonstrating a substantial correlation (i.e., $r > 0.4$) with a corresponding p below 0.02 were retained. Subsequently, the intra-feature correlation (r_{ff}) was computed. For feature pairs in which the inter-correlation surpassed the intra-correlation (i.e., $r_{ff} > r_{fo}$), the feature demonstrating weaker correlation with the target variable was selectively removed. This process ensured the preservation of the most pertinent and informative features for subsequent analysis.

Classification

The study conducted automatic binary classification to distinguish subjects characterized by either good or poor sleep quality. Labels were determined by applying a common clinical practice approach, where a threshold on the continuous sPSQI score was set to demarcate the two categories, as documented in clinical guidelines [281]. Specifically, the quality threshold was established at 5.0, as detailed in Section 6.2.3. Subjects with a continuous sPSQI score above this threshold were categorized as poor sleepers.

To ensure the model generalization capabilities and mitigate overfitting, the corpus was randomly split into two subsets: 80% used for training and validation, and the remaining 20% for testing. The process of feature selection, model selection, and model optimization was exclusively carried out on the training and validation set, leaving the 20% of subjects for testing purposes.

Seven different ML classifiers (i.e., NB, SVM, KNN, RF, ADA, GB, BAG) were employed and compared, and the classifier achieving the highest F-1 score was selected for further optimization. Given the imbalanced dataset, where the count of poor sleepers outnumbered good sleepers, the F-1 score was preferred over the classification accuracy for performance evaluation. Due to the random dataset split, each experiment was conducted 10 times on 10 randomly drawn subsets, and the performance metrics averaged to facilitate classifier comparison. Furthermore, hyperparameter optimization employing a Grid Search approach was applied to the best-performing model.

6.2.4 Results

In this study, an examination of feature significance was conducted by assessing the Pearson correlation coefficients between each feature and the class variable. For the female population, the three most correlated acoustic features were identified as ΔΔMFCC12 std ($r: -0.70, p < 0.001$), ΔMFCC12 std ($r: -0.67, p < 0.001$), and ΔΔMFCC10 std ($r: 0.60, p < 0.001$). Similarly, among the male population, significant features were ΔMFCC3 mean ($r: 0.50, p < 0.001$) and ΔΔ MFCC6 mean ($r: -0.46, p < 0.004$), along with the SLEEPQ parameter from the sleep questionnaire ($r: -0.56, p < 0.001$).

A similar analysis was performed on the sleep scores obtained from the SLEEPS questionnaires to identify potential significant factors contributing to reduced sleep quality. Among the observed factors, a noteworthy connection was found with COVID-19 positivity. In light of this, the sPSQI scores were analyzed, taking into account the COVID-19 parameter, which distinguished subjects as having past positivity, current positivity, or no prior diagnosis. Among males, those with a history of COVID infection exhibited notably worse sPSQI scores, indicating poorer sleep quality. Conversely, for individuals in the N-condition (never diagnosed), the scores were more evenly distributed within the data range. In contrast, no similar trend was observed in the female group. Furthermore, in male subset sPSQI and SLEEPQ, both scores exhibited a degree of correlation. However, within the female group, some subjects demonstrated a tendency to overrate their sleep quality. Specifically, 35% of these individuals rated their sleep as *Average*, despite having sPSQI scores of 7.0 ± 1.53, and 32.4% assessed their sleep as *Above Average* with sPSQI scores of 6.0 ± 0.95—both indicating sleep quality below the average threshold. Lastly, the items collected via the SLEEPS questionnaire were ranked based on their correlation with the SLEEPQ parameter. As anticipated, factors such as the frequency of nocturnal awakenings, the presence of insomnia, and the total hours of sleep exhibited a strong correlation with perceived sleep quality. In contrast, no significant correlation was identified with variables such as remote working (or learning) or the use of a blue light filter. Instead, moderate correlations were observed with variables like the time of electronic device usage cessation, morning drowsiness, and difficulties in falling asleep.

As for the classification step on the female subgroup, among the seven different models the top-performing classifiers were BAG ($88\% \pm 3.4$), KNN ($87\% \pm 3.7$),

and SVM (86% $\pm$ 3.7). Given the absence of a definitively best-performing model, the optimization process was carried out on all these three classifiers, and the models with the highest performance following hyperparameter tuning were selected. For SVM, the optimal configuration included parameters C = 10, gamma = 0.001, and a RBF kernel. The same procedure was applied to the male subgroup, where BAG, KNN, and SVM classifiers exhibited the best performance with F-1 scores of 77% $\pm$ 4.4, 77% $\pm$ 5.8, and 80% $\pm$ 4.6, respectively. Notably, a KNN model (k = 7, Chebyshev distance metric) emerged as the most suitable configuration.

The final performance of the optimized models is presented in Table 6.3, while an overview of the selected features resulting from the feature selection process for both male and female subgroups is reported in Table 6.4.

Table 6.3 Classification performance of the optimized models, averaged over 10 iterations employing a 5-fold cross-validation.

	Female		**Male**	
	Validation	**Test**	**Validation**	**Test**
Accuracy	0.83 ± 0.044	0.86 ± 0.090	0.82 ± 0.060	0.84 ± 0.113
Precision	0.84 ± 0.034	0.85 ± 0.082	0.83 ± 0.046	0.88 ± 0.111
F1-score	0.88 ± 0.030	0.91 ± 0.058	0.85 ± 0.048	0.87 ± 0.091
Sens.	0.96 ± 0.034	0.98 ± 0.060	0.91 ± 0.052	0.88 ± 0.133
Specificity	0.60 ± 0.126	0.55 ± 0.267	0.69 ± 0.105	0.77 ± 0.213
AUC	0.92 ± 0.030	0.76 ± 0.140	0.84 ± 0.070	0.82 ± 0.120

Table 6.4 Overview of features selected in the final model

Female	**Male**
12^{th} MFCC; 10^{th} 12^{th} ΔMFCC	3^{rd} Formant; 1^{st} MFCC
5^{th}, 6^{th}, 10^{th}-13^{th} $\Delta\Delta$ MFCC	3^{rd} , 6^{th}, 7^{th} ΔMFCC
4^{th} ΔPLP ; 5^{th}, 8^{th} $\Delta\Delta$ PLP	1^{st} PLP ; 1^{st} Δ PLP
Spectral Flux; Spectral Roll-off point	1^{st} $\Delta\Delta$ PLP

6.2.5 Discussion

The analysis of feature significance, carried out through Pearson correlation coefficients between acoustic features and the class variable, revealed notable associations

between specific vocal features and sleep quality. Among the studied population, the most influential parameters were identified as the MFCC, implying potential alterations in the articulatory capability of subjects following poor sleep quality. Moreover, the general high correlation observed between these acoustic features and the validated clinical scores demonstrates the potential of vocal features in assessing sleep quality.

Analysis of sleep scores obtained from the SLEEPS questionnaires provided valuable insights into factors affecting sleep quality. A significant correlation was identified with COVID-19 positivity. Males with a history of COVID-19 exhibited considerably worse scores on the sPSQI, suggesting a potential impact of the virus on sleep quality. Conversely, individuals with no prior COVID-19 diagnosis displayed sPSQI scores more evenly distributed within the data range. This gender-specific pattern was not observed in the female group. Furthermore, a noteworthy discrepancy emerged between self-reported sleep quality (SLEEPQ) and sPSQI scores among females, with some of them overestimating their sleep quality despite objective measures indicating sub-optmial quality. Interestingly, variables like remote working or the use of blue light filters exhibited no significant correlation with sleep quality. In contrast, variables such as the cessation time of electronic device usage, morning drowsiness, and difficulties falling asleep exhibited moderate correlations, suggesting their potential influence with the application at hand.

Moving to the classification phase, the effectiveness in distinguishing between poor and good sleepers, coupled with the absence of performance degradation when transitioning from validation data to entirely new samples, indicates the feasibility of an automated sleep quality evaluation and robust model generalization. It is worth noting that the models exhibited relatively low classification specificity, primarily attributed to class imbalances in the dataset. However, the good overall performance suggests that this issue may be mitigated when training the algorithm on larger datasets.

6.2.6 Conclusion and Future Works

This study introduced a systematic workflow for SQ classification based on vocal analysis, integrating ML techniques. Vocal recordings were collected from personal computers or smartphones. Despite the absence of professional-grade microphones

and specific task-training the ML models employed demonstrated remarkable efficiency, yielding F-1 scores of 88% for females and 85% for males. This disparity in performance, with higher scores in the female subgroup, may be attributed to the inherent anatomical and physiological differences in the female vocal apparatus, making it potentially more susceptible to vocal alterations induced by sleep disturbances, as also observed in previous research [278].

This study is not without limitations. Firstly, the classification target was identified using the subjective sPSQI score, a clinically validated index that relies solely on self-reported items. Future research endeavors will need to address this limitation by incorporating objective parameters, such as actigraphy-derived measures [282], into the analysis. Additionally, enhancing the model performance might be achieved by expanding the dataset and conducting a stratified analysis based on various factors, including age or specific physiological characteristics (e.g., consistent sleep-wake routines, comorbidities, and other relevant physiological or demographic parameters).

Chapter 7

Application III: Alcohol Intoxication

7.1 Alcohol Intoxication

7.1.1 Statistics

Alcohol intoxication refers to a physiological and behavioral state resulting from the excessive consumption of alcoholic substances. This condition is recognized as a significant risk factor contributing to mortality and disability, as indicated in studies by Muller [78] and the WHO reports [283]. Additionally, alcohol consumption is associated with an increased risk of drowning and injuries resulting from violence, falls, and motor vehicle accidents, as documented by Taylor [284, 285] and Cherpitel [286].

Various sociodemographic factors, including economic wealth and cultural habits, exert a substantial influence on alcohol use. The highest percentages of alcohol consumption are found in high-income countries in Western Europe, North and South America, Australia, and New Zealand, as reported by the WHO [283]. Gender also plays a significant role, with women more frequently being abstainers than men and consuming less alcohol when they do drink.

As for age, the WHO report emphasizes that the drinking habits of young individuals often mirror the overall population. There are lower prevalence rates of current drinking among individuals (15-19 years). However, by the age of 20-24

years, young people often exhibit equal or higher alcohol consumption than the non-stratified population [287].

A notable indicator of alcohol consumption patterns is heavy episodic drinking (HED), defined as the consumption of 60 or more grams of pure alcohol on a single occasion at least once a month. This phenomenon is more prevalent in parts of Eastern Europe and some sub-Saharan countries. Moreover, the prevalence of HED is lower among adolescents aged 15-19 years compared to the overall population, but it peaks at the age of 20-24 years. Current epidemiological data show that individuals in this age range typically engage in heavy drinking sessions [287].

7.1.2 Pathophysiology

Upon the consumption of alcoholic beverages, the ethanol molecules they contain are absorbed via the gastrointestinal tract. Less than 10% of this ethanol is excreted by the kidneys, expelled through the lungs, or the skin, ultimately being identified in urine, exhaled air, and sweat, as discussed in studies by Jones [288] and Hyun [289]. The remaining 90-95% circulates throughout the body and eventually reaches the liver via the portal vein. The liver, with its high levels of alcohol-metabolizing enzymes, plays a central role in alcohol metabolism, involving a series of reactions primarily mediated by alcohol dehydrogenase (ADH) and cytochrome P450 2E1 (CYP2E1). The primary pathway in ethanol metabolism is the oxidative process in which ADH converts alcohol to acetaldehyde. Subsequently, aldehyde dehydrogenase (ALDH) promptly converts acetaldehyde into acetate, which is further metabolized in peripheral tissues into carbon dioxide (CO_2), fatty acids (FAs), and water (H_2O), as outlined in the work of Hyun [289] and Jones [288].

The mechanism involved in ethanol metabolism indicates a dose-dependent pharmacokinetics. This is due to the saturation of the hepatic ADH enzyme with substrate at Blood Alcohol Concentration (BAC) levels above 15–20 mg/100 mL. Consequently, for lower rates of alcohol intake, ethanol absorption follows first-order kinetics, where a constant proportion relative to the original concentration is eliminated from the body over time. However, when BAC exceeds the threshold (15–20 mg%), the BAC decreases at a constant rate over time, displaying zero-order kinetics [288].

7.1.3 Symptoms

Alcohol consumption exerts intricate effects on the human body, involving complex interactions within various physiological systems. These effects can manifest as both acute and long-term consequences.

The causal relationship between alcohol consumption and liver diseases is well-established. Given the crucial role of the liver in ethanol metabolism, chronic alcohol consumption can result in conditions such as steatosis (fatty liver), alcoholic hepatitis, fibrosis, and ultimately cirrhosis, as detailed by Hyun [289] and WHO reports [287]. Alcohol impact also extends to the cardiovascular system. Acute alcohol consumption can lead to peripheral vasodilation, temporarily reducing blood pressure. Moreover, numerous epidemiological studies have linked chronic and excessive alcohol intake to conditions like hypertension, cardiomyopathy, arrhythmias, and an increased risk of stroke, as reported by the WHO [287].

Furthermore, the WHO report highlights a causal connection between alcohol use and the development of cancer in various regions of the body, including the oropharynx, larynx, esophagus, liver, colon, rectum, and the female breast. Other symptoms associated with alcohol consumption may encompass increased urine production, dehydration, and irritation of the gastrointestinal tract, potentially leading to gastritis, ulcers, and gastrointestinal bleeding. Prolonged and excessive alcohol consumption can also result in neurotoxic effects, including neuronal damage and an elevated susceptibility to neurodegenerative disorders [287].

In addition to the cardinal effects primarily associated with ethanol metabolism, alcohol consumption can lead to intoxication, characterized by a loss of control over actions or behavior due to the impact on various neural pathways and brain regions, as explained in the WHO reports [287] and in Paprocki's research [290]. The process of ethanol-induced intoxication can be divided into three main stages: initial absorption, peak intoxication, and the decay stage. During initial absorption, BAC rapidly increases after ethanol ingestion, with peak intoxication typically occurring within 30 to 60 minutes. Subsequently, peak BAC levels begin to decline, and intoxication gradually subsides, with BAC returning to zero within six to eight hours of initial consumption [290].

Recent evidence consistently suggests that when individuals consume small doses of ethanol, resulting in blood BAC levels ranging from 30 to 50 mg/dL, they

commonly experience increased talkativeness, reduced inhibitions, and a mild sense of euphoria. However, as BAC levels rise further, entering the range 80-120 mg/dL, noticeable signs of impairment become more pronounced, encompassing slower reaction times, challenges in processing information, impaired motor coordination, slurred speech and an unsteady gait.

It is worth noting that the effects of alcohol can vary based on individual factors such as metabolism, genetic predisposition, overall health, gender, and patterns of consumption. Understanding the intricate mechanisms underlying alcohol effects on the body is vital for the development of preventive measures, interventions, and treatment strategies aimed at mitigating the negative health consequences associated with alcohol consumption, as highlighted in [287, 288, 290].

7.1.4 Assessment Criteria and Complicating Factors

BAC serves as a widely used metric for assessing the degree of alcohol intoxication, typically expressed as a percentage that represents the ratio of alcohol to blood volume. BAC is influenced by various factors, including the quantity and speed of alcohol consumption, an individual body weight, metabolism, and the presence of food in the stomach [288, 290].

In several scenarios, such as the need to evaluate alcohol impairment in individuals suspected of driving under the influence or other alcohol-related offenses, quicker procedures are essential. In such contexts, Breath alcohol tests (BrAC), commonly referred to as *breathalyzer tests*, are frequently employed to estimate an individual BAC by analyzing the alcohol content in their exhaled breath [291]. These tests leverage the quantifiable dose-effect relationship between an individual BAC and the resulting effects. They operate on the principle of measuring the quantity of alcohol vapor present in exhaled breath, which correlates with the alcohol concentration in the bloodstream.

It is worth noting that breath alcohol tests provide an estimation of BAC and are subject to certain limitations. Factors such as the calibration and accuracy of the device and potential interfering substances (e.g., mouthwash or specific medications) can affect the accuracy of the results. Consequently, for legal purposes, confirmatory testing employing more precise methods such as blood tests may be necessary [290].

Recent advancements have introduced techniques for estimating the level of intoxication based on changes associated with common symptoms, offering the advantage of truly measuring the effects of intoxication. Among these emerging techniques, some propose algorithms relying on the photoplethysmography signal [290], while others rely on in-vehicle cameras focused on the driver face, as seen in the case of Volvo [290].

7.2 Automated Vocal Analysis for Alcohol Intoxication Assessment

Detecting a person state concerning possible alcohol inebriation holds paramount importance for social safety and prevention. Inebriation is associated with a range of effects, including reduced attention, increased drowsiness, impaired balance and coordination, and slowed or slurred speech. These impairments pose substantial risks, especially in contexts like driving, as outlined in the work of Davies [292].

According to the WHO Global status report on road safety in 2018 [293], alcohol is a contributing factor in a significant percentage of road traffic fatalities worldwide, ranging from 5% to 35%. In an effort to reduce traffic-related fatalities, EU Regulation 2019/2144 mandates that newer vehicles, as of July 6, 2022, must be equipped with a device that assesses the driver alcohol consumption and prevents the vehicle from starting. [294].

Within this context, this section delves into the impact of alcohol consumption on the process of speech production and assesses the feasibility of using speech-based models to identify intoxication. In more detail, a comprehensive series of statistical analyses was conducted to uncover significant patterns in acoustic features that manifest after alcohol consumption. Furthermore, a specific investigation was carried out to examine how various speaker characteristics, such as gender, age, and drinking habits, may contribute to distinct trends.

7.2.1 Related Literature

Vocal analysis, particularly when coupled with AI and ML techniques, emerges as a promising solution for the preliminary identification of an individual inebriation

condition, given that altered speech production is one of the most important changes observable after excessive alcohol consumption.

The 2011 INTERSPEECH Challenge [295] was primarily concerned with the identification of intoxication or inebriation through speech analysis. The study made use of a dataset called the Alcohol Language Corpus (ALC), which encompassed audio recordings from 162 individuals under both sober and alcohol-influenced conditions. In general, the results of this research indicated an increase in F0 in subjects exhibiting signs of intoxication (with a blood alcohol concentration of greater than or equal to 0.5%), with this effect being more pronounced in female subjects [296].

However, an extensive review of the literature surrounding ALC-based studies, including works by [296–309] unveiled a lack of consensus with respect to the significance of other acoustic features or specific speech dimensions in the context of intoxication detection.

As of the present, the highest documented performance in this particular domain utilizing the ALC dataset was reported in Bone et al.'s research [298]. In their study, a comprehensive model was introduced to enhance the task of identifying intoxication using speech data. This approach encompassed several techniques, such as the implementation of hierarchical acoustic features, which were extracted at varying temporal scales to capture moment-to-moment speech changes. Additionally, they utilized iterative speaker normalization, a method aimed at repetitively normalizing features to reduce inter-speaker variations while preserving discriminative information. As a result, the model they developed effectively discriminated between sober and inebriated subjects (with a BAC of 0.5% or higher). On the test set, this model achieved a classification accuracy of 70.54%, thereby establishing a benchmark for performance in this specific research domain.

7.2.2 Materials

The corpus utilized in the experiment is the ALC, which is accessible through the Bavarian Archive for Speech Signals as mentioned in [310]. This dataset comprises speech recordings of individuals in both sober and intoxicated states, encompassing a diverse range of speaking styles. The type of vocal tasks performed include simple

Table 7.1 Demographics of participants included in the study N: Numerosity.

	N subjects	Age (years)	BMI
Tot	162	30.99 ± 9.49	22.82 ± 2.84
Male	55	29.05 ± 0.51	21.57 ± 2.75
Female	77	32.75 ± 9.88	23.95 ± 2.41

digit strings, sentence repetition, tongue twisters, specific prompts for application commands, monologues, and natural conversational speech.

The ALC dataset encompasses speech samples from a total of 162 German speakers, with ages ranging from 21 to 64 years. Extensive metadata for each individual, including pertinent personal information, has been meticulously documented. Additionally, comprehensive data related to each recording session has been collected, covering factors such as the presence of external noise and the emotional state of the subjects. Detailed demographics of the participants involved in the study are provided in Table 7.1.

Additional participant characteristics that could potentially influence pronunciation were also documented, including their educational level, profession, and region of origin. Detailed information on the drinking habits of the participants was collected. During the interviews, participants were required to describe the frequency and quantity of their alcohol consumption. They were specifically asked about how often they typically consumed alcohol (e.g., daily, more than once a week, once a week, or less than once a week) and the quantity consumed in a single session (measured in units, such as glasses of wine or beers). Based on their responses, the researchers categorized both the amount of alcohol intake (categorized as sparse: 1-2 units of beer or wine, or plenty: more than 2 units) and the frequency of consumption (classified as seldom: once a week or less, or often: more than once a week or daily). Subsequently, participants' drinking habits were classified as light (sparse and seldom), moderate (sparse and often, or plenty and seldom), or heavy (often and plenty). The results indicated that participants primarily reported light to moderate alcohol consumption habits, with a preponderance of females in the light category and males in the moderate one. Less than 10 subjects were eventually categorized as heavy alcohol consumers, with only one female subject among them.

As for the data collection procedure, two distinct microphones were employed for all recordings: a Beyerdynamic Opus54.16/3 headset microphone and an AKG

Q400 mouse microphone positioned in the middle of the vehicle front ceiling. The recordings were sampled at a rate of 44.1 kHz, utilizing a 16-bit PCM format. In this specific study presented in this dissertation, an exclusive use of vocal samples recorded with head-mounted microphones was employed. This approach aimed to maintain a consistent and high-quality signal, thereby avoiding the introduction of complexities associated with diminished signal quality. This choice aligns with the practices of the majority of studies involving the ALC corpus, as documented in works such as [296–309]. These studies also reported using samples acquired through high-quality equipment, ensuring a fair and meaningful basis for comparison between this work and similar investigations.

To investigate potential influences from the recording environment, samples were either performed in a Volkswagen Passat Variant Diesel 134PS 2004 or in an Opel Astra (GM) Astra Coupe 22 AUT 2001, which had varying interior sizes. For safety reasons, the vehicle engine was turned off for approximately two-thirds of the speech exercises during recording sessions. Notably, no recordings occurred while the vehicle was in motion. Each test, conducted in both sober and intoxicated conditions, was controlled by an operator who also assumed the role of the conversational partner during the dialogues.

Before the test, each participant chose their target BAC for the intoxication test. Subsequently, the necessary quantity of alcohol was determined using the Widmark formula, denoted as in Equation 7.1.

$$c = \frac{V}{mr} \tag{7.1}$$

where c represents the alcohol concentration, V stands for the quantity of alcohol in grams (g), m refers to the individual body mass in kilograms (kg), and r is the reduction factor that accounts for the specific characteristics of the subjects. The calculation of this reduction factor is based on the Watson formula, as indicated by Equation 7.2.

$$r = \frac{\rho_b g}{fm} \tag{7.2}$$

In this context, $\rho_b = 1.055 \frac{g}{cm^3}$ represents the density of blood, and f is the fraction of water in blood, with a constant value of 0.8. The body water content g can

be determined according to the formulas provided in 7.3 and 7.4, where t denotes the age in years and h signifies the body height in centimeters (cm), as specified in [310].

$$g_{male} = 2.447 - 0.09516t + 0.1074h + 0.3362m \tag{7.3}$$

$$g_{female} = 0.203 - 0.07t + 0.1069h + 0.2466m \tag{7.4}$$

Following the determination of the required alcohol intake, all participants underwent BAC, BRAC, and speech recordings. BAC measurements were conducted using Head-Space Gas Chromatography, while BRAC was computed through the use of the Drager Alcotest 7410 and Envitec Alcotest devices. Notably, comparable BAC and BRAC distributions were achieved for both female (BAC: 0.086 ± 0.029, BRAC: 0.086 ± 0.032) and male (BAC: 0.091 ± 0.028, BRAC: 0.091 ± 0.030) participants.

Speech recordings were conducted immediately after the alcohol assessments, involving a set of 30 speech exercises. Subsequently, each participant was required to undergo a second recording session in a sober state, which included an expanded set of 60 speech exercises, after a minimum of two weeks had passed.

All participants voluntarily took part in an intoxication test supervised by staff from the Institute of Legal Medicine. Each participant who contributed to the ALC dataset provided their informed consent, documented in a legal consent form. This form granted permission for the scientific and technical use of their recorded speech, with the assurance that the contents of the corpus would not be linked to their personal data. The list of speech exercises performed in both the A (intoxicated) and NA (sober) states, along with their corresponding English translations, can be found in Table 7.2.

Table 7.2 List of tasks in the ALC dataset along with corresponding Englis translations

ID NA	ID A	Prompt (IPA Translation)	Translation to English
1	1	Bitte lesen Sie die Telefonnummer: +491763582901	Please read the phone number: +491763582901

Table 7.2 continued

8	8	Bitte lesen Sie die Adresse: ʃpˈɔrtplatsv,eːk ʌʏ, mˈarktɣraɪts	Please read the address: Sportplatzweg 27, Marktgraitz
12	12	Bitte lesen: bˈakɜ, mˈuːzə(ə)lm,an, mˈɛnʃə(ə)n, mˈasə(ə)n, mˈɔrdɜ, mˈoːrə(ə)n, mˈʊtɜ, mˈɑ(ɑ)nuːm,ɛntə(ə)nm,axɜ	Please read: Baker, Muselmann, People, Masses, Murderers, Moors, Mothers, Manumentmakers
13	13	Bitte lesen Sie die Adresse: 18546 zˈasnɪts	Please read the address: 18546 Sassnitz
15	15	Bitte lesen: 06271 57390	Please read: 06271 57390
19	19	Bitte Adresse lesen: mˈ adap,akaːb ə(ə) tˈeːɣɪnd,ɪsʃtrˈafl ə(ə) 77b	Please read address: Madapaka-Betegindis-Stra e 77b
20	20	Bitte so schnell wie möglich lesen: kˈɛtsɜ kʁˈyçtsɜp,ɛtstə(ə)n jˈɛtst klˈɛːklɪç, lˈɛtstlɪç plˈœ(œ)tslɪç lˈaɪçt skˈɛptɪʃ	Please read as soon as possible: Heretics are now complaining, in the end suddenly slightly skeptical.
23	7	Bitte so schnell wie möglich lesen: Bemoost wächst nächst dem Strom ein Stamm, feststämmig stolz strebt sein Geäst stromwärts, und weist nach Ost und West.	Please read as fast as possible: Mossy grows next to the stream a trunk, firmly proud strives for its current, and points to east and west.
24	17	Bitte lesen Sie die Adresse: bə(ə)mˈoːst vˈɛkst nˈɛçst dˈeːm ʃtʁˈoːm ˈaɪn ʃtˈam, fˈɛstʃt,ɛmɪç ʃtˈɔlts ʃtʁˈeːpt zˈaɪn ɣə(ə)ˈɛst ʃtʁˈɔmvɛʁts, ˈʊnt vˈaɪst nˈɑ(ɑ)ːx ˈɔst ˈʊnt vˈɛst.	Please read the address: Sister-Hermenegildis-Street
26	6	Bitte Telefonnummer lesen: 0862359286	Please read the phone number: 0862359286

Table 7.2 continued

29	9	Bitte lesen Sie die Kreditkartennummer: 1390 7516 0281 9357	Please read the credit card number: 1390 7516 0281 9357
31	11	Bitte lesen Sie das Autokennzeichen: STA-PB 2759	Please read the license plate: STA-PB 2759
32	3	Bitte so schnell wie möglich lesen: mˈɛsvɛksə(ə)l, vˈaxsmaskə(ə), vˈaxsmaskə(ə), mˈɛsvɛksə(ə)l	Please read as fast as possible: Measurement change, Wax mask, Wax mask, Measurement change.
36	16	Bitte so schnell wie möglich lesen: dˈiː kˈœ(œ)çm mˈɪt dˈeːm tˈʊpfə(ə)nk,ɔpftʊx kˈɔxt kˈaɾpfə(ə)n ˈɪn dˈeːm kˈʊpfɜk,ɔxtɔpf	Please read as fast as possible: The kitchen with the polka dot headscarf cooks carp in the copper saucepan.
41	21	Bitte Steuerbefehl lesen: tˌɛmpeːratˈuːɾ 23 °C	Please read the control command: Temperature 23°C
50	29	Bitte Steuerbefehl lesen: ˈaʊtoːbˌɑ(ɑ)ːnə(ə)n mˈaɪdə(ə)n	Please read the control command: Avoid motorways
51	24	Bitte Steuerbefehl lesen: freːkvˈɛnts 92,2 MHz	Please read the control command: Frequency 92.2 MHz
59	23	Bitte Steuerbefehl lesen: nˈɛçstɜ tˈiːtə(ə)l	Please read the control order: next title
60	30	Bitte buchstabieren: M A R K T G R A I T Z	Please spell: M A R K T G R A I T Z
10	10	Sprechen Sie mit dem Versuchsleiter über das Bild - tat_3GF	Talk to the investigator about the picture - tat_3GF
30	5	Erzählen Sie eine Geschichte zum Bild - tat_13MF	Tell a story about the picture - tat_13MF
34	2	Bitte sprechen Sie mit dem Versuchsleiter: Erzählen Sie von einem Ihrer Urlaube.	Please talk to the investigator: Tell us about one of your vacations.

Table 7.2 continued

42	22	Bitte stellen Sie sich vor, Sie wollten den Radiosender auf Speicherplatz FM3 einstellen. Geben Sie dem Auto den Befehl, dies zu tun.	Please imagine that you want to tune the radio station to storage space FM3. Give the car the command to do this.
46	25	Bitte stellen Sie sich vor, sie wollen ihre Sitzheizung auf Stufe 2 schalten. Geben Sie dem Auto den Befehl, dies zu tun.	Please imagine that you want to switch your seat heating to level 2. Give the car the command to do this.
48	27	Sie wollen Ihre Klimaanlage anschalten. Geben Sie Ihrem Auto den Befehl, dies zu tun.	You want to turn on your air conditioning. Give your car the command to do this.
49	28	Sie möchten zum Hilton Nürnberg und dafür Ihr Navigationssystem benutzen. Geben Sie Ihrem Auto den entsprechenden Befehl.	You want to go to the Hilton Nürnberg and use your navigation system. Give your car the appropriate command.
55	26	Sie wollen den 9. Titel auf der 6. CD Ihres CD-Wechslers hören. Geben Sie Ihrem Auto den Befehl dazu.	You want to listen to the 9th track on the 6th CD of your CD changer. Give your car the command to do so.
14	14	Bitte beantworten Sie folgende Frage: Was war bisher das schönst Geschenk, das Sie bekommen haben und warum hat es Ihnen so gefallen?	Please answer the following question: What has been the best gift you have received so far and why did you like it so much?
38	18	Erzählen Sie eine Geschichte zum Bild - tat_12M	Tell a story about the picture - tat_12M
4*	4*	Bitte lesen Sie die Adresse: Laurentiusbergstraße 27, Tauberbischofsheim	Please read the address: Laurentiusbergstraße 27, Tauberbischofsheim

Table 7.2 continued

2	Bitte sprechen Sie mit dem Versuchsleiter: Was halten Sie von Weihnachten?	Please talk to the experimenter: What do you think of Christmas?
3	Bitte so schnell wie möglich lesen: Fischers frisch frisierter Fritze frisst frisch frittierte Frisch-Fisch-Frikadellen.	Please read as soon as possible: Fischer's freshly coiffed Fritze eats freshly.
5	Erzählen Sie eine Geschichte zum Bild - tat_18GF	Tell a story about the picture - tat_18GF
6	Bitte Telefonnummer lesen: 073952863491	Please read the phone number: 073952863491
7	Bitte so schnell wie möglich lesen:zˈɔnst nˈɪstə(ə)n ʃtˈɑ(ɑ)ːrə(ə) ʃtˈeːts ˈɪm ʃtˈam, dˈɔx ʃtˈ??m tsɛɾʃtˈœ(œ)ɾtə(ə) ˈast ˈʊm ˈast, dˈas lˈɛŋst tsɛɾʃtˈoːp dˈas	Please read as soon as possible: Otherwise starlings always nest in the trunk, but a storm burst branch after branch, which long ago destroyed the starling nest.
9	Bitte lesen Sie die Kreditkartennummer: 1835 0117 2839 9602	Please read the credit card number: 1835 0117 2839 9602
11	Bitte lesen Sie das Autokennzeichen: BGL-KP 397	Please read the license plate: BGL-KP 397
16	Bitte so schnell wie möglich lesen: kˈalə(ə) kˈɑ(ɑ)ːleːk,atsə(ə)nɣl,atsə(ə)nkɾ,atsə kɾˈatst kˈɑ(ɑ)ːlə(ə) kˈatsə(ə)nɣl,atsə(ə)n.	Please read as soon as possible: Kalle bald cat bald scratcher scratches bald cat bald heads.
17	Bitte Adresse lesen: ɣrˈuːɪsˈʃˈeːs ɣˈɛsçə(ə)n 5	Please read the address: Gruis'sches Gässchen 5
18	Erzählen Sie eine Geschichte zum Bild - tat_5	Tell a story about the picture - tat_5
21	Bitte lesen Sie die Telefonnummer: +491623792048	Please read the phone number: +491623792048

Table 7.2 continued

22	Bitte sprechen Sie mit dem Versuchsleiter: Sprechen Sie über den Trinkversuch, an dem Sie vor ein paar Wochen teilgenommen haben.	Please talk to the investigator: Talk about the drinking experiment you took part in a few weeks ago.
25	Sprechen Sie mit dem Versuchsleiter über das Bild - tat_19	Talk to the investigator about the picture - tat_19
27	Bitte so schnell wie möglich lesen: dˈeːn vˈɪntsɪɣə(ə)n tsvˈɛɾk tsˈɪmpə(ə)lp,ʊm tsvˈɪkt zˈaɪmə(ə) tsˈɪpfə(ə)lm,ytsə(ə). ˈɛɾ tsˈʊpft, ˈɛɾ tsˈiːt, ˈʊnt tsˈɛɾt tsuːlˈɛtst fˈɔl tsˈɔɾn zˈiː ˈɪn dˈiː pfˈytsə(ə).	Please read as soon as possible: The tiny dwarf Zimpelpum pinches his pointed cap. He plucks, he pulls, and finally, full of anger, drags her into the puddle.
28	Bitte lesen Sie die Adresse: 25557 hˈɑ(ɑ)neːr,aʊ-hˈɑ(ɑ)deːm,aɾʃə(ə)n	Please read the address: 25557 Hanerau-Hademarschen
33	Bitte lesen Sie die Adresse: ʃvˈɑ(ɑ)rə(ə)nbɛrɣstr,ɑ(ɑ)ːsə(ə) 2, Stuttgart	Please read the address: Schwarenbergstraße 2, Stuttgart
35	Bitte lesen: 81637 05249	Please read: 81637 05249
37	Bitte Adresse lesen: lˈiːsl-kˈaɾlstat-ʃtrˈɑ(ɑ)ːsə(ə)1a	Please read address: Liesl-Karlstadt-Straße 1a
39	Bitte Adresse lesen: tˈɪpə(ə)n-tˈapə(ə)n-tˈøːnçə(ə)n, v,ʊpɜtˈɑ(ɑ)ːl	Please read address: Tippen-Tappen-Tönchen, Wuppertal

Table 7.2 continued

40	Bitte so schnell wie möglich lesen: tsvˈɪʃə(ə)n tsvˈaɪ tsvɛtʃɣˈɛntsvaɪɣə(ə)n zˈɪtsə(ə)n tsvˈaɪ tsˈɛçə(ə)nʃv,aɾtsə(ə) tsˈɛçɪʃ tsvˈɪtʃɛrndə(ə) tsvˈɛrkʃvalbə(ə)n	Please read as soon as possible: Between two plum branches sit two black Czech chirping dwarf swallows.
43	Stellen Sie sich vor, sie wollen mit Hilfe Ihres Navigationssystems zum Tierpark Hellabrunn. Geben Sie Ihrem Auto den Befehl dazu.	Imagine you want to go to Hellabrunn Zoo with the help of your navigation system. Give your car the command to do so.
44	Bitte Steuerbefehl lesen: freːkvˈɛnts 87,7 MHz	Please read the control command: Frequency 87.7 MHz
45	Sie hören gerade Radio und wollen auf CD wechseln. Geben Sie Ihrem Auto den Befehl, dies zu tun.	You are listening to the radio and want to switch to CD. Give your car the command to do this.
47	Sie wollen Ihr Tempomat auf 120 km/h einstellen. Geben Sie Ihrem Auto den entsprechenden Befehl.	You want to set your cruise control to 120 km/h. Give your car the appropriate command.
52	Bitte Steuerbefehl lesen: Sender als Preset speichern.	Please read the control command: Save the transmitter as a preset.
53	Sie wollen den Repeat-Modus Ihres CD-Spielers aktivieren. Geben Sie Ihrem Auto den Befehl dazu.	You want to activate the repeat mode of your CD player. Give your car the command to do so.
54	Sie wollen ihren Kollegen Herbert Schuster anrufen. Geben Sie Ihrem Auto den Befehl dazu.	You want to call your colleague Herbert Schuster. Give your car the command to do so.

Table 7.2 continued

56	Bitte Steuerbefehl lesen: t,ɛmpoːmˈɑ(ɑ)ːt d,eːaktiːvˈiːrə(ə)n	Please read the control command: Deactivate cruise control
57	Bitte Steuerbefehl lesen: kˈartə(ə) ˈɪm nˈɔrdrɪçt,ʊŋ	Please read the control command: Map in northerly direction
58	Bitte Steuerbefehl lesen: r,uːtə(ə)nɔptsjˈoːnə(ə)n ˈɛndɜn	Please read the control command: Change route options

7.2.3 Methods

Pre-processing and Feature Extraction

An acoustic examination of the speech samples within the ALC dataset revealed the presence of initial and final silent regions in the recordings. Consequently, a customized algorithm, based on the Praat Voice Activity Detector (VAT), was implemented to eliminate non-relevant regions. Additionally, signal amplitudes were normalized in the range [0,1] to mitigate the impact of speaker-to-microphone distance.

Following the pre-processing procedures, a comprehensive set of features was extracted from each speech signal using the OpenSmile toolkit and using the ComParE2016 feature set with default hyperparameters. Furthermore, to delve into the rhythmic aspects of speech, additional 12 parameters were computed using Praat-based algorithms implemented in Python through the Parselmouth library. These additional parameters encompassed DPI, Max Duration of pauses, Duration of Unvoiced regions, Max Duration of Voiced regions, Percentage of pause intervals, Percentage of unvoiced regions, Percentage of voiced regions, RST, Total Energy, Total Duration of Voiced frames, and Mean absolute pitch slope. In total, the feature set comprised 6385 distinct features. To ensure consistency and comparability, Z-score standardization was applied to all features, thereby scaling them to a common range.

Experiment 1

In the first experiment, the primary objective was to evaluate the influence of alcohol consumption on speech samples and investigate how various speaker characteristics, including age, gender, and drinking habits, affected acoustic features. Additionally, one of the main aims was to explore the feasibility of binary classification through the automatic analysis of specific vocal samples. To streamline the analysis and maintain a focus on the intended scope, the first experiment was conducted on a single speech task, thus reducing the complexity of the investigation. Results from prior studies suggested improved performance with tongue-twisters, as mentioned in [307], hence this task was selected for analysis:

Kalle Kahlekatzenglatzen kratzer kratzt kahle Katzenglatzen.

A differential examination was conducted to identify potential trends in features that exhibited similar changes across different subjects when transitioning from the NA to the A status. This analysis involved calculating the percentage variation of each feature between the two states and determining whether there was an increase or decrease by applying the sign function. The signs of each feature were summed across all subjects, and the absolute value was computed. A higher absolute value of this sum indicated a more systematic variation in the feature. In other words, a feature that consistently increased or decreased after inebriation would yield higher values because there were numerous positive results of the sign function applied to the percentage variation. The metric derived from this process was referred to as Absolute Systematic Variation (ASV), as defined in Equation 7.5 and it is expressed as a percentage of the total cardinality of the subgroup investigated.

$$ASV_i = \left| \sum_{j=1}^{N} \operatorname{sgn}\left(\frac{f_{ij}(y_1) - f_{ij}(y_0)}{f_{ij}(y_0)} \right) \right| \tag{7.5}$$

In the equation, i identifies each feature and j identifies each instance, for a total of N instances; f_{ij} (y_1) represents thus the i-th feature computed for the j-th subject in the status/class 1 (in our case A), and f_{ij} (y_0) is the same feature for the status/class 0 (NA).

A comprehensive analysis was conducted by repeating the procedure with different stratifications. This approach allowed for the evaluation of the impact of external factors on feature trends. These stratifications were established based on various

covariates, with the condition that only groups comprising at least 28 subjects were considered for analysis, ensuring the attainment of statistically significant results.

- Gender: female (F, 77 subjects) and male (M, 85 subjects);
- Age:18-40 years old range (139 subjects);
- BMI: normo-weight (116 subjects) and overweight to moderate obese (37 subjects) ranges (Table 5.1).
- BRAC: 0.5-0.79 (light: 162 NA, 49 A) and 0.8+ (moderate: 162 NA, 92 A).

Classification was conducted on the complete tongue-twister dataset, involving a 75-25 split between the training and test sets. Care was taken to ensure that different individuals were present in each set to avoid data leakage. Additionally, a statistical feature selection was implements through K-Best ANOVA, setting the value of K to the square root of the number of data instances, which in our case amounted to 162 x 0.75 x 2, resulting in K = 16.

Three distinct classifiers, namely SVM, GB, and RF were testes. All statistical and ML analyses were conducted using Python with the scikit-learn and scipy libraries, employing default hyperparameters or sub-models.

Experiment 2

In this second phase of the study, the goal was to create a lightweight model capable of distinguishing between sober and intoxicated speakers, regardless of the specific speech tasks and individual characteristics. To achieve this objective, a customized classification model was utilized, based on the Discriminative Adversarial Neural Network (DANN) framework. Figure 7.1 provides a representation of this architecture.

DANNs are commonly employed to acquire a feature representation that is domain-independent [311]. A typical DANN architecture consists of three key components: an Encoder block, responsible for extracting relevant information from input arrays; a Task block, used to classify instances into the desired categories (e.g., A or NA); and a Discriminator block, tasked with predicting the domains of instances. In line with the approach [312], the adversarial block was customized to perform

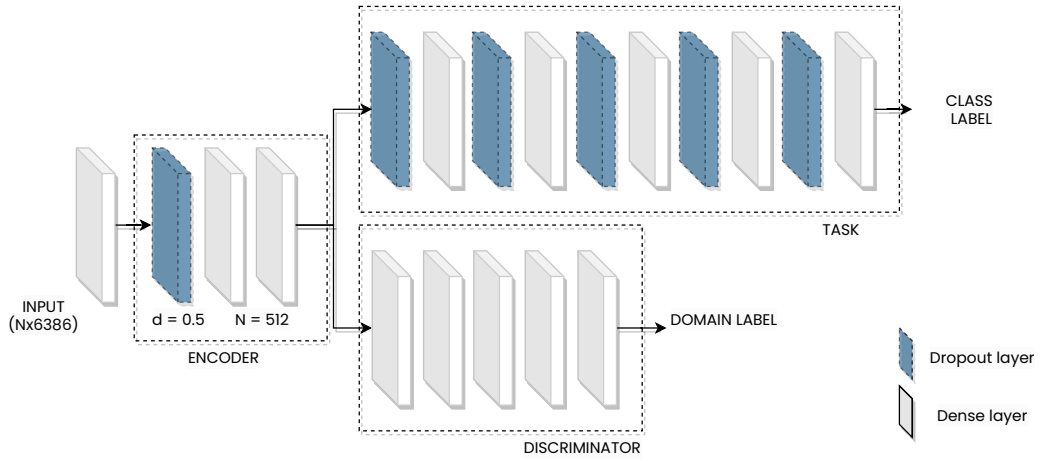


Fig. 7.1 Schematic architecture of a Discriminative Adversarial Neural Network

subject ID classification, effectively reducing inter-speaker feature variability within the overall model.

In this architecture, the Encoder plays a pivotal role in feature extraction, facilitating the learning of domain-invariant features. During the training phase, the network encounters samples from various speakers, and through a gradient reversal-based technique, the model weights are adjusted to enhance its ability to distinguish between the two relevant categories while reducing its capacity to differentiate between different speakers, as described in [311]. To achieve this goal, the encoder loss is computed at each iteration, as shown in Equation 7.6:

$$loss_{Enc} = loss_{Task} - \lambda * loss_{Disc} \tag{7.6}$$

where λ is a stabilizing factor that is usually used to minimize unwanted oscillations due to complex domains classifications.

For this work, a customized model based on the Adapt Python library [313] was employed. To prevent potential issues related to the model generalization capability, the training process incorporated 75% of the original dataset, with performance evaluation carried out on the remaining 25%, without further optimization. To ensure speaker independence, the splitting procedure was applied based on the subjects' IDs, thereby including recordings from the same speakers in either the training or test sets. Additionally, to minimize the influence of external factors such as age, gender, BMI,

and alcohol consumption habits, a stratified train-test split was conducted, resulting in balanced subsets.

It is essential to acknowledge that, due to the inherent data collection process, the ALC dataset displayed an imbalanced distribution between sober (60 tasks per subject) and intoxicated speakers (30 tasks per subject). As a result, following the approach of similar studies [297–300, 296, 301–309], it was chosen to prioritize balanced accuracy over the traditional accuracy metric to evaluate the performance of the final classification models. Balanced accuracy, calculated as the average of sensitivity and specificity, offers a fair assessment in scenarios of class imbalance.

Regarding the model architecture and its hyperparameters, an empirical trial-and-error approach was employed to identify the optimal parameter set. The balanced accuracy on a validation set, which constituted 10% of the original training set, served as the guiding metric for assessing the model ability to predict new, unseen samples. The determination of the optimal number of iterations and the need for early stopping were made through a careful comparison of training and validation curves.

7.2.4 Results

Experiment 1

In Table 7.3, the top 10 features with the highest ASV values are presented for the entire dataset. The table also includes the ASV values for these features across various stratifications. These results demonstrate the consistent increase or decrease in specific features when comparing subjects in their NA and A states.

Table 7.4 provides an overview of the model performance on the test set for binary classification, distinguishing between sober and intoxicated speakers using tongue-twister speech samples. The table includes metrics such as balanced accuracy, precision, F1 score, sensitivity, specificity, and AUC to facilitate a comprehensive comparison among the various models that were tested. All classifiers were trained on a subset of the original set of features derived from the applied feature selection process. This subset encompasses the following features in OpenSmile nomenclature:

- $audspecRasta\_lengthL1norm\_sma\_upleveltime25$

- *audspecRasta_lengthL*1*norm_sma_upleveltime*50
- *mfcc_sma*[12]*_minSegLen*
- *pcm_fftMag_spectralRollOff*50.0*_sma_de_minSegLen*
- *pcm_fftMag_spectralRollOff*75.0*_sma_de_minSegLen*
- *pcm_fftMag_spectralCentroid_sma_de_minSegLen*
- *pcm_fftMag_spectralVariance_sma_de_minSegLen*
- *audspecRasta_lengthL*1*norm_sma_linregc*
- *pcm_RMSenergy_sma_meanPeakDist*
- *audSpec_Rfilt_sma*[7]*_linregc*2
- *audSpec_Rfilt_sma*[10]*_peakMeanRel*
- *audspec_lengthL*1*norm_sma_de_meanPeakDist*
- *audSpec_Rfilt_sma_de*[6]*_flatness*
- *audSpec_Rfilt_sma_de*[7]*_flatnes*
- *audSpec_Rfilt_sma_de*[8]*_flatness*

Table 7.3 Identification of the top 10 features in terms of non-stratified ASV, alongside the ASV for each stratification

Feature Name	Stratification							
	None	Female subjects	Male subjects	BMI Normo-weight	BMI Over-weight	Age 18-40	BRAC Light	BRAC Moderate
audSpec_Rfilt_sma[13]_qregc1	70.37	61.04	78.82	68.70	78.95	69.78	83.67	58.70
mfcc_sma_de[2]_quartile2	69.14	66.23	71.76	70.43	73.68	68.35	59.18	71.74
mfcc_sma_de[7]_lpc2	64.20	68.83	60.00	60.00	68.42	62.59	71.43	63.04
jitterDDP_sma_qregc1	62.96	55.84	69.41	60.00	68.42	61.15	55.10	65.22
audSpec_Rfilt_sma[11]_qregc1	61.73	63.64	60.00	60.00	68.42	66.91	55.10	67.39
jitterLocal_sma_qregc2	61.73	42.86	78.82	65.22	52.63	61.15	51.02	65.22
jitterLocal_sma_qregc1	61.73	48.05	74.12	61.74	63.16	62.59	42.86	71.74
audSpec_Rfilt_sma[3]_qregc2	60.49	58.44	62.35	63.48	57.89	59.71	67.35	58.70
audSpec_Rfilt_sma[24]_qregc1	60.49	45.45	74.12	56.52	73.68	59.71	59.18	56.52
mfcc_sma_de[11]_skewness	60.49	66.23	55.29	65.22	36.84	58.27	67.35	50.00

Table 7.4 Classification performance for three different classifiers, reported on the independent test set.

	Classifier Names		
	SVM	GB	RF
Balanced accuracy	0.707	0.682	0.682
Precision	0.718	0.692	0.674
F1 score	0.7	0.675	0.69
Sensitivity	0.683	0.658	0.707
Specificity	0.732	0.707	0.658

Experiment 2

This section provides a detailed description of the model architecture and hyperparameters resulting from the optimization process for the DANN model.

The Adam optimization algorithm was employed with a learning rate of 0.001, and an exponential decay rate for the first moment β was set to 0.7 for the Encoder, Task, and Decoder blocks. A maximum of 200 epochs was chosen as a result of an early stopping procedure, which aimed to prevent excessive weight adaptation on training samples.

The input layer was deployed in order to accepts data with a shape of (1, 6386), where 6386 is the number of features in the dataset. The Encoder block starts with a Dropout layer (dropout rate = 0.5) followed by two densely connected layers with ReLU activation functions. The Task sub-network includes a Flatten layer to prepare the features for classification. It is followed by five Dropout-Dense layer pairs, each characterized by a dropout rate of 0.5 and ReLU activations. Given the primary objective of this block (i.e., learning binary discrimination between sober and intoxicated speakers), the last layer employs a sigmoid activation function. Similarly, the Discriminator network comprises five densely connected layers with ReLU activation functions, concluding with a softmax output layer to predict domain labels for the distinct domains in the dataset (i.e., subjects' IDs). Additionally, to maintain stable gradients of the loss function during training and ensure faster and more stable convergence, the weights of each fully connected layer were initialized using the Xavier Uniform initializer.

In terms of the loss functions, a Binary Cross-Entropy loss was utilized for binary classification in the task block, while a Sparse Categorical Cross-Entropy was

employed for multiclass classification. The λ value used in the computation of the Encoder loss was set to 0.01, in accordance with [313].

A schematic of the optimized DANN architecture is provided in Figure 7.2.

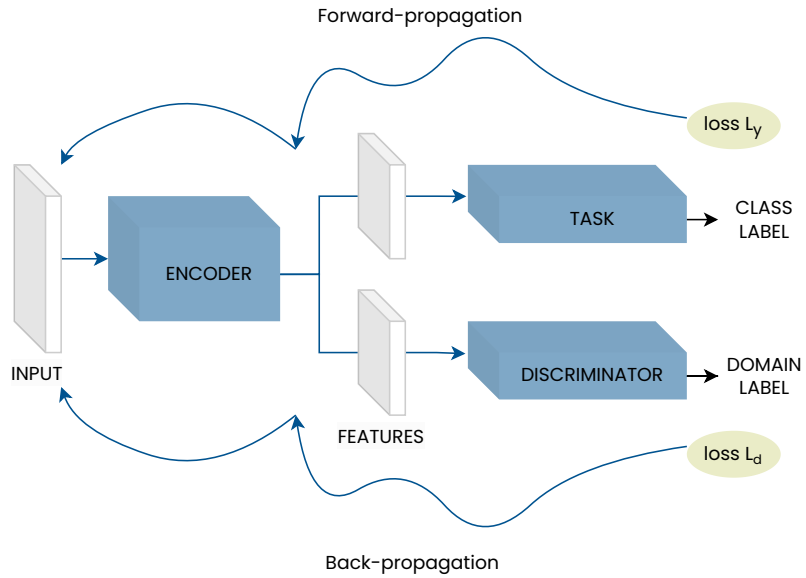


Fig. 7.2 Schematic architecture of the optimized Discriminative Adversarial Neural Network

Table 7.5 presents the model performance on both the training and test sets in the binary classification task of distinguishing between A and NA speakers. Due to the imbalanced distribution between the two classes, the primary focus was on evaluating the model effectiveness using balanced accuracy and AUC metrics. These metrics offer a more reliable assessment of the model performance, especially in scenarios where class imbalances could potentially introduce biases into the results.

Table 7.5 Performance of the Domain Adversarial Neural Network model on both train and test sets.

	Train set	Test set
Balanced accuracy	0.981	0.709
F1 score	0.981	0.610
AUC	0.969	0.709

Table 7.6 offers a comparative analysis the present study and related research that has tackled the binary classification task using the ALC dataset. The evaluation of each approach is based on the balanced accuracy ,which was measured on a separate

test set consisting of entirely new and unseen samples. This comparative examination sheds light on the efficacy of the proposed model in relation with prior investigations on the same dataset.

7.2.5 Discussions

This study had two primary objectives: first, to evaluate the potential for automatic classification of individuals as either sober or intoxicated based on vocal samples, and second, to propose a model that minimizes the impact of speech content and individual speaker characteristics.

The results of the analysis on the effects of alcohol intake on speech parameters are significant. The findings, as presented in Table 7.3, highlight that the impact of alcohol on speech parameters is highly personalized, with only a small subset of features exhibiting consistent trends in over 60% of the participants. Further analyses stratified by factors like gender, age, and BMI demonstrate that individuals with similar characteristics tend to display more consistent speech feature trends after alcohol consumption. These results emphasize the importance of developing algorithms that either take individual characteristics into account or are designed to be minimally influenced by speaker diversity. Notably, the most effective feature domains for discrimination were related to the RASTA-filtered auditory spectrum, a perceptual representation of perceived loudness obtained through signal filtering. Low-level spectral characteristics also proved to be significant in distinguishing between sober and intoxicated speakers.

In the binary classification using tongue-twister speech samples, the SVM model outperformed the others, achieving an accuracy of 70.7% on an independent test set. The classifiers demonstrated to be robust, with consistent F1-scores, precision, and recall in both A and NA classes. Nevertheless, it is important to note that this model was trained on a single speech task and, therefore, on a limited set of samples, which limits its real-world applicability.

Due to the limitations of the initial model and the insights gained from the feature analysis, a second experiment was carried out with the aim of creating a model minimally dependent on individual characteristics. Therefore, the DANN architecture was deployed, seeking to develop a feature representation that is less influenced by the subjects' characteristics. To achieve this, the classifier was trained

Table 7.6 Comparison of performance between the present study and similar studies involving the ALC corpus. Results are expressed as balanced accuracy on the test set. N: numerosity

Study ID	Accuracy	Model characteristics	N subjects
Present study	0.709	OpenSmile and Praat features; DANN classification model	162
[297]	0.67	OpenSmile features; Partial Least Square (PLS) classification model	154
[298]	0.705	OpenSmile and Praat features; Gaussian Mixture Model based on SVM	154
[299]	0.645	OpenSmile features; SVM embedding classification model	154
[300]	0.675	OpenSmile features; Universal Background Model (UBM); Hidden Markov Model (HMM)	154
[301]	0.688	OpenSmile features; Gaussian Mixture Model based on SVM	154
[302]	0.676	OpenSmile features; SVM and LDA classification model	154
[304]	0.666	OpenSmile features; SVM classification model	154
[305]	0.601	OpenSmile features; HMM classification model	145
[306]	0.68	Spectrogram-based features; CNN classification model	162
[307]	0.683	WAV2LETTER and Deepspeech-based; Recurrent Neural Network (RNN) classification model	154
[308]	0.666	OpenSmile features; SVM classification model	162
[309]	0.677	OpenSmile features; SVM and ResNET classification model	162

on the entire set of speech samples in the ALC corpus, making the model independent of the specific prompts spoken by the speakers. The DANN architecture yielded promising results, achieving a classification accuracy of 70.91% (Table 7.5) on a test set consisting of entirely new subjects. Although the results on the test set indicate strong performance and robust classification capabilities, the reduction in performance from training to testing suggests the need for an expanded sample size to further assess the model generalization ability.

When comparing the results to existing literature (as shown in Table 7.6), it is important to note that the examined studies primarily report balanced accuracy on the test set, omitting details about the training phase. This absence makes comprehensive comparisons of the model generalization capacity with other studies challenging. Additionally, while the models in these studies were trained on similar datasets with comparable features,the current study utilizes a slightly larger corpus compared to the best-performing study by Bone et al. [298]. Indeed, the present corpus includes additional 720 samples (8 subjects * 90 tasks), providing a more comprehensive investigation of the problem.

The comparison with other articles addressing similar objectives using the ALC dataset generally indicates performance levels below the 70% threshold, with only one study by Bone et al. [298] surpassing this benchmark and being comparable to the results presented in this study. The comparison between the two models reveals the commonality of both to include a step to minimize inter-subject differences. This finding, in line with the statistical analysis conducted in this study, suggests the critical need to develop algorithms that are minimally influenced by individual subject characteristics. Besides this, the approach presented in this dissertation presents crucial differences from the methodology employed in [298]. Indeed the model here proposed does not require intricate data-preparation techniques and it relies solely on DANN, thus significantly reducing the complexity. Moreover, the proposed model offers a key advantage by creating a prompt and subject-independent system, thus showing a high generalization capability. In addition, while this work focused on a specific aim, it showcases potential applicability across various domains. Indeed, it provides a prospective solution to the dependence on speaker-specific characteristics that often affects voice based models, holding promise for widespread applications.

7.2.6 Conclusions and Future Works

The study presented demonstrated the feasibility of an automatic classification based on vocal samples for the detection of intoxication, with the DANN architecture showing potential in creating models that are independent of individual characteristics and vocal prompts.

These encouraging results suggest the prospect of developing automatic devices suitable for the automotive context. These devices could gather information from any driver and dissuade them from starting the car if there is a potential risk of intoxication. Nevertheless, it is crucial to underscore that further refinement and validation of these findings are necessary. Firstly, expanding the dataset could mitigate issues related to overfitting and enhance the model ability to generalize. Indeed, intricate architectures such as DANN typically necessitate extensive, high-dimensional datasets for effective training. Thus, expanding the dataset employed in this study could lead to better and more generalized outcomes. Additionally, delving into advanced techniques in deep learning and transfer learning may lead to enhancements in classification performance.

Moreover, addressing the challenge of real-world applicability is of paramount importance. Developing models capable of handling diverse and uncontrolled environments, where speech patterns can vary considerably, is essential. Finally, research efforts should prioritize refining the interpretability and explainability of the models to facilitate their adoption in practical applications, particularly in safety-critical contexts such as automotive environments.

Regarding the classification model, an innovative, speaker-neutral ML algorithm was introduced. This model was specifically designed to overcome the inherent challenges associated with individual characteristics, providing a more robust and universally applicable approach to intoxication detection through speech analysis. This research not only deepens the understanding of the effects of alcohol on speech but also opens promising perspectives for the development of practical and widely applicable intoxication detection systems.

Chapter 8

Final Remarks

This dissertation effectively demonstrated the potential of voice analysis across diverse domains, encompassing the examination of neurodegenerative diseases and transient conditions arising from altered sleep quality or alcohol intoxication. The conducted experiments further delved into the impact of the simultaneous presence of two or more pathologies, providing a deeper understanding of how comorbidities may alter the vocal signal. Additionally, a series of investigations explored the influence of individual-specific characteristics, including age, gender, and language, among others.

In the experiments exploring Parkinson's disease, the efficacy of specific vocal tasks, emphasizing the importance of a concise protocol involving vowels, particularly /a/, and sentences, was revealed. Notably, occlusive sounds in the Italian language demonstrated superior capacity for capturing PD-associated impairments. The analysis of acoustic parameters, including MFCC coefficients, F0, Shimmer, and Jitter, highlighted their robustness and effectiveness for the application at hand. The promising outcomes, although based on limited sample sizes, suggested the feasibility for cross-lingual models.

Similarly, investigations into obesity and GERD classification showcased the efficiency of ML models, addressing limitations related to dataset size and gender-age imbalances. The study's applicability to more impactful health conditions, like PD, discloses the option for simultaneous evaluations of co-existing conditions.

The exploration of sleep quality classification demonstrated remarkable efficiency despite recording challenges, particularly for females. Limitations involving

subjective sPSQI scores were identified, revealing the need for future incorporation of objective parameters and expanded datasets for improved model performance.

In the context of alcohol intoxication, the study exhibited the potential of DANN architecture for speaker-neutral models, emphasizing the need for refinement, expanded datasets, and real-world applicability considerations.

While the experiments presented in this dissertation successfully demonstrated the potential of voice analysis across diverse domains, it is crucial to acknowledge certain limitations and offer critical insights to guide future research and applications. Specifically, findings from the experiments conducted were promising but often relied on limited sample sizes. To ensure the robustness of analyses, future research endeavors should prioritize expanding datasets, allowing for more comprehensive and reliable conclusions.

Moreover, despite extensive efforts to prioritize model interpretability through the predominant use of shallow algorithms and post hoc analysis of extracted acoustic parameters, it is crucial to note that these endeavors have inherent limitations. Indeed, there exists an unavoidable absence of a direct correspondence between certain specific features (e.g., MFCC(6)) and a particular aspect of vocal production. Future studies could delve deeper into this aspect and, with the support of specialized professionals, shed more light on aspects characterized by lower interpretability.

blackAs for sleep quality classification, while efficient, the experiment revealed limitations related to subjective sPSQI scores. Incorporating objective parameters and larger datasets could address these shortcomings, ensuring more reliable models. Similary, the study on alcohol intoxication presented promising outcomes with the DANN architecture. However, the call for refinement, expanded datasets, and considerations for real-world applicability should guide future investigations.

The overall evidence highlighted the inherent variability in vocal samples due to individual characteristics and the influence of recording modalities. While proposed solutions, such as covariate inclusion and domain adversarial networks, show promise, further research should delve into refining these techniques and expanding their applicability across diverse clinical conditions.

For each of the experiments conducted and reported, a detailed literature analysis was carried out, and comparisons with similar articles were presented. In these comparisons, special attention was given to ensuring consistency and robust compa-

rability by utilizing similar data and training conditions. However, the diversity of experimental conditions, such as corpora and validation methods, precluded a precise comparison that would allow for a comprehensive evaluation of the proposed model. Hence, future research endeavors should consider incorporating publicly available models or those with detailed descriptions to ensure reproducibility. Employing identical data and training/testing conditions in these studies would facilitate a clear and comprehensive comparison, eliminating potential biases that could lead to the overestimation of performance.

In summary, the outcomes of this dissertation collectively advocated for the possibility of mitigating the influence of potentially confounding factors by constructing specialized and highly tailored models for specific applications.

The findings underscored the significance of developing models that account for the intricacies of individual conditions, enabling a more accurate and context-specific analysis. This approach minimizes the impact of confounding variables, thereby enhancing the reliability, applicability, and interpretability of the generated models. Such an improved methodology not only contributes to the robustness of the results obtained but also lays a solid foundation for a more effective implementation of voice analysis techniques in various real-world scenarios.

References

[1] John Field. *Psycholinguistics: The Key Concepts*. Routledge, 2004.

[2] Gregory Hickok. Computational neuroanatomy of speech production. *Nature Reviews Neuroscience*, 13:135–145, 2 2012.

[3] Benesty Jacob, Sondhiand Mohan, and Huang Yiteng. *Springer Handbook of Speech Processing*. Springer, 2008.

[4] Mills Wesley. *Voice production in singing and speaking based on scientific principles*. J. B. Lippincott & Co., 2006.

[5] Olek Remesz. Larynx-antero-lateral view, with external muscles of larynx visible, modified. SVG version of PD picture from Gray's Anatomy, 2008. CC-BY-SA ver. 2.5, 2.0, 1.0.

[6] Ludlow Christy. Central nervous system control of voice and swallowin. *Journal of clinical neurophysiology : official publication of the American Electroencephalographic Society*, 32:294–303, 2015.

[7] Binkofski Ferdinand and Buccino Giovanni. Motor functions of the broca's region. *Brain and language*, 89:362–369, 2004.

[8] Juan Rafael Orozco-Arroyave andlorian F Hönig, Julian D Arias-Londoño, J Francisco Vargas-Bonilla, Khaled Daqrouq, Sabin Skodda, Jan Rusz, and Elmar Nöth. Automatic detection of Parkinson's disease in running speech spoken in three different languages. *The Journal of the Acoustical Society of America*, pages 481–500.

[9] Juan Ignacio Godino-Llorente, Víctor Osma-Ruiz, Nicolás Sáenz-Lechón, Pedro Gómez-Vilda, Manuel Blanco-Velasco, and Fernando Cruz-Roldán. The effectiveness of the glottal to noise excitation ratio for the screening of voice disorders. *Journal of Voice*, 24:47–56, 2010.

[10] Jorge Andrés Gómez-García, Laureano Moro-Velázqueze, and Juan ignacio Godino Llorente. On the design of automatic voice condition analysis systems. part ii: Review of speaker recognition techniques and study on the effects of different variability factors. *Biomedical Signal Processing and Control*, 48:128–143, 2019.

[11] David M Howard and Jamie Andrea-Shyla Angus. *Introduction to human speech production, human hearing and speech analysis*. Chapman Hall, 1998.

[12] Gunnar Fant. The source filter concept in voice production. 1981.

[13] Kristina Simonyan, Hermann Ackermann, Edward F. Chang, and Jeremy D. Greenlee. New developments in understanding the complexity of human speech production. *Journal of Neuroscience*, 36:11440–11448, 11 2016.

[14] Federica Amato, Luigi Borzi, Gabriella Olmo, Carlo Alberto Artusi, Gabriele Imbalzano, and Leonardo Lopiano. Speech impairment in parkinson's disease: acoustic analysis of unvoiced consonants in italian native speakers. *IEEE Access*, 9:1–1, 2021.

[15] Antonio Suppa, Francesco Asci, Giovanni Saggio, Pietro Di Leo, Zakarya Zarezadeh, Gina Ferrazzano, Giovanni Ruoppolo, Alfredo Berardelli, and Giovanni Costantini. Voice Analysis with Machine Learning: One Step Closer to an Objective Diagnosis of Essential Tremor. *Movement Disorders*, 36(6):1401–1410, June 2021.

[16] Giovanni Saggio Antonio Suppa Francesco Asci, Giovanni Costantini. Fostering Voice Objective Analysis in Patients with Movement Disorders. *Movement Disorders*, 36(4):1041, 2021.

[17] Gavin Doherty Anja Thieme, Danielle Belgrave. Machine Learning in Mental Health: A systematic review of the HCI literature to support the development of effective and implementable ML Systems. *ACM Transactions on Computer-Human Interaction*, 27(5), 2020.

[18] Rosa M Bermúdez de Alvear, Francisco J Barón-López, María D Alguacil, and Mark S Dawid-Milner. Interactions between voice fundamental frequency and cardiovascular parameters. Preliminary results and physiological mechanisms. *Logopedics Phoniatrics Vocology*, 38(2):52–58, 2013.

[19] Pavel Kukharchik. Speech signal processing based on wavelets and SVM for vocal tract pathology detection. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5099 LNCS:192–199, 2008.

[20] Lourdes Bernadete Rocha de Souza and Marquiony Marques dos Santos. Body mass index and acoustic voice parameters: is there a relationship? *Brazilian Journal of Otorhinolaryngology*, 84(4):410–415, 2018.

[21] Ana Luara Ferreura Fonseca, Wilson Salgado, and Roberto Oliveira Dantas. Maximum Phonation Time in People with Obesity Not Submitted or Submitted to Bariatric Surgery. *Journal of Obesity*, 2019:14–19, 2019.

[22] Adolfo M. García, Daniel Escobar-Grisales, Juan Camilo Vásquez Correa, Yamile Bocanegra, Leonardo Moreno, Jairo Carmona, and Juan Rafael Orozco-Arroyave. Detecting parkinson's disease and its cognitive phenotypes

via automated semantic analyses of action stories. *npj Parkinson's Disease*, 8, 12 2022.

[23] Laureano Moro-Velazquez, Jorge Andres Gomez-Garcia, Juan Ignacio Godino-Llorente, Jesús Villalba, Jan Rusz, Stephanie Shattuck-Hufnagel, and Najim Dehak. A forced gaussians based methodology for the differential evaluation of Parkinson's disease by means of speech processing. *Biomedical Signal Processing and Control*, 48:205–220, 2019.

[24] Juan Rafael Orozco-Arroyave. *Analysis of speech of people with Parkinson's disease*, volume 41. Logos Verlag Berlin GmbH, 2016.

[25] Karen Chenausky, Joel MacAuslan, and Richard Goldhor. Acoustic analysis of pd speech. *Parkinson's Disease*, 2011, 2011.

[26] Javier Hernando Mireia Farrús and Pascual Ejarque. Jitter and shimmer measurements for speaker recognition. *Proceedings of the Interspeech 2007*, 2007.

[27] Elkyn A. Belalcazar-Bolaños, Juan Rafael Orozco, Julian D. Arias-Londoño, and Elmar Noeth. Automatic detection of Parkinson's disease using noise measures of speech. *Symposium of Signals, Images and Artificial Vision - 2013, STSIVA 2013*, 2013.

[28] Alan V. Oppenheim and Ronald W. Schafer. From frequency to quefrency: a history of the cepstrum. *IEEE Signal Processing Magazine*, 2004.

[29] Namrata Dave. Feature extraction methods lpc , plp and mfcc in speech recognition. *International Journal for Advance Research in Engineering and Technology*, 1:1–5, 2013.

[30] Hynek Hermansky and Nelson Morgan. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2:578–589, 1994.

[31] Floris Takens. Detecting strange attractors in turbulence. In David Rand and Lai-Sang Young, editors, *Dynamical Systems and Turbulence, Warwick 1980*, volume 898, pages 366–381. Springer Berlin Heidelberg, Berlin, Heidelberg, 1981. Series Title: Lecture Notes in Mathematics.

[32] Herbert Edelsbrunner, David G. Kirpatrick, and Raimund Seidel. On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29(4):551–559, July 1983.

[33] Jorge Andrés Gómez-García, Laureano Moro-Velázqueze, and Juan ignacio Godino Llorente. On the design of automatic voice condition analysis systems. part iii: review of acoustic modelling strategies. *Biomedical Signal Processing and Control*, 66, 2021.

[34] Max A. Little, Patrick E. McSharry, Stephen J. Roberts, Declan A.E. Costello, and Irene M. Moroz. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *BioMedical Engineering Online*, 6:1–19, 2007.

[35] Jan Rusz, Jan Hlavnicka, Tereza Tykalova, Michal Novotny, Petr Dusek, Karel Sonka, and Evzen Ruzicka. Smartphone Allows Capture of Speech Abnormalities Associated with High Risk of Developing Parkinson's Disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(8):1495–1507, 2018.

[36] Jan Hlavnika, Roman Cmejla, Tereza Tykalová, Karel Šonka, Evzen Ruzicka, and Jan Rusz. Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder. *Scientific Reports*, 7:1–13, 2017.

[37] Tereza Tykalova, Jan Rusz, Jiri Klempir, Roman Cmejla, and Evzen Ruzicka. Distinct patterns of imprecise consonant articulation among Parkinson's disease, progressive supranuclear palsy and multiple system atrophy. *Brain and Language*, 165:1–9, 2017.

[38] Paul Boersma and Vincent van Heuven. Speak and unSpeak with PRAAT. 5(9):8, 2001.

[39] Yannick Jadoul, Bill Thompson, and Bart de Boer. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15, November 2018.

[40] Jan G. Švec and Svante Granqvist. Guidelines for Selecting Microphones for Human Voice Production Research. *American Journal of Speech-Language Pathology*, 19(4):356–368, November 2010.

[41] Theodoros Giannakopoulos. pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. *PLOS ONE*, 10(12):e0144610, December 2015.

[42] Athanasios Tsanas, Max A. Little, Patrick E. McSharry, and Lorraine O. Ramig. Accurate telemonitoring of parkinsons disease progression by non-invasive speech tests. *IEEE Transactions on Biomedical Engineering*, 57, 2010.

[43] Athanasios Tsanas, Max A. Little, Patrick E. McSharry, and Lorraine O. Ramig. Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. *Journal of the Royal Society Interface*, 8:842–855, 2011.

[44] Athanasios Tsanas. Accurate telemonitoring of parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning. *Thesis*, 2012.

[45] Pedro Gómez, Victoria Rodellar, Víctor Nieto, Agustín Álvarez, Bartolomé Scola, Carlos Ramírez, Daniel Poletti, and Mario Fernández. BioMet®Phon: A System to Monitor Phonation Quality in the Clinics. page 6, 2013.

[46] Yien-Liang Shue. *The voice source in speech production: data, analysis and models*. PhD thesis, 2010.

[47] Yen-Liang Shue, Patricia Keating, and Chad Vicenik. VOICESAUCE: A program for voice analysis. *The Journal of the Acoustical Society of America*, 126(4):2221, 2009.

[48] Florian Eyben, Martin Wollmer, and Bjorn Schuller. OpenEAR — Introducing the munich open-source emotion and affect recognition toolkit. pages 1–6, September 2009.

[49] Bjorn Schuller, Florian Eyben, and Gerhard Rigoll. Fast and Robust Meter and Tempo Recognition for the Automatic Discrimination of Ballroom Dance Styles. pages I–217–I–220, April 2007.

[50] Bjorn Schuller, Stefan Steidl, and Anton Batliner. The INTERSPEECH 2009 Emotion Challenge. page 4.

[51] Juan Rafael Orozco-Arroyave, Juan Camilo Vásquez-Correa, Jesús Francisco Vargas-Bonilla, R. Arora, N. Dehak, P.S. Nidadavolu, H. Christensen, F. Rudzicz, M. Yancheva, H. Chinaei, A. Vann, N. Vogler, T. Bocklet, M. Cernak, J. Hannink, and Elmar Nöth. NeuroSpeech: An open-source software for Parkinson's speech analysis. *Digital Signal Processing*, 77:207–221, June 2018.

[52] Matti Airas, Hannu Pulakka, Tom Bäckström, and Paavo Alku. A toolkit for voice inverse filtering and parametrisation. pages 2145–2148, September 2005.

[53] Ali Samii, John G. Nutt, and Bruce R. Ransom. Parkinson's disease. *Lancet*, 363:1783–1793, 2004.

[54] Amy Reeve, Eve Simcox, and Doug Turnbull. Ageing and Parkinson's disease: why is advancing age the biggest risk factor? *Ageing Res Rev*, 14:19–30, Mar 2014.

[55] Philip Cole Dimitrios Trichopoulos Jack Mandel Karin Wirdefeldt, Hans-Olov Adami. Epidemiology and etiology of Parkinson's disease: a review of the evidence. *Eur J Epidemiol*, 26 Suppl 1:1–58, Jun 2011.

[56] E. Ray Dorsey, Todd Sherer, Michael S. Okun, , and Bastiaan R. Bloemd. The Emerging Evidence of the Parkinson Pandemic. *J Parkinsons Dis*, 8(s1):S3–S8, 2018.

[57] Bastiaan R. Bloem, Michael S. Okun, and Christine Klein. Parkinson's disease. *The Lancet*, 397:2284–2303, 6 2021.

[58] Lorraine V. Kalia and Anthony E. Lang. Parkinson's disease. *The Lancet*, 386:896–912, 2015.

[59] Jacob Oliver Day and Stephen Mullin. The genetics of parkinson's disease and implications for clinical practice. *Genes*, 12, 7 2021.

[60] J. Jankovic. Parkinson's disease: clinical features and diagnosis. *J Neurol Neurosurg Psychiatry*, 79(4):368–376, Apr 2008.

[61] Michael S Okun Melissa J Armstrong. Diagnosis and Treatment of Parkinson Disease: A Review. *JAMA*, 323(6):548–560, 02 2020.

[62] Carlos H Schenck, Scott R Bundlie, Milton G Ettinger, and Mark W Mahowald. Chronic behavioral disorders of human rem sleep: A new category of parasomnia. 1986.

[63] José Haba-Rubio, Birgit Frauscher, Pedro Marques-Vidal, Jérôme Toriel, Nadia Tobback, Daniela Andries, Martin Preisig, Peter Vollenweider, Ronald Postuma, and Raphaël Heinzer. Prevalence and determinants of rapid eye movement sleep behavior disorder in the general population. *Sleep*, 41, 2 2018.

[64] Andrea Galbiati, Laura Verga, Enrico Giora, Marco Zucconi, and Luigi Ferini-Strambi. The risk of neurodegeneration in rem sleep behavior disorder: A systematic review and meta-analysis of longitudinal studies. *Sleep Medicine Reviews*, 43:37–46, 2 2019.

[65] Giovanni Defazio, Marta Guerrieri, Daniele Liuzzi, Angelo Fabio Gigante, and Vincenzo di Nicola. Assessment of voice and speech symptoms in early Parkinson's disease by the robertson dysarthria profile. *Neurological Sciences*, 37:443–449, 2016.

[66] Tobias Bocklet, Elmar Nöth, Georg Stemmer, Hana Ruzickova, and Jan Rusz. Detection of persons with Parkinson's disease by acoustic, vocal, and prosodic analysis. *2011 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2011, Proceedings*, 2011.

[67] Brian T. Harel, Michael S. Cannizzaro, Henrí Cohen, Nicole Reilly, and Peter J. Snyder. Acoustic characteristics of Parkinsonian speech: A potential biomarker of early disease progression and treatment. *Journal of Neurolinguistics*, 2004.

[68] João Massano and Kailash P. Bhatia. Clinical approach to parkinson's disease: Features, diagnosis, and principles of management. *Cold Spring Harbor Perspectives in Medicine*, 2, 2012.

[69] Alberto Albanese. Standard strategies for diagnosis and treatment of patients with newly diagnosed Parkinson disease: ITALY. *Neurol Clin Pract*, 3(6):476–477, Dec 2013.

[70] Carlos Pérez-López Marti Pie Joan Calvet Albert Samà Chiara Capra Andreu Català Alejandro Rodríguez-Molinero Daniel Rodríguez-Martín, Joan Cabestany. A new paradigm in parkinson's disease evaluation with wearable medical devices: A review of stat-ontm. *Frontiers in Neurology*, 13, 2022.

[71] José Obeso Álvaro Sánchez-Ferro Mariana H G Monje, Guglielmo Foffani. New sensor and wearable technologies to aid in the diagnosis and treatment monitoring of parkinson's disease. *Annual review of biomedical engineering*, 21:111–143, 2019.

[72] Roongroj Bhidayasiri and Pablo Martinez-Martin. Clinical Assessments in Parkinson's Disease: Scales and Monitoring. *International review of neurobiology*, 132:129–182, 2017.

[73] Lorenzo Morlan Gracia Josè Balzeiro Gómez J Francisco Martínez-Sarriés Feliz Bermejo Pablo Martínez-Martín, Antonio Gil-Nagel. Unified parkinson's disease rating scale characteristics and structure. *Movement disorders : official journal of the Movement Disorder Society*, 9:76–83, 1 1994.

[74] Margaret M Hoehn and Melvin D Yahr. Parkinsonism: onset, progression, and mortality. *Neurology*, 17(5):427–442, 1967.

[75] Stephen G. Reich and Joseph M. Savitt. Parkinson's Disease. *Medical Clinics of North America*, 103(2):337–350, 03 2019.

[76] Roberta Balestrino and Anthony H. V. Schapira. Parkinson disease. *European journal of neurology*, 27(1):27–42, 01 2020.

[77] Joseph Jankovic. Motor fluctuations and dyskinesias in parkinson's disease: clinical manifestations. *Movement disorders: official journal of the Movement Disorder Society*, 20(S11):S11–S16, 2005.

[78] Thomas Müller. Catechol-o-methyltransferase inhibitors in parkinson's disease. *Drugs*, 75:157–174, 2015.

[79] Regina Katzenschlager, Cristina Sampaio, João Costa, and Andrew Lees. Anticholinergics for symptomatic management of parkinson´s disease. *Cochrane Database of Systematic Reviews*, 2010, 7 2002.

[80] Thomas Wichmann and Mahlon R. DeLong. Deep brain stimulation for movement disorders of basal ganglia origin: Restoring function or functionality? *Neurotherapeutics*, 13:264–283, 4 2016.

[81] Stefan Jun Groiss, Lars Wojtecki, Martin Sudmeyer, and Arthur Schnitzler. Deep brain stimulation in parkinson-s disease. *Therapeutic Advances in Neurological Disorders*, 2:379–391, 2009.

[82] Alexandra-Maria Tăuţan, Bogdan Ionescu, and Emiliano Santarnecchi. Artificial intelligence in neurodegenerative diseases: A review of available tools with a focus on machine learning techniques. *Artificial Intelligence in Medicine*, 117:102081, July 2021.

[83] Jorge Andrés Gómez-García, Laureano Moro-Velázqueze, and Juan ignacio Godino Llorente. On the design of automatic voice condition analysis systems. Part III: review of acoustic modelling strategies. *Biomedical Signal Processing and Control*, 66:102049, April 2021.

[84] Andrew Ma, Kenneth K. Lau, and Dominic Thyagarajan. Voice changes in Parkinson's disease: What are they telling us? *Journal of Clinical Neuroscience*, 72:1–7, 2020.

[85] L. Brabenec, J. Mekyska, Z. Galaz, and Irena Rektorova. Speech disorders in Parkinson's disease: early diagnostics and effects of medication and brain stimulation. *Journal of Neural Transmission*, 124(3):303–334, March 2017.

[86] Laureano Moro-Velazquez, Jorge A. Gomez-Garcia, Julian D. Arias-Londoño, Najim Dehak, and Juan I. Godino-Llorente. Advances in Parkinson's Disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects. *Biomedical Signal Processing and Control*, 66:102418, April 2021.

[87] Federica Amato, Giovanni Saggio, Valerio Cesarini, Gabriella Olmo, and Giovanni Costantini. Machine learning- and statistical-based voice analysis of Parkinson's disease patients: A survey. *Expert Systems with Applications*, 219, 6 2023.

[88] Laureano Moro-Velazquez, Jorge Gomez-Garcia, Najim Dehak, and Juan Ignacio Godino-Llorente. New tools for the differential evaluation of Parkinson's disease using voice and speech processing. pages 165–169, 2021.

[89] Achraf Benba, Abdelilah Jilbab, and Ahmed Hammouch. Voice analysis for detecting patients with Parkinson's disease using the hybridization of the best acoustic features. *International Journal on Electrical Engineering and Informatics*, 8:108–116, 3 2016.

[90] Adedolapo Aishat Toye and Suryaprakash Kompalli. Comparative study of speech analysis methods to predict Parkinson's disease. 11 2021.

[91] Jidde Jacobi, Teja Rebernik, Roel Jonkers, Michael Proctor, Ben Maassen, and Martijn Wieling. The effect of levodopa on vowel articulation in Parkinson's disease: A cross-linguistic study. *19th ICPhS*, pages 1069–1073, 2019.

[92] Juan Camilo Vásquez-Correa, Tomas Arias-Vergara, Cristian D. Rios-Urrego, Maria Schuster, Jan Rusz, Juan Rafael Orozco-Arroyave, and Elmar Nöth. Convolutional neural networks and a transfer learning strategy to classify Parkinson's disease from speech in three different languages. *Lecture Notes in*

Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11896 LNCS:697–706, 2019.

[93] Abner Hernandez and Minhwa Chung. Dysarthria classification using acoustic properties of fricatives dysarthria classification using acoustic properties of fricatives. pages 4–6, 2019.

[94] David Montaña, Yolanda Campos-Roca, and Carlos J. Pérez. A Diadochokinesis-based expert system considering articulatory features of plosive consonants for early detection of Parkinson's disease. *Computer Methods and Programs in Biomedicine*, 154:89–97, February 2018.

[95] Jan Rusz, Jan Hlavnička, Michal Novotný, Tereza Tykalová, Amelie Pelletier, Jacques Montplaisir, Jean Francois Gagnon, Petr Dušek, Andrea Galbiati, Sara Marelli, Paul C. Timm, Luke N. Teigen, Annette Janzen, Mahboubeh Habibi, Ambra Stefani, Evi Holzknecht, Klaus Seppi, Elisa Evangelista, Anna Laura Rassu, Yves Dauvilliers, Birgit Högl, Wolfgang Oertel, Erik K. St. Louis, Luigi Ferini-Strambi, Evžen Růžička, Ronald B. Postuma, and Karel Šonka. Speech biomarkers in rapid eye movement sleep behavior disorder and Parkinson disease. *Annals of Neurology*, 90:62–75, 7 2021.

[96] Daniel Kovac, Jiri Mekyska, Vered Aharonson, Pavol Harar, Zoltan Galaz, Steven Rapcsak, Juan Rafael Orozco-Arroyave, Lubos Brabenec, and Irena Rektorova. Exploring language-independent digital speech biomarkers of hypokinetic dysarthria. 2022.

[97] Margherita Fabbri, Isabel Guimarães, Rita Cardoso, Miguel Coelho, Leonor Correia Guedes, Mario M. Rosa, Catarina Godinho, Daisy Abreu, Nilza Gonçalves, Angelo Antonini, and Joaquim J. Ferreira. Speech and Voice Response to a Levodopa Challenge in Late-Stage Parkinson's Disease. *Frontiers in Neurology*, 8:432, August 2017.

[98] Hannah Im, Scott Adams, Anita Abeyesekera, Marcus Pieterman, Greydon Gilmore, and Mandar Jog. Effect of Levodopa on Speech Dysfluency in Parkinson's Disease. *Movement Disorders Clinical Practice*, 6(2):150–154, February 2019.

[99] Nemuel D. Pah, Mohammod A. Motin, Peter Kempster, and Dinesh K. Kumar. Detecting Effect of Levodopa in Parkinson's Disease Patients Using Sustained Phonemes. *IEEE Journal of Translational Engineering in Health and Medicine*, 9:1–9, 2021.

[100] J. R. Orozco-Arroyave, J. C. Vdsquez-Correa, F. Honig, J. D. Arias-Londono, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, and E. Noth. Towards an automatic monitoring of the neurological state of parkinson's patients from speech. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2016-May, 2016.

[101] Juan Rafael Orozco-Arroyave, Juan Camilo Vásquez-Correa, Philipp Klumpp, Paula Andrea Pérez-Toro, Daniel Escobar-Grisales, Nils Roth, Cristian David Ríos-Urrego, Martin Strauss, Helber Andrés Carvajal-Castaño, Sebastian Bayerl, Luis Reinel Castrillón-Osorio, Tomas Arias-Vergara, Arne Künderle, Felipe Orlando López-Pabón, Luis Felipe Parra-Gallego, Björn Eskofier, Luis Felipe Gómez-Gómez, Maria Schuster, and Elmar Nöth. Apkinson: The smartphone application for telemonitoring parkinson's patients through speech, gait and hands movement. *Neurodegenerative Disease Management*, 10, 2020.

[102] Javier Carrón, Yolanda Campos-Roca, Mario Madruga, and Carlos J. Pérez. A mobile-assisted voice condition analysis system for Parkinson's disease: assessment of usability conditions. *BioMedical Engineering OnLine*, 20(1):114, December 2021.

[103] Pedro Gómez-Vilda, Andrés Gómez-Rodellar, Daniel Palacios-Alonso, Victoria Rodellar-Biarge, and Agustín Álvarez Marquina. The Role of Data Analytics in the Assessment of Pathological Speech—A Critical Appraisal. *Applied Sciences*, 12(21):11095, November 2022.

[104] Giovanni Dimauro, Vincenzo Di Nicola, Vitoantonio Bevilacqua, Danilo Caivano, and Francesco Girardi. Assessment of speech intelligibility in Parkinson's disease using a speech-to-text system. *IEEE Access*, 5:22199–22208, 2017.

[105] Juan Rafael Orozco-Arroyave, Julián David Arias-Londoño, Jesús Francisco Vargas-Bonilla, María Claudia González-Rátiva, and Elmar Nöth. New spanish speech corpus database for the analysis of people suffering from Parkinson's disease. *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pages 342–347, 2014.

[106] Jan Hlavnicka, Roman Cmejla, Jiri Klempir, Evzen Ruzicka, and Jan Rusz. Acoustic tracking of pitch, modal, and subharmonic vibrations of vocal folds in Parkinson's disease and parkinsonism. *IEEE Access*, 7:150339–150354, 2019.

[107] Antonio Suppa, Giovanni Costantini, Francesco Asci, Pietro Di Leo, Mohammad Sami Al-Wardat, Giulia Di Lazzaro, Simona Scalise, Antonio Pisani, and Giovanni Saggio. Voice in Parkinson's disease: A machine learning study. *Frontiers in Neurology*, 13, 2 2022.

[108] Evaldas Vaiciukynas, Antanas Verikas, Adas Gelzinis, and Marija Bacauskiene. Detecting Parkinson's disease from sustained phonation and speech signals. *PLOS ONE*, 12(10):e0185613, October 2017.

[109] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. *MM 2013 - Proceedings of the 2013 ACM Multimedia Conference*, 2013.

[110] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José Zapata, and Xavier Serra. Essentia: An audio analysis library for music information retrieval. *Proceedings of the 14th International Society for Music Information Retrieval Conference, ISMIR 2013*, 2013.

[111] Daniel McEnnis, Cory McKay, and Ichiro Fujinaga. jaudio: Additions and improvements. *ISMIR 2006 - 7th International Conference on Music Information Retrieval*, 2006.

[112] Francesco Piazza Holger Crysand, Giovanni Tummarello. Mpeg-7 encoding and processing : Mpeg7audioenc+mpeg7audiodb. 2004.

[113] Benoit Mathieu, Slim Essid, Thomas Fillon, Jacques Prado, and Gaël Richard. Yaafe, an easy to use and efficient audio feature extraction software. *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010*, 2010.

[114] Max A. Little, Patrick E. McSharry, Eric J. Hunter, Jennifer Spielman, and Lorraine O. Ramig. Suitability of dysphonia measurements for telemonitoring of parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 56, 2009.

[115] C. Okan Sakar, Gorkem Serbes, Aysegul Gunduz, Hunkar C. Tunc, Hatice Nizam, Betul Erdogdu Sakar, Melih Tutuncu, Tarkan Aydin, M. Erdem Isenkul, and Hulya Apaydin. A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform. *Applied Soft Computing*, 74:255–263, 2019.

[116] Betul Erdogdu Sakar, M. Erdem Isenkul, C. Okan Sakar, Ahmet Sertbas, Fikret Gurgen, Sakir Delil, Hulya Apaydin, and Olcay Kursun. Collection and Analysis of a Parkinson Speech Dataset With Multiple Types of Sound Recordings. *IEEE Journal of Biomedical and Health Informatics*, 17(4):828–834, July 2013.

[117] Lizbeth Naranjo, Carlos J. Pérez, Yolanda Campos-Roca, and Jacinto Martín. Addressing voice recording replications for Parkinson's disease detection. *Expert Systems with Applications*, 46:286–292, March 2016.

[118] Brian M. Bot, Christine Suver, Elias Chaibub Neto, Michael Kellen, Arno Klein, Christopher Bare, Megan Doerr, Abhishek Pratap, John Wilbanks, E. Ray Dorsey, Stephen H. Friend, and Andrew D. Trister. The mPower study, Parkinson disease mobile data collected using ResearchKit. *Scientific Data*, 3(1):160011, December 2016.

[119] John Prince, Siddharth Arora, and Maarten de Vos. Big data in Parkinson's disease: using smartphones to remotely detect longitudinal disease phenotypes. *Physiological Measurement*, 39(4):044005, April 2018.

[120] Sabrina Scimeca, Federica Amato, Gabriella Olmo, Francesco Asci, Antonio Suppa, Giovanni Costantini, and Giovanni Saggio. Robust and language-independent acoustic features in parkinson's disease. *Frontiers in Neurology*, 14, 6 2023.

[121] Paul Boersma. Praat, a system for doing phonetics by computer. *Glot International*, 5, 2001.

[122] Wouter Gevaert, Georgi Tsenov, and Valeri Mladenov. Neural networks used for speech recognition. *Journal of Automatic Control*, 20:1–7, 2010.

[123] Biswajit Karan, Sitanshu Sekhar Sahu, Juan Rafael Orozco-Arroyave, and Kartik Mahto. Hilbert spectrum analysis for automatic detection and evaluation of Parkinson's speech. *Biomedical Signal Processing and Control*, 61:102050, 2020.

[124] Tamer A. Mesallam, Mohamed Farahat, Khalid H. Malki, Mansour Alsulaiman, Zulfiqar Ali, Ahmed Al-Nasheri, and Ghulam Muhammad. Development of the arabic voice pathology database and its evaluation by using speech features and machine learning algorithms. *Journal of Healthcare Engineering*, 2017, 2017.

[125] Juan Camilo Vásquez-Correa, Tomas Arias-Vergara, J. R. Orozco-Arroyave, Björn Eskofier, Jochen Klucken, and Elmar Nöth. Multimodal assessment of Parkinson's disease: A deep learning approach. *IEEE Journal of Biomedical and Health Informatics*, 23:1618–1630, 2019.

[126] Betul Erdogdu Sakar, Gorkem Serbes, and C. Okan Sakar. Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease. *PloS one*, 12:e0182428, 2017.

[127] C. Okan Sakar and Olcay Kursun. Telediagnosis of Parkinson's disease using measurements of dysphonia. *Journal of Medical Systems*, 34:591–599, 2010.

[128] Cees G M Snoek, Marcel Worring, and Arnold W M Smeulders. Early versus late fusion in semantic video analysis. *Proceedings of the 13th ACM International Conference on Multimedia, MM 2005*, pages 399–402, 2005.

[129] Federica Amato, Luigi Borzì, Gabriella Olmo, Juan Rafael, and Orozco Arroyave. An algorithm for parkinson ' s disease speech classification based on isolated words analysis. *Health Information Science and Systems*, 2021.

[130] Jerome H. Friedman, Jon Louis Bentley, and Raphael Ari Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software (TOMS)*, 3:209–226, 1977.

[131] J R Orozco-Arroyave, Florian Hönig, J D Arias-Londoño, J F Vargas-Bonilla, and Elmar Nöth. Spectral and cepstral analyses for Parkinson's disease detection in spanish vowels and words. *Expert Systems*, 32:688–697, 2015.

[132] Laiba Zahid, Muazzam Maqsood, Mehr Yahya Durrani, Maheen Bakhtyar, Junaid Baber, Habibullah Jamal, Irfan Mehmood, and Oh Young Song. A spectrogram-based deep feature assisted computer-aided diagnostic system for Parkinson's disease. *IEEE Access*, 8:35482–35495, 2020.

[133] Miron B. Kursa and Witold R. Rudnicki. Feature selection with the boruta package. *Journal of Statistical Software*, 36(11):1–13, 2010.

[134] Javier Carrón, Yolanda Campos-Roca, Mario Madruga, and Carlos J. Pérez. A mobile-assisted voice condition analysis system for Parkinson's disease: assessment of usability conditions. *BioMedical Engineering Online*, 20, 12 2021.

[135] Laureano Moro-Velazquez, Jorge A. Gomez-Garcia, Juan I. Godino-Llorente, Francisco Grandas-Perez, Stefanie Shattuck-Hufnagel, Virginia Yagüe-Jimenez, and Najim Dehak. Phonetic relevance and phonemic grouping of speech in the automatic detection of Parkinson's disease. *Scientific Reports*, 9:1–16, 2019.

[136] Federica Amato, Giovanni Saggio, Valerio Cesarini, Gabriella Olmo, and Giovanni Costantini. Machine learning- and statistical-based voice analysis of Parkinson's disease patients: A survey. *Expert Systems with Applications*, 219:119651, June 2023.

[137] Luca Parisi, Amir Zaernia, Renfei Ma, and Mansour Youseffi. m-ark-Support Vector Machine for Early Detection of Parkinson's Disease from Speech Signals. *International Journal of Mathematics and Computers in Simulation*, 15:34–41, April 2021.

[138] Nazmun Nahar, Ferdous Ara, Md Arif Istiek Neloy, Anik Biswas, Mohammad Shahadat Hossain, and Karl Andersson. Feature selection based machine learning to improve prediction of Parkinson disease. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12960 LNAI:496–508, 2021.

[139] Laetitia Jeancolas, Graziella Mangone, Dijana Petrovska-Delacrétaz, Habib Benali, Badr-Eddine Benkelfat, Isabelle Arnulf, Jean-Christophe Corvol, Marie Vidailhet, and Stéphane Lehéricy. Voice characteristics from isolated rapid eye movement sleep behavior disorder to early Parkinson's disease. *Parkinsonism & Related Disorders*, 95:86–91, February 2022.

[140] Muntasir Hoq, Mohammed Nazim Uddin, and Seung-Bo Park. Vocal Feature Extraction-Based Artificial Intelligent Model for Parkinson's Disease Detection. *Diagnostics*, 11(6):1076, June 2021.

[141] Jan Rusz, Jan Hlavnička, Michal Novotný, Tereza Tykalová, Amelie Pelletier, Jacques Montplaisir, Jean-Francois Gagnon, Petr Dušek, Andrea Galbiati, Sara Marelli, Paul C. Timm, Luke N. Teigen, Annette Janzen, Mahboubeh Habibi, Ambra Stefani, Evi Holzknecht, Klaus Seppi, Elisa Evangelista, Anna Laura

Rassu, Yves Dauvilliers, Birgit Högl, Wolfgang Oertel, Erik K. St. Louis, Luigi Ferini-Strambi, Evžen Růžička, Ronald B. Postuma, and Karel Šonka. Speech Biomarkers in Rapid Eye Movement Sleep Behavior Disorder and Parkinson Disease. *Annals of Neurology*, 90(1):62–75, July 2021.

[142] Mehran Sahandi Far, Simon B Eickhoff, Maria Goni, and Juergen Dukart. Exploring Test-Retest Reliability and Longitudinal Stability of Digital Biomarkers for Parkinson Disease in the m-Power Data Set: Cohort Study. *Journal of Medical Internet Research*, 23(9):e26608, September 2021.

[143] Jan Rusz, Tereza Tykalová, Michal Novotný, Evžen Růžička, and Petr Dušek. Distinct patterns of speech disorder in early-onset and late-onset de-novo Parkinson's disease. *npj Parkinson's Disease*, 7(1):98, December 2021.

[144] Michal Šimek and Jan Rusz. Validation of cepstral peak prominence in assessing early voice changes of Parkinson's disease: Effect of speaking task and ambient noise. *The Journal of the Acoustical Society of America*, 150(6):4522–4533, December 2021.

[145] Qian Yu, Xiaoya Zou, Fengying Quan, Zhaoying Dong, Huimei Yin, Jinjing Liu, Hongzhou Zuo, Jiaman Xu, Yu Han, Dezhi Zou, Yongming Li, and Oumei Cheng. Parkinson's disease patients with freezing of gait have more severe voice impairment than non-freezers during "ON state". *Journal of Neural Transmission*, 129(3):277–286, March 2022.

[146] Amr Gaballah, Vijay Parsa, Daryn Cushnie-Sparrow, and Scott Adams. Improved Estimation of Parkinsonian Vowel Quality through Acoustic Feature Assimilation. *The Scientific World Journal*, 2021:1–11, July 2021.

[147] Wasifur Rahman, Sangwu Lee, Md Saiful Islam, Victor Nikhil Antony, Harshil Ratnu, Mohammad Rafayet Ali, Abdullah Al Mamun, Ellen Wagner, Stella Jensen-Roberts, Emma Waddell, Taylor Myers, Meghan Pawlik, Julia Soto, Madeleine Coffey, Aayush Sarkar, Ruth Schneider, Christopher Tarolli, Karlo Lizarraga, Jamie Adams, Max A Little, E Ray Dorsey, and Ehsan Hoque. Detecting Parkinson Disease Using a Web-Based Speech Task: Observational Study. *Journal of Medical Internet Research*, 23(10):e26305, October 2021.

[148] Ilias Tougui, Abdelilah Jilbab, and Jamal El Mhamdi. Analysis of smartphone recordings in time, frequency, and cepstral domains to classify Parkinson's disease. *Healthcare Informatics Research*, 26:274–283, 10 2020.

[149] Jan Rusz, Tereza Tykalová, Michal Novotný, David Zogala, Evžen Růžička, and Petr Dušek. Automated speech analysis in early untreated Parkinson's disease: Relation to gender and dopaminergic transporter imaging. *European Journal of Neurology*, 29(1):81–90, January 2022.

[150] Siddharth Arora and Athanasios Tsanas. Assessing Parkinson's Disease at Scale Using Telephone-Recorded Speech: Insights from the Parkinson's Voice Initiative. *Diagnostics*, 11(10):1892, October 2021.

[151] Md. Sakibur Rahman Sajal, Md. Tanvir Ehsan, Ravi Vaidyanathan, Shouyan Wang, Tipu Aziz, and Khondaker Abdullah Al Mamun. Telemonitoring Parkinson's disease using machine learning by combining tremor and voice analysis. *Brain Informatics*, 7(1):12, December 2020.

[152] Pawichaya Suphinnapong, Onanong Phokaewvarangkul, Nuttakorn Thubthong, Arporn Teeramongkonrasmee, Patnarin Mahattanasakul, Preeya Lorwattanapongsa, and Roongroj Bhidayasiri. Objective vowel sound characteristics and their relationship with motor dysfunction in Asian Parkinson's disease patients. *Journal of the Neurological Sciences*, 426:117487, July 2021.

[153] Ina Kodrasi and Herve Bourlard. Spectro-Temporal Sparsity Characterization for Dysarthric Speech Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1210–1222, 2020.

[154] Siddharth Arora, Christine Lo, Michele Hu, and Athanasios Tsanas. Smartphone Speech Testing for Symptom Assessment in Rapid Eye Movement Sleep Behavior Disorder and Parkinson's Disease. *IEEE Access*, 9:44813–44824, 2021.

[155] Jan Rusz, Jan Hlavnicka, Tereza Tykalova, Michal Novotny, Petr Dusek, Karel Sonka, and Evzen Ruzicka. Smartphone Allows Capture of Speech Abnormalities Associated With High Risk of Developing Parkinson's Disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(8):1495–1507, August 2018.

[156] Fatih Demir, Abdulkadir Sengur, Ali Ari, Kamran Siddique, and Mohammed Alswaitti. Feature Mapping and Deep Long Short Term Memory Network-Based Efficient Approach for Parkinson's Disease Diagnosis. *IEEE Access*, 9:149456–149464, 2021.

[157] Daniel Palacios-Alonso, Guillermo Melendez-Morales, Agustin Lopez-Arribas, Carlos Lazaro-Carrascosa, Andres Gomez-Rodellar, and Pedro Gomez-Vilda. MonParLoc: A Speech-Based System for Parkinson's Disease Analysis and Monitoring. *IEEE Access*, 8:188243–188255, 2020.

[158] Yuanyuan Liu, Nelly Penttilä, Tiina Ihalainen, Juulia Lintula, Rachel Convey, and Okko Räsänen. Language-Independent Approach for Automatic Computation of Vowel Articulation Features in Dysarthric Speech Assessment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2228–2243, 2021. arXiv:2108.06943 [cs, eess].

[159] Yanhao Xiong and Yaohua Lu. Deep Feature Extraction From the Vocal Vectors Using Sparse Autoencoders for Parkinson's Classification. *IEEE Access*, 8:27821–27830, 2020.

[160] Hakan Gunduz. Deep Learning-Based Parkinson's Disease Classification Using Vocal Feature Sets. *IEEE Access*, 7:115540–115551, 2019.

[161] Oliver Y. Chen, Florian Lipsmeier, Huy Phan, John Prince, Kirsten I. Taylor, Christian Gossens, Michael Lindemann, and Maarten de Vos. Building a Machine-Learning Framework to Remotely Assess Parkinson's Disease Using Smartphones. *IEEE Transactions on Biomedical Engineering*, 67(12):3491–3500, December 2020.

[162] Liaqat Ali, Ce Zhu, Zhonghao Zhang, and Yipeng Liu. Automated Detection of Parkinson's Disease Based on Multiple Types of Sustained Phonations Using Linear Discriminant Analysis and Genetically Optimized Neural Network. *IEEE Journal of Translational Engineering in Health and Medicine*, 7:1–10, 2019.

[163] Mittapalle Kiran Reddy and Paavo Alku. A Comparison of Cepstral Features in the Detection of Pathological Voices by Varying the Input and Filterbank of the Cepstrum Computation. *IEEE Access*, 9:135953–135963, 2021.

[164] Changqin Quan, Kang Ren, and Zhiwei Luo. A Deep Learning Based Method for Parkinson's Disease Detection Using Dynamic Features of Speech. *IEEE Access*, 9:10239–10252, 2021.

[165] Amira S. Ashour, Majid Kamal A. Nour, Kemal Polat, Yanhui Guo, Wafaa Alsaggaf, and Amira El-Attar. A Novel Framework of Two Successive Feature Selection Levels Using Weight-Based Procedure for Voice-Loss Detection in Parkinson's Disease. *IEEE Access*, 8:76193–76203, 2020.

[166] Michael Saxon, Ayush Tripathi, Yishan Jiao, Julie Liss, and Visar Berisha. Robust Estimation of Hypernasality in Dysarthria with Acoustic Model Likelihood Features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2511–2522, 2020.

[167] Hunkar C. Tunc, C. Okan Sakar, Hulya Apaydin, Gorkem Serbes, Aysegul Gunduz, Melih Tutuncu, and Fikret Gurgen. Estimation of Parkinson's disease severity using speech features and extreme gradient boosting. *Medical & Biological Engineering & Computing*, 58(11):2757–2773, November 2020.

[168] Xu Zhang, Yaming Wang, Liang Zhang, Bo Jin, and Hongzhe Zhang. Exploring unsupervised multivariate time series representation learning for chronic disease diagnosis. *International Journal of Data Science and Analytics*, November 2021.

[169] Jie Ma, Yuanfan Zhang, Yongming Li, Lang Zhou, Lingyun Qin, Yuwei Zeng, Pin Wang, and Yan Lei. Deep dual-side learning ensemble model for Parkinson speech recognition. *Biomedical Signal Processing and Control*, 69:102849, August 2021.

[170] Lizbeth Naranjo, Carlos J. Pérez, and Yolanda Campos-Roca. Monitoring Parkinson's disease progression based on recorded speech with missing ordinal responses and replicated covariates. *Computers in Biology and Medicine*, 134, 7 2021.

[171] Patricia Klobusiakova, Jiri Mekyska, Lubos Brabenec, Zoltan Galaz, Vojtech Zvoncak, Jan Mucha, Steven Z. Rapcsak, and Irena Rektorova. Articulatory network reorganization in Parkinson's disease as assessed by multimodal MRI and acoustic measures. *Parkinsonism & Related Disorders*, 84:122–128, March 2021.

[172] Hakan Gunduz. An efficient dimensionality reduction method using filter-based feature selection and variational autoencoders on Parkinson's disease classification. *Biomedical Signal Processing and Control*, 66:102452, April 2021.

[173] Gabriel Solana-Lavalle and Roberto Rosas-Romero. Analysis of voice as an assisting tool for detection of Parkinson's disease and its subsequent clinical interpretation. *Biomedical Signal Processing and Control*, 66:102415, April 2021.

[174] Yuchuan Liu, Yongming Li, Xiaoheng Tan, Pin Wang, and Yanling Zhang. Local discriminant preservation projection embedded ensemble learning based dimensionality reduction of speech data of Parkinson's disease. *Biomedical Signal Processing and Control*, 63:102165, January 2021.

[175] Biswajit Karan, Sitanshu Sekhar Sahu, Juan Rafael Orozco-Arroyave, and Kartik Mahto. Hilbert spectrum analysis for automatic detection and evaluation of Parkinson's speech. *Biomedical Signal Processing and Control*, 61:102050, August 2020.

[176] John M. Tracy, Yasin Özkanca, David C. Atkins, and Reza Hosseini Ghomi. Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson's disease. *Journal of Biomedical Informatics*, 104:103362, April 2020.

[177] Juan Camilo Vásquez-Correa, Cristian D. Rios-Urrego, Alice Rueda, Juan Rafael Orozco-Arroyave, Sri Krishnan, and Elmar Nöth. Articulation and Empirical Mode Decomposition Features in Diadochokinetic Exercises for the Speech Assessment of Parkinson's Disease Patients. In Ingela Nyström, Yanio Hernández Heredia, and Vladimir Milián Núñez, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, volume 11896, pages 688–696. Springer International Publishing, Cham, 2019. Series Title: Lecture Notes in Computer Science.

[178] Vladimir Despotovic, Tomas Skovranek, and Christoph Schommer. Speech Based Estimation of Parkinson's Disease Using Gaussian Processes and Automatic Relevance Determination. *Neurocomputing*, 401:173–181, August 2020.

[179] Moumita Pramanik, Ratika Pradhan, Parvati Nandy, Akash Kumar Bhoi, and Paolo Barsocchi. The ForEx++ based decision tree ensemble approach for robust detection of Parkinson's disease. *Journal of Ambient Intelligence and Humanized Computing*, February 2022.

[180] Rohit Lamba, Tarun Gulati, and Anurag Jain. A Hybrid Feature Selection Approach for Parkinson's Detection Based on Mutual Information Gain and Recursive Feature Elimination. *Arabian Journal for Science and Engineering*, January 2022.

[181] Biswajit Karan, Sitanshu Sekhar Sahu, Juan Rafael Orozco-Arroyave, and Kartik Mahto. Non-negative matrix factorization-based time-frequency feature extraction of voice signal for Parkinson's disease prediction. *Computer Speech & Language*, 69:101216, September 2021.

[182] Adolfo M. García, Tomás Arias-Vergara, Juan Vasquez-Correa, Elmar Nöth, Maria Schuster, Ariane E. Welch, Yamile Bocanegra, Ana Baena, and Juan R. Orozco-Arroyave. Cognitive Determinants of Dysarthria in Parkinson's Disease: An Automated Machine Learning Approach. *Movement Disorders*, 36(12):2862–2873, December 2021.

[183] Mehedi Masud, Parminder Singh, Gurjot Singh Gaba, Avinash Kaur, Roobaea Alrobaea Alghamdi, Mubarak Alrashoud, and Salman Ali Alqahtani. CROWD: Crow Search and Deep Learning based Feature Extractor for Classification of Parkinson's Disease. *ACM Transactions on Internet Technology*, 21(3):1–18, June 2021.

[184] Biswajit Karan and Sitanshu Sekhar Sahu. An improved framework for Parkinson's disease prediction using Variational Mode Decomposition-Hilbert spectrum of speech signal. *Biocybernetics and Biomedical Engineering*, 41(2):717–732, April 2021.

[185] Atiqur Rahman, Sanam Shahla Rizvi, Aurangzeb Khan, Aaqif Afzaal Abbasi, Shafqat Ullah Khan, and Tae-Sun Chung. Parkinson's Disease Diagnosis in Cepstral Domain Using MFCC and Dimensionality Reduction with SVM Classifier. *Mobile Information Systems*, 2021:1–10, March 2021.

[186] Francesco Cavallieri, Carla Budriesi, Annalisa Gessani, Sara Contardi, Valentina Fioravanti, Elisa Menozzi, Serge Pinto, Elena Moro, Franco Valzania, and Francesca Antonelli. Dopaminergic Treatment Effects on Dysarthric Speech: Acoustic Analysis in a Cohort of Patients With Advanced Parkinson's Disease. *Frontiers in Neurology*, 11:616062, February 2021.

[187] Attila Zoltán Jenei, Gábor Kiss, Miklós Gábriel Tulics, and Dávid Sztahó. Separation of Several Illnesses Using Correlation Structures with Convolutional Neural Networks. *Acta Polytechnica Hungarica*, 18(7):47–66, 2021.

[188] N.P. Narendra, Bjorn Schuller, and Paavo Alku. The Detection of Parkinson's Disease From Speech Using Voice Source Information. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1925–1936, 2021.

[189] Tao Zhang, Yajuan Zhang, Hao Sun, and Haoran Shan. Parkinson disease detection using energy direction features based on EMD from voice signal. *Biocybernetics and Biomedical Engineering*, 41(1):127–141, January 2021.

[190] Jinee Goyal, Padmavati Khandnor, and Trilok Chand Aseri. A Comparative Analysis of Machine Learning classifiers for Dysphonia-based classification of Parkinson's Disease. *International Journal of Data Science and Analytics*, 11(1):69–83, January 2021.

[191] Ibrahim Karabayir, Samuel M. Goldman, Suguna Pappu, and Oguz Akbilgic. Gradient boosting for Parkinson's disease diagnosis from voice recordings. *BMC Medical Informatics and Decision Making*, 20(1):228, December 2020.

[192] Kemal Polat and Majid Nour. Parkinson disease classification using one against all based data sampling with the acoustic features from the speech signals. *Medical Hypotheses*, 140:109678, July 2020.

[193] Thomas Arias-Vergara, J. Camillo Vásquez-Correa, and J. Rafael Orozco-Arroyave. Parkinson's Disease and Aging: Analysis of Their Effect in Phonation and Articulation of Speech. *Cognitive Computation*, 9(6):731–748, December 2017.

[194] Carlos M. Travieso, Jesús B. Alonso, J.Rafael Orozco-Arroyave, J. Fracisco Vargas-Bonilla, Elmar Nöth, and Antonio G. Ravelo-García. Detection of different voice diseases based on the nonlinear characterization of speech signals. *Expert Systems with Applications*, 82:184–195, October 2017.

[195] Achraf Benba, Abdelilah Jilbab, and Ahmed Hammouch. Voice assessments for detecting patients with neurological diseases using PCA and NPCA. *International Journal of Speech Technology*, 20(3):673–683, September 2017.

[196] Betul Erdogdu Sakar, Gorkem Serbes, and C. Okan Sakar. Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease. *PLOS ONE*, 12(8):e0182428, August 2017.

[197] Lizbeth Naranjo, Carlos J. Pérez, and Jacinto Martín. Addressing voice recording replications for tracking Parkinson's disease progression. *Medical & Biological Engineering & Computing*, 55(3):365–373, March 2017.

[198] Elmehdi Benmalek, Jamal Elmhamdi, and Abdelilah Jilbab. Multiclass classification of Parkinson's disease using different classifiers and LLBFS feature selection algorithm. *International Journal of Speech Technology*, 20(1):179–184, March 2017.

[199] Lizbeth Naranjo, Carlos J. Pérez, Jacinto Martín, and Yolanda Campos-Roca. A two-stage variable selection and classification approach for Parkinson's disease detection by using voice recording replications. *Computer Methods and Programs in Biomedicine*, 142:147–156, April 2017.

[200] Jan Hlavnička, Roman Čmejla, Tereza Tykalová, Karel Šonka, Evžen Růžička, and Jan Rusz. Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder. *Scientific Reports*, 7(1):12, December 2017.

[201] Pedro Gómez, Jiri Mekyska, Andrés Gómez, Daniel Palacios, Victoria Rodellar, and Agustín Álvarez. Monitoring Parkinson Disease from speech articulation kinematics. *Loquens*, 4(1):036, December 2017.

[202] Ye N. Zhang. Can a Smartphone Diagnose Parkinson Disease? A Deep Neural Network Method and Telediagnosis System Implementation. *Parkinson's Disease*, 2017:1–11, 2017.

[203] Qi Wei Oung, Hariharan Muthusamy, Shafriza Nisha Basah, Hoileong Lee, and Vikneswaran Vijean. Empirical Wavelet Transform Based Features for Classification of Parkinson's Disease Severity. *Journal of Medical Systems*, 42(2):29, February 2018.

[204] Cüneyt Yücelbaş. A new approach: information gain algorithm-based k-nearest neighbors hybrid diagnostic system for Parkinson's disease. *Physical and Engineering Sciences in Medicine*, 44(2):511–524, June 2021.

[205] Abdellah Kacha, Christophe Mertens, Francis Grenez, Sabine Skodda, and Jean Schoentgen. On the harmonic-to-noise ratio as an acoustic cue of vocal timbre of Parkinson speakers. *Biomedical Signal Processing and Control*, 37:32–38, August 2017.

[206] Razieh Sheibani, Elham Nikookar, and SeyedEnayatollah Alavi. An ensemble method for diagnosis of Parkinson's disease based on voice measurements. *Journal of Medical Signals & Sensors*, 9(4):221, 2019.

[207] Liang Zhang, Yue Qu, Bo Jin, Lu Jing, Zhan Gao, and Zhanhua Liang. An Intelligent Mobile-Enabled System for Diagnosing Parkinson Disease: Development and Validation of a Speech Impairment Detection System. *JMIR Medical Informatics*, 8(9):e18689, September 2020.

[208] Siddharth Arora, Ladan Baghai-Ravary, and Athanasios Tsanas. Developing a large scale population screening tool for the assessment of Parkinson's disease using telephone-quality voice. *The Journal of the Acoustical Society of America*, 2871, 2019.

[209] Sanjana Singh and Wenyao Xu. Robust Detection of Parkinson's Disease Using Harvested Smartphone Voice Data: A Telemedicine Approach. *Telemedicine and e-Health*, 26(3):327–334, March 2020.

[210] John Prince, Fernando Andreotti, and Maarten De Vos. Multi-Source Ensemble Learning for the Remote Prediction of Parkinson's Disease in the Presence of Source-Wise Missing Data. *IEEE Transactions on Biomedical Engineering*, 66(5):1402–1411, May 2019.

[211] Jan Hlavnička, Tereza Tykalová, Olga Ulmanová, Petr Dušek, Dana Horáková, Evžen Růžička, Jiří Klempíř, and Jan Rusz. Characterizing vocal tremor in progressive neurological diseases via automated acoustic analyses. *Clinical Neurophysiology*, 131(5):1155–1165, May 2020.

[212] Milos Cernak, Juan Rafael Orozco-Arroyave, Frank Rudzicz, Heidi Christensen, Juan Camilo Vásquez-Correa, and Elmar Nöth. Characterisation of voice quality of Parkinson's disease using differential phonological posterior features. *Computer Speech & Language*, 46:196–208, November 2017.

[213] Markus Brückl, Alain Ghio, and François Viallet. Measurement of Tremor in the Voices of Speakers with Parkinson's Disease. *Procedia Computer Science*, 128:47–54, 2018.

[214] Savitha S. Upadhya and A.N. Cheeran. Discriminating Parkinson and Healthy People Using Phonation and Cepstral Features of Speech. *Procedia Computer Science*, 143:197–202, 2018.

[215] Savitha S. Upadhya, Alice Cheeran, and Jagannath Nirmal Nirmal. Thomson Multitaper MFCC and PLP voice features for early detection of Parkinson disease. *Biomedical Signal Processing and Control*, 46:293–301, September 2018.

[216] Kebin Wu, David Zhang, Guangming Lu, and Zhenhua Guo. Learning acoustic features to detect Parkinson's disease. *Neurocomputing*, 318:102–108, November 2018.

[217] Brittany R. Burk and Christopher R. Watts. The Effect of Parkinson Disease Tremor Phenotype on Cepstral Peak Prominence and Transglottal Airflow in Vowels and Speech. *Journal of Voice*, 33(4):580.e11–580.e19, July 2019.

[218] Patricia Gillivan-Murphy, Nick Miller, and Paul Carding. Voice Tremor in Parkinson's Disease: An Acoustic Study. *Journal of Voice*, 33(4):526–535, July 2019.

[219] Pablo Rodríguez-Pérez, Rubén Fraile, Miguel García-Escrig, Nicolás Sáenz-Lechón, Juana M. Gutiérrez-Arriola, and Víctor Osma-Ruiz. A transversal study of fundamental frequency contours in parkinsonian voices. *Biomedical Signal Processing and Control*, 51:374–381, May 2019.

[220] Elmehdi Benmalek, Jamal Elmhamdi, and Abdelilah Jilbab. Multiclass classification of Parkinson's disease using cepstral analysis. *International Journal of Speech Technology*, 21(1):39–49, March 2018.

[221] Zoltan Galaz, Jiri Mekyska, Vojtech Zvoncak, Jan Mucha, Tomas Kiska, Zdenek Smekal, Ilona Eliasova, Martina Mrackova, Milena Kostalova, Irena Rektorova, Marcos Faundez-Zanuy, Jesus Alonso-Hernandez, and Pedro Gomez-Vilda. Changes in Phonation and Their Relations with Progress of Parkinson's Disease. *Applied Sciences*, 8(12):2339, November 2018.

[222] Salim Lahmiri, Debra Ann Dawson, and Amir Shmuel. Performance of machine learning methods in diagnosing Parkinson's disease based on dysphonia measures. *Biomedical Engineering Letters*, 8(1):29–39, February 2018.

[223] Laureano Moro-Velázquez, Jorge Andrés Gómez-García, Juan Ignacio Godino-Llorente, Jesús Villalba, Juan Rafael Orozco-Arroyave, and Najim Dehak. Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's Disease. *Applied Soft Computing*, 62:649–666, January 2018.

[224] Luca Parisi, Narrendar RaviChandran, and Marianne Lyne Manaog. Feature-driven machine learning to improve early diagnosis of Parkinson's disease. *Expert Systems with Applications*, 110:182–190, November 2018.

[225] Savitha S. Upadhya and Alice N. Cheeran. Performance comparison of regression techniques in predicting Parkinson's disease severity score using speech features. *Biomedical Engineering: Applications, Basis and Communications*, 30(04):1850025, August 2018.

[226] Savitha S. Upadhya and Alice N. Cheera. Investigation of pitch and noise features extracted from voice samples of healthy and Parkinson affected people using statistical tests. *Journal of Engineering Science and Technology*, 13:13, 2018.

[227] Achraf Benba, Abdelilah Jilbab, Sara Sandabad, and Ahmed Hammouch. Voice signal processing for detecting possible early signs of Parkinson's disease in patients with rapid eye movement sleep behavior disorder. *International Journal of Speech Technology*, 22(1):121–129, March 2019.

[228] Pedro Gómez-Vilda, Zoltan Galaz, Jiri Mekyska, José M. Ferrández Vicente, Andrés Gómez-Rodellar, Daniel Palacios-Alonso, Zdenek Smekal, Ilona Eliasova, Milena Kostalova, and Irena Rektorova. Vowel Articulation Dynamic Stability Related to Parkinson's Disease Rating Features: Male Dataset. *International Journal of Neural Systems*, 29(02):1850037, March 2019.

[229] Fredrik Karlsson and Lena Hartelius. How Well Does Diadochokinetic Task Performance Predict Articulatory Imprecision? Differentiating Individuals with Parkinson's Disease from Control Subjects. *Folia Phoniatrica et Logopaedica*, 71(5-6):251–260, 2019.

[230] Salim Lahmiri and Amir Shmuel. Detection of Parkinson's disease based on voice patterns ranking and optimized support vector machine. *Biomedical Signal Processing and Control*, 49:427–433, March 2019.

[231] Antonio Suppa, Giovanni Costantini, Francesco Asci, Pietro Di Leo, Mohammad Sami Al-Wardat, Giulia Di Lazzaro, Simona Scalise, Antonio Pisani, and Giovanni Saggio. Voice in Parkinson's Disease: A Machine Learning Study. *Frontiers in Neurology*, 13:831428, February 2022.

[232] Juan Camillo Vásquez-Correa, Jordi Serra, Juan Rafael Orozco-Arroyave, J. Fracisco Vargas-Bonilla, and Elmar Nöth. Effect of acoustic conditions on algorithms to detect Parkinson's disease from speech. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 5065–5069, 2017.

[233] Liaqat Ali, Zhiquan He, Wenming Cao, Hafiz Tayyab Rauf, Yakubu Imrana, and Md Belal Bin Heyat. MMDD-Ensemble: A Multimodal Data–Driven Ensemble Approach for Parkinson's Disease Detection. *Frontiers in Neuroscience*, 15:754058, November 2021.

[234] Jefferson S. Almeida, Pedro P. Rebouças Filho, Tiago Carneiro, Wei Wei, Robertas Damaševičius, Rytis Maskeliūnas, and Victor Hugo C. de Albuquerque. Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques. *Pattern Recognition Letters*, 125:55–62, July 2019.

[235] Ilias Tougui, Abdelilah Jilbab, and Jamal El Mhamdi. Analysis of Smartphone Recordings in Time, Frequency, and Cepstral Domains to Classify Parkinson's Disease. *Healthcare Informatics Research*, 26(4):274–283, October 2020.

[236] Lizbeth Naranjo, Carlos J. Pérez, and Yolanda Campos-Roca. Monitoring Parkinson's disease progression based on recorded speech with missing ordinal responses and replicated covariates. *Computers in Biology and Medicine*, 134:104503, July 2021.

[237] Zoltan Galaz, Jiri Mekyska, Zdenek Mzourek, Zdenek Smekal, Irena Rektorova, Ilona Eliasova, Milena Kostalova, Martina Mrackova, and Dagmar Berankova. Prosodic analysis of neutral, stress-modified and rhymed speech in patients with Parkinson's disease. *Computer Methods and Programs in Biomedicine*, 127:301–317, April 2016.

[238] Hunkar C. Tunc, C. Okan Sakar, Hulya Apaydin, Gorkem Serbes, Aysegul Gunduz, Melih Tutuncu, and Fikret Gurgen. Estimation of Parkinson's disease severity using speech features and extreme gradient boosting. *Medical and Biological Engineering and Computing*, 58(11):2757–2773, 2020.

[239] Federica Amato, Valerio Cesarini, Luca Pietrosanti, Giovanni Costantini, Gabriella Olmo, and Giovani Saggio. Hallmarks of parkinson's disease progression determined by temporal evolution of speech attractors in the reconstructed phase-space. *2023 IEEE International Workshop on Metrology for Industry 4.0 and IoT, MetroInd4.0 and IoT 2023 - Proceedings*, pages 270–274, 2023.

[240] A.C. Lindgren, M.T. Johnson, and R.J. Povinelli. Speech recognition using reconstructed phase space features. 1:I–60–3, 2003.

[241] James D. Gardiner, Julia Behnsen, and Charlotte A. Brassey. Alpha shapes: determining 3D shape complexity across morphologically diverse structures. *BMC Evolutionary Biology*, 18(1):184, December 2018.

[242] Pedro Gómez-Vilda, Andrés Gómez-Rodellar, Daniel Palacios-Alonso, Victoria Rodellar-Biarge, and Agustín Álvarez Marquina. The role of data analytics in the assessment of pathological speech: A critical appraisal. *Applied Sciences*, 12(21), 2022.

[243] Giovanni Costantini, Valerio Cesarini, Pietro Di Leo, Federica Amato, Antonio Suppa, Francesco Asci, Antonio Pisani, Alessandra Calculli, and Giovanni Saggio. Artificial intelligence-based voice assessment of patients with Parkinson's disease off and on treatment: Machine vs. deep-learning comparison. *Sensors*, 23, 2 2023.

[244] Bo Chen, Heung Kou, Bowen Hou, and Yanbing Zhou. Music feature extraction method based on internet of things technology and its application. *Computational Intelligence and Neuroscience*, 2022, 2022.

[245] Markus Brückl. Vocal tremor measurement based on autocorrelation of contours. *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, 1:714–717, 2012.

[246] Athanasios Tsanas, Max A. Little, and Patrick Mcsharry. A methodology for the analysis of medical data. *Handbook of Systems and Complexity in Health*, pages 797–813, 1 2013.

[247] Patrícia Pinho, Larissa Monteiro, Maria Francisca de Paula Soares, Lorena Tourinho, Ailton Melo, and Ana Caline Nóbrega. Impact of levodopa treatment in the voice pattern of Parkinson's disease patients: A systematic review and meta-analysis. *Codas*, 30:1–7, 2018.

[248] World Health Organization. Regional Office for Europe. Who european regional obesity : Report 2022. Technical report, World Health Organization, 2022.

[249] Xihua Lin and Hong Li. Obesity: Epidemiology, pathophysiology, and therapeutics. *Frontiers in Endocrinology*, 12, 9 2021.

[250] Steven B. Heymsfield and Thomas A. Wadden. Mechanisms, pathophysiology, and management of obesity. *New England Journal of Medicine*, 376:254–266, 1 2017.

[251] Xavier Formiguera and Ana Cantón. Obesity: Epidemiology and clinical aspects. *Best Practice and Research: Clinical Gastroenterology*, 18(6 SPEC.ISS.):1125–1146, 2004.

[252] Andrew S. Jackson, Philip Stanforth, Jacques Gagnon, T Rankinen, Arthur Leon, D C Thirupathi Rao, James S Skinner, Claude Bouchard, and J H Wilmore. The effect of sex, age and race on estimating percentage body fat from body mass index: The Heritage Family Study. *International Journal of Obesity*, 26(6):789–796, 2002.

[253] Caroline M Apovian Jamy D Ard Anthony G Comuzzie-Karen A Donato Frank B Hu Van S Hubbard John M Jakicic Robert F Kushner Catherine M Loria Barbara E Millen Cathy A Nonas F Xavier Pi-Sunyer June Stevens Victor J Stevens Thomas A Wadden Bruce M Wolfe Susan Z Yanovski Harmon S Jordan Karima A Kendall Linda J Lux Roycelynn Mentor-Marcel Laura C Morgan Michael G Trisolini Janusz Wnek Jeffrey L Anderson Jonathan L

Halperin Nancy M Albert Biykem Bozkurt Ralph G Brindis Lesley H Curtis David DeMets Judith S Hochman Richard J Kovacs E Magnus Ohman Susan J Pressler Frank W Sellke Win-Kuang Shen Sidney C Smith Jr Gordon F Tomaselli; American College of Cardiology/American Heart Association Task Force on Practice Guidelines; Obesity Society Michael D Jensen, Donna H Ryan. 2013 AHA/ACC/TOS guideline for the management of overweight and obesity in adults: A report of the American College of cardiology/American Heart Association task force on practice guidelines and the obesity society. *Circulation*, 129(25 SUPPL. 1):102–138, 2014.

[254] World Health Organization. Who technical report series obesity: Preventing and managing the global epidemic. Technical report, World Health Organization, 2000.

[255] David E. Arterburn, Dana A. Telem, Robert F. Kushner, and Anita P. Courcoulas. Benefits and risks of bariatric surgery in adults: A review. *JAMA - Journal of the American Medical Association*, 324:879–887, 9 2020.

[256] Bruce M Wolfe, Elizaveta Kvach, and Robert H Eckel. Treatment of obesity. *Circulation Research*, 118:1844–1855, 5 2016.

[257] Joel E. Richter and Joel H. Rubenstein. Presentation and epidemiology of gastroesophageal reflux disease. *Gastroenterology*, 154:267–276, 1 2018.

[258] John Maret-Ouda, Sheraz R. Markar, and Jesper Lagergren. Gastroesophageal reflux disease a review. *JAMA - Journal of the American Medical Association*, 324:2536–2547, 12 2020.

[259] Catiele Antunes, Abdul Aleem, and Sean A. Curtis. *Gastroesophageal Reflux Disease*. StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing, 2022.

[260] Paul Chang and Frank Friedenberg. Obesity & GERD. *Gastroenterology Clinics of North America*, 43(1):161–173, 2014.

[261] Philip O. Katz, Lauren B. Gerson, and Marcelo F. Vela. Guidelines for the diagnosis and management of gastroesophageal reflux disease. *American Journal of Gastroenterology*, 108:308–328, 2013.

[262] Federica Amato, Maria Fasani, Glauco Raffaelli, Valerio Cesarini, Gabriella Olmo, Nicola Di Lorenzo, Giovanni Costantini, and Giovanni Saggio. Obesity and gastro-esophageal reflux voice disorders: a machine learning approach. *2022 IEEE International Symposium on Medical Measurements and Applications, MeMeA 2022 - Conference Proceedings*, 2022.

[263] A. T. Murugan and Gulshan Sharma. Obesity and respiratory diseases. *Chronic Respiratory Disease*, 5(4):233–242, 2008.

[264] Abdul-Latif Hamdan, Randa Al-Barazi, Dollen Tabri, Rami Saade, Issa Kutkut, Solara Sinno, and Jihad Nassar. Relationship between acoustic parameters and body mass analysis in young males. *Journal of Voice*, 26(2):144–147, 2012.

[265] Gonzàlez Julio. Correlations between speaker ' s body size and acoustic parameters of voice. *Perceptual and motor skills*, 105(1):1–11, 2007.

[266] Maria Gabriela Bernardo da Cunha, Gustavo Haruo Passerotti, Raimar Weber, Bruno Zilberstein, and Ivan Cecconello. Voice feature characteristic in morbid obese population. *Obesity Surgery*, 21(3):340–344, 2011.

[267] M. G.Manisha Milani, Murugaiya Ramashini, and Murugiah Krishani. A real-time application to detect human voice disorders. *2020 International Conference on Decision Aid Sciences and Application, DASA 2020*, pages 979–984, 2020.

[268] Serdar Akyildiz, Fatih Ogut, Ahmet Varis, Tayfun Kirazli, and Serhat Bor. Impact of laryngeal findings on acoustic parameters of patients with laryngopharyngeal reflux. *ORL; journal for oto-rhino-laryngology and its related specialties*, 74(4):215–219, 2012.

[269] Laureano Moro-Velázquez, Jorge Andrés Gómez-García, Juan Ignacio Godino-Llorente, Jesús Villalba, Juan Rafael Orozco-Arroyave, and Najim Dehak. Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's Disease. *Applied Soft Computing Journal*, 62:649–666, 2018.

[270] Sumitha Ramanathan and Gopi Ayyasamy. Voice Disorders and Reflux Disease – A Prospective Study. *International Journal of Scientific Study*, 8(10):8–11, 2021.

[271] Jong Cheol Shin, Julia Kim, and Diana Grigsby-Toussaint. Mobile phone interventions for sleep disorders and sleep quality: Systematic review. *JMIR Mhealth Uhealth*, 2017.

[272] Danan Gu, Jessica Sautter, Robin Pipkin, and Yi Zeng. Sociodemographic and health correlates of sleep quality and duration among very old chinese. *Sleep*, 33(5):601–610, 2010.

[273] Natalie L Hauglund, Chiara Pavan, and Maiken Nedergaard. Cleaning the sleeping brain–the potential restorative function of the glymphatic system. *Current Opinion in Physiology*, 15:1–6, 2020.

[274] Lulu Xie, Hongyi Kang, Qiwu Xu, Michael J Chen, Yonghong Liao, Meenakshisundaram Thiyagarajan, John O'Donnell, Daniel J Christensen, Charles Nicholson, Jeffrey J Iliff, et al. Sleep drives metabolite clearance from the adult brain. *science*, 342(6156):373–377, 2013.

[275] Francesc X Gamez-Oliva Margaret Thorogood Ngianga-Bakwin Kandala Saverio Stranges, William Tigbe. Sleep problems: An emerging global epidemic? findings from the indepth who-sage study among more than 40,000 older adults from 8 countries across africa and asia. Technical report, World Health Organization, 2012.

[276] Jessica Vensel Rundo and Ralph Downey III. Polysomnography. *Handbook of clinical neurology*, (160):381–392, 2019.

[277] Daniel J Buysse, Charles F Reynolds III, Timothy H Monk, Susan R Berman, and David J Kupfer. The pittsburgh sleep quality index: a new instrument for psychiatric practice and research. *Psychiatry research*, 28(2):193–213, 1989.

[278] Michal Icht, Gil Zukerman, Shir Hershkovich, Tal Laor, Yuval Heled, Nir Fink, and Leah Fostick. The "Morning Voice": The Effect of 24 Hours of Sleep Deprivation on Vocal Parameters of Young Adults. *Journal of Voice*, 34(3):489.e1–489.e9, 2018.

[279] Samuel Kim, Namhee Kwon, Henry O'Connell, Nathan Fisk, Scott Ferguson, and Mark Bartlett. How are you? Estimation of anxiety, sleep quality, and mood using computational voice analysis. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2020-July:5369–5373, 2020.

[280] M. Catarina Botelho, Isabel Trancoso, Alberto Abad, and Teresa Paiva. Speech as a biomarker for obstructive sleep apnea detection. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5851–5855, 2019.

[281] Janet S Carpenter and Michael A Andrykowski. Psychometric evaluation of the pittsburgh sleep quality index. *Journal of psychosomatic research*, 45(1):5–13, 1998.

[282] Sakari Lemola, Thomas Ledermann, and Elliot M Friedman. Variability of sleep duration is related to subjective sleep quality and subjective well-being: an actigraphy study. *PloS one*, 8(8):e71292, 2013.

[283] World Health Organization. World health statistics 2023 monitoring health for the sdgs sustainable development goals health for all. Technical report, 2023.

[284] Benjamin Taylor, Hyacinth M. Irving, Fotis Kanteres, Robin Room, Guilherme Luiz Guimaraes Borges, Cheryl J.Stephens Cherpitel, Thomas Kennedy Greenfield, and Jurgen T. Rehm. The more you drink, the harder you fall: A systematic review and meta-analysis of how acute alcohol consumption and injury or collision risk increase together. *Drug and Alcohol Dependence*, 110:108–116, 7 2010.

[285] Benjamin Taylor and Jürgen Rehm. The relationship between alcohol consumption and fatal motor vehicle injury: High risk at low alcohol levels. *Alcoholism: Clinical and Experimental Research*, 36:1827–1834, 10 2012.

[286] Cheryl J. Cherpitel. Alcohol and injuries: A review of international emergency room studies since 1995. *Drug and Alcohol Review*, 26:201–214, 3 2007.

[287] World Health Organization. Global status report on alcohol and health 2018. Technical report, World Health Organization, 2018.

[288] Alan W. Jones. Alcohol, its absorption, distribution, metabolism, and excretion in the body and pharmacokinetic calculations. *WIREs Forensic Science*, 1, 9 2019.

[289] Jeongeun Hyun, Jinsol Han, Chanbin Lee, Myunghee Yoon, and Youngmi Jung. Pathophysiological aspects of alcohol metabolism in the liver. *International Journal of Molecular Sciences*, 22, 6 2021.

[290] Szymon Paprocki, Meha Qassem, and Panicos A. Kyriacou. Review of ethanol intoxication sensing technologies and techniques. *Sensors*, 22, 9 2022.

[291] Alan W. Jones. Alcohol, its analysis in blood and breath for forensic purposes, impairment effects, and acute toxicity. *WIREs Forensic Science*, 1, 11 2019.

[292] Becky T Davies and Charles K Bowen. Estimation of peal blood alcohol concentration in research and highway safety. *Annu Proc Assoc Adv Automot Med*, 43:251–2634, 1999.

[293] World Health Organization. Global status report on road safety 2018. Technical report, World Health Organization, 2018.

[294] Alcohol interlock installation facilitation eu. Technical report, accessed May 10, 2023.

[295] Björn Schuller, Stefan Steidl, Anton Batliner, Florian Schiel, and Jarek Krajewski. The interspeech 2011 speaker state challenge *. *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 2011.

[296] Florian Schiel. Perception of alcoholic intoxication in speech. *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 2011.

[297] Dong-Yan Huang, Shuzhi Sam Ge, and Zhengchen Zhang. Speaker state classification based on fusion of asymmetric simpls and support vector machines. *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 2011.

[298] Daniel Bone, Matthew P Black, Ming Li, Angeliki Metallinou, Sungbok Lee, and Shrikanth S Narayanan. Intoxicated speech detection by fusion of speaker normalized hierarchical features and gmm supervectors. *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 2011.

[299] Claude Montacié and Marie-José Caraty. Combining multiple phoneme-based classifiers with audio feature-based classifier for the detection of alcohol intoxication. *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 2011.

[300] Rok Gajšek, Simon Dobrišek, and France Mihelič. University of ljubljana system for interspeech 2011 speaker state challenge. *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 2011.

[301] Florian Schiel, Christian Heinrich, Sabine Barfüsse, and Th Gilg. Alcohol and speech. 1997.

[302] Florian Hönig, Anton Batliner, and Elmar Nöth. Does it groove or does it stumble-automatic classification of alcoholic intoxication using prosodic features. *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 2011.

[303] Fadi Biadsy, William Yang Wang, Andrew Rosenberg, and Julia Hirschberg. Intoxication detection using phonetic, phonotactic and prosodic cues. *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 2011.

[304] Stefan Ultes, Alexander Schmitt, and Wolfgang Minker. Attention, sobriety checkpoint! can humans determine by means of voice, if someone is drunk.. . and can automatic classifiers compete? *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 2011.

[305] Albino Nogueiras Rodríguez. An hmm-based approach to the interspeech 2011 speaker state challenge. *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 2011.

[306] Abraham Albert Bonela, Zhen HE, Aiden Nibali, Thomas Norman, Peterg Miller, and Emmanuel Kuntsche. Audio-based deep learning algorithm to identify alcohol inebriation (adlaia). *Alcohol*, 12 2022.

[307] Benjamin Sertolli, Zhao Ren, Björn W. Schuller, and Nicholas Cummins. Representation transfer learning from deep end-to-end speech recognition networks for the classification of health states from speech. *Computer Speech and Language*, 68, 7 2021.

[308] Yue Zhang, Felix Weninger, and Björn W. Schuller. Cross-domain classification of drowsiness in speech: The case of alcohol intoxication and sleep deprivation. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017-August:3152–3156, 2017.

[309] Deep neural networks with batch speaker normalization for intoxicated speech detection. *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2019*, pages 1323–1327, 11 2019.

[310] Florian Schiel, Christian Heinrich, and Sabine Barfüsser. The first public corpus of alcoholized german speech. *Lang Resources Evaluation*, 46:503–521, 2011.

[311] Yaroslav Ganin, Evgeniya Ustinova, Pascal Germain Hana Ajakan, Hugo Larochelle, François Laviolette, and Mario Marchand a d Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17:1–35, 2016.

[312] Zhong Meng, Jinyu Li, Zhuo Chen, Yang Zhao, Vadim Mazalov, Yifan Gong, and Biing-Hwang Juang. Speaker-invariant training via adversarial learning. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5969–5973, 2018.

[313] Antoine de Mathelin, Franccois Deheeger, Guillaume Richard, Mathilde Mougeot, and Nicolas Vayatis. Adapt : Awesome domain adaptation python toolbox. *ArXiv*, abs/2107.03049, 2021.