UNIVERSIDAD DE EXTREMADURA

TESIS DOCTORAL

Análisis y mejora de la robustez de sistemas automáticos de ayuda al diagnóstico de enfermedades detectables por voz

Robustness analysis and improvement of automatic diagnostic aid systems for detectable-by-voice diseases

Mario Madruga Escalona

Programa de Doctorado interuniversitario en Tecnología Aeroespacial: Ingenierías Electromagnética, Electrónica, Informática y Mecánica

Conformidad del director y codirectora:

Fdo: Dr. Carlos Javier Pérez Sánchez Fdo: Dra. Yolanda Campos Roca

Esta tesis cuenta con la autorización del director y codirectora de la misma y de la Comisión Académica del programa. Dichas autorizaciones constan en el Servicio de la Escuela Internacional de Doctorado de la Universidad de Extremadura.

2023

Dedicado a María Agustina y Celestino.

Acknowledgements

Thanks to Dr. Carlos J. Pérez and Dr. Yolanda Campos for their advise, support, and hard work, which allowed me to finish this work. I would also like to thank all the teachers and professors who led me to this moment, specially to Dr. Carmen Ortiz for encouraging me to always go further.

I would like to thank Dr. Moreno for his medical advising, Sandra Paniagua and Esther de la O. for recording part of the speech database on organic voice disorders, and Prof. Gómez for advising and allowing the use of specialized acoustic material. It is acknowledged the Otorhinolaryngology Unit of the Hospital San Pedro de Alcántara (Cáceres) as well as all the voluntary individuals who participated in the experiments.

I would also like to thank Rosa Muñoz for registering the complementary information about the Parkinson's disease participants as well as providing her neurological advising, and Diego Santiago for recording part of the Parkinson's disease speech database. It is also acknowledged the Regional Association for Parkinson's Disease of Extremadura and the patients and healthy people who voluntarily participated in this study.

Finally, thanks to my parents, from whom I learnt to strive and persevere, for giving me the tools to navigate life.

This Doctoral Thesis has been financially supported by:

- Ministerio de Ciencia e Innovación and Agencia Estatal de Investigación: projects MTM2017-86875-C3-2-R and PID2021-122209OB-C32.

- Ministerio de Universidades: FPU18/03274 predoctoral grant.

- Junta de Extremadura: projects IB16054, GR18108, GR18055, GR21057, and GR21072.

- European Union: projects IB16054, GR18108, GR18055, GR21057, GR21072, MTM2017-86875-C3-2-R, and PID2021-122209OB-C32.

Resumen

El objeto de esta tesis es el estudio de sistemas automáticos de ayuda al diagnóstico, concretamente los dirigidos a enfermedades que afectan a la producción vocal del sujeto considerando patologías orgánicas de las cuerdas vocales (nódulos, pólipos y edema de Reinke) y enfermedades de origen neurológico, como la enfermedad de Parkinson (EP). El objetivo principal de estos sistemas es la detección automática de las variaciones que dichas enfermedades inducen en la producción vocal mediante el análisis y procesado de señal.

A partir de la grabación de la voz de un sujeto, un conjunto de algoritmos analiza la serie temporal que representa la grabación. Como resultado de este análisis de señal se obtiene un conjunto de características numéricas, cada una de las cuales representa un aspecto concreto de la grabación. Suponiendo que existan diferencias significativas entre voces sanas y patológicas, y que esas diferencias se reflejen en dichas características numéricas, se puede entrenar un sistema de aprendizaje automático como herramienta de ayuda en el proceso de diagnóstico.

Este enfoque presenta muchas ventajas en el entorno clínico. La prueba es fácil de realizar, no es invasiva, tiene bajo coste, y los resultados se obtienen rápidamente. Sin embargo, también presenta el inconveniente de que la transición desde el entorno de investigación a una situación clínica real puede resultar complicada debido a diversos factores.

En estos sistemas, el canal de comunicación tiene un gran impacto en la predicción. Desde la producción vocal, y hasta la forma de onda en bruto obtenida, cada elemento del canal tiene influencia en el resultado. La suma de todas estas contribuciones es lo suficientemente grande como para que las grabaciones tomadas bajo las mismas condiciones constituyan un subdominio del grupo objetivo: El conjunto de todas las grabaciones vocales, tanto sanas como afectadas por la enfermedad que se quiere diagnosticar. Por lo tanto, entrenar un sistema de aprendizaje automático utilizando grabaciones obtenidas en condiciones específicas induce un sesgo en el modelo.

En este trabajo se ha analizado el impacto que tiene el canal (ruido, dispositivos de grabación) en la detección automática de enfermedades detectables a través de grabaciones de voz. El entrenamiento multicondición se propone como una posible solución a este problema. Entrenar el clasificador con un conjunto de grabaciones obtenidas en una variedad de entornos ayuda a evitar el sesgo del modelo. Para probarlo se han estudiado diferentes elementos del canal de comunicación de forma aislada. Todos ellos provocan individualmente un sesgo en el modelo. Para todos ellos, la estrategia de entrenamiento multicondición ha demostrado ser una solución válida.

Para evaluar la capacidad de generalización de los resultados se han usado varias bases de datos, que contienen grabaciones reales. Se han utilizado datos públicos o con acuerdo de transferencia, así como datos propios recogidas con la colaboración del Servicio de Otorrinolaringología del Hospital San Pedro de Alcántara (Cáceres) y de la Asociación Regional de Parkinson de Extremadura (Cáceres y Mérida). La utilización de la base de datos mPower ha permitido estudiar también el impacto en los resultados de detección automática de la EP de la ausencia de supervisión cualificada en las grabaciones.

Además de la variabilidad que introduce el canal, se ha tenido en cuenta la variabilidad intra-sujeto, proponiendo métodos de regularización que tienen en cuenta esta variabilidad para mejorar el funcionamiento de un sistema automático de detección del edema de Reinke.

Palabras clave

Características acústicas
Diagnóstico asistido por ordenador
Entrenamiento multicondición
Aprendizaje automático
Enfermedad de Parkinson
Alteraciones orgánicas de la voz

Summary

This thesis subject of study consists on automatic diagnostic aid systems, specifically those aiming to detect diseases that affect the subject vocal production, considering both organic pathologies of the vocal folds (nodules, polyps, and Reinke's edema), and diseases of neurological origin, such as Parkinson's disease (PD). The main goal of these systems is the automatic detection of the variations that those diseases induce in the vocal production by means of signal analysis and processing.

On the basis of a subject's vocal recording, a collection of algorithms analyze the time series representing the recording. The outcome of those analysis is a set of numerical features, each one of them representing a specific aspect of the recording. Assuming that there exist significant differences between healthy and pathological voices, and that those differences are reflected in those numerical features, machine learning systems can be trained to help in the diagnostic process.

This approach shows many advantages in the clinical environment. The test is easy to perform, noninvasive, low cost, and the results are quickly obtained. However, it also has the disadvantage that the transition from the research environment to a realistic clinical setup can be difficult due to several factors.

In these systems, the communication channel has a large impact on the prediction. Between the vocal production and the raw waveform obtained, every element of the channel has an influence. The sum of all these contributions is large enough to allow recordings taken under the same conditions to constitute a subdomain of the target group: The set of every vocal recording, both healthy or affected by the disease being assessed. Therefore, training a machine learning system using recordings obtained under specific conditions induce a model bias.

The present work has analyzed the impact that the channel (noise, recording devices) has in the automatic detection of diagnosable by voice diseases. Multicondition training technique is proposed as a solution to this problem. Training the classifier with a recording set showing a variety of conditions helps avoiding the bias problem. Different elements of the communication channel have been isolated. All of them individually lead to model bias. For all of them, multicondition training strategy has proven to be a valid technique to overcome this problem.

Different databases, containing real voice recordings, have been used to evaluate de generalizability of the results. Public databases or databases with transfer agreement have been considered, as well as in-house collected databases built in collaboration with *Servicio de Otorrinolaringología del Hospital San Pedro de Alcántara* (Cáceres), and *Asociación Regional de Parkinson de Extremadura* (Cáceres and Mérida). mPower database has allowed to consider the impact that the absence of professional supervision in the recording process has in the automatic detection of PD outcome.

In addition to variability induced by the channel, intra-subject variability has also been considered. Regularization methods that improve the machine learning system performance by handling the intra-subject variability have been proposed, and successfully improved the performance of an automatic Reinke's edema detection system.

Keywords

Acoustic features
Computer aided diagnosis (CAD)
Multicondition training
Machine learning
Parkinson's disease
Organic voice disorders

Contents

Chapter 1

Introduction

1.1 Background

Since the proposal of machine learning techniques, parallel with the development of the first computer systems circa the decade of 1950's, a huge amount of research and development on these techniques has been conducted. The ever increasing computing capacity along with the recent increase in information availability has enabled the spread of these techniques. One of the topics of machine learning is the classification problem. Given a population composed of a set of categories, the objective is to identify which one new observations belong to.

The potential applications for these techniques are almost unlimited, and they have been widely used in many knowledge fields. One of most obvious candidates is medical diagnosis. There exist differences between healthy and pathological subjects caused by the disease. If these differences are measurable, the classification problem becomes clear. For each subject, the features that characterize the disease must be obtained. Those measurements feed a machine learning system that tries to predict the category the sample belongs to based on previous observations (Shehab et al., 2022).

Using this approach, research on many diseases has been conducted. Medical diagnosis counts with a wide range of tests of different nature. Image diagnosis, tissue and blood samples, or physiological measurements are some examples of useful features that have been considered in the scientific literature. Biomedical signal acquisition, and the subsequent processing and analysis, is one of the many tools that can be successfully used for medical diagnosis. Computer and electrical engineering tools developed for signal processing are commonly used in the biomedical field. For example, signals coming from electrocardiograms and electroencephalograms are widely used in this context. However, many other signals can be obtained from the physiological processes of the human body. One of these signals comes from voice production. Voice is the main communication tool for human beings, and its production involves complex physiological and neurological processes. As such, it is a good candidate to be used as a descriptor for diseases affecting those processes.

In the physiological side, modifications of the mechanical aspects of voice production will change its properties. Changes in the mechanical aspects of the vocal folds like mass or rigidity lead to alterations in vocal production. Reinke's space diseases share common pathologic features, although the lesions' etiologic factors differ. Hantzakos et al. (2009) described the differences between the usual conditions associated to Reinke's space. The main diagnostic methods involve direct examination of the vocal folds; Laryngoscopy and videostroboscopy performed by qualified otolaryngologists are the most usual techniques for diagnosis (Echternach et al., 2020).

On the other hand, neurological disorders can also affect vocal production in a significant yet different way. Parkinson's disease (PD) is a good example. PD is a neurodegenerative disorder which is usually classified as a motor function disease. People suffering from PD usually present bradykinesia, rigidity and tremor (Tysnes and Storstein, 2017). Its diagnosis is complex and many approaches have been proposed. Electroencephalograms (Oh et al., 2020), magnetic resonance imaging (Amoroso et al., 2018) or motion and gait analysis (Belić et al., 2019) have been considered, among others.

The proposed techniques for both physiological and neurological diseases share great disadvantages for detection and progression tracking of voice-detectable diseases. They are usually expensive, cumbersome, and require the expertise of a well trained specialist. Therefore they are not widely available and can delay the diagnostic process. For that reason, the search for new, fast, low-cost, noninvasive, and reliable tests is of great interest.

Voice analysis arises as a good candidate for that task. The affection of physiological diseases on the mechanical aspects of voice production are reflected on the vocal signal. Thus, voice analysis is a good candidate as Computer Aided Diagnostic (CAD) tool. On the neurological side, the motor function alteration produced by PD is also reflected on the vocal production. Pawlukowska et al. (2018) stated that 75%-95% of people with PD suffer from some sort of speech impairment. In this case, many aspects of vocal production can be studied, like the ability to sustain a steady voice production (Zhang et al., 2021), running speech analysis (Rahman et al., 2021), or diadochokinesis tests (Montaña et al., 2018). This makes voice analysis a good candidate for new developments on automatic diagnostic tools of PD.

Extensive research on automatic diagnostic aid tools has been made. A large number of machine learning techniques have been proposed and tested. Hidden Markov Models, Gaussian Mixture Models, Support Vector Machines, Random Forests, Artificial Neural Networks, or Deep Neural Networks are examples of recently applied techniques for the classification task. Input data selection for these techniques is another factor under research in the topic.

Although the target diseases for these systems share a common symptom like voice production alteration, the materialization of the effects is different attending to the specific disease. Physiological conditions mostly affect frequency related properties like fundamental frequency, or harmonic structure. Neurological diseases on their side can prevent the patient's pitch control, normal articulation, and prosodic abilities. PD patients usually refer speech impairment, which is beyond voice production. Therefore, each pathology requires a specific set of descriptors, or features, which also receive great research attention. Gómez-García et al. (2019a,b) and Hegde et al. (2019) offer an overview of both features and classification tools used in this area.

1.2 Objective

More than two decades after the first efforts in automatic diagnostic aid tools for detectable-by-voice diseases, it does not exist a reliable commercial application available yet. One of the possible reasons is population mismatch. Research is usually made by strictly controlling the setup so the number of variables is minimized. In this case, the research teams usually want to fix every recording aspect like equipment, location, and even medication in the case of neurological disorders to find significant differences between groups in the analysis. Therefore, generalization is not an easy task.

The general purpose of this PhD thesis is studying and analyzing some of the factors that hinder the application in clinical environments of automatic diagnostic aid systems for detectable-by-voice diseases and propose solutions that help preventing them.

The main specific objectives of this PhD thesis are summarized as follows:

O1. Investigate, implement and test speech feature extraction algorithms that show efficient performance in computer aided diagnosis systems for detectable-by-voice diseases.

O2. Investigate variable selection and classification algorithms that provide efficient performance in the same application context.

O3. Analyze the impact of channel mismatch effects (noise, different recording devices) on the performance of computer aided diagnosis systems for detectable-by-voice diseases.

O4. Perform healthy/disease-affected discrimination experiments based on real voice recordings including diseases with organic and neurologic origin, and including the smartphone as a recording device in some experiments.

O5. Address intra-subject variability by proposing replication-based regularization approaches and test them to detect pathologies affecting speech.

O6. In healthy/disease-affected discrimination experiments, address result generalization by the use of different databases in a comparative way.

O7. Propose novel multicondition training based approaches to address channel mismatches and investigate the performance improvement versus single-condition training systems for the same applications.

1.3 Thesis development

This PhD thesis, presented by a compendium of publications, contains seven chapters, including the introduction and a chapter of summary of results, conclusions and future research. The main

body of the work (chapters 2-6) consists of five publications in journals indexed in the Journal of Citation Reports. This subsection provides a short overview of how these five articles contribute to the thesis goals stated in the previous subsection.

Suboptimal recording conditions affect the performance of the output machine learning model. The presence of noise makes it harder to perform the signal analysis step and therefore the subsequent steps are affected and performance is degraded (Gómez-García et al., 2021). Chapter 2 (Madruga et al., 2021a) studied the impact of noise in the accuracy of a machine learning system trained to detect some Reinke's space diseases, i.e. nodules, and Reinke's edema. It uses MEEI database (Massachusetts Eye & Ear Infirmary, 1994), a well known voice recordings database long considered as the test case for voice disease computer analysis, along with an at-home voice recording database collected in collaboration with the Hospital San Pedro de Alcántara (Cáceres). Noise is added to the original recordings and the consequent performance drop is measured. This article is linked to the specific objectives O1-O4 and O6.

Chapter 3 (Carrón et al., 2021) analyzed the diagnostic ability of systems built using non curated datasets. In this case, in addition to an at-home database from healthy and people suffering from Parkinson disease (collected in collaboration with the Asociación Regional de Parkinson de Extremadura, ARPE), mPower recordings were used (Bot et al., 2016). Both databases were recorded in non-optimal conditions, by using a mobile phone as recording device. Furthermore, mPower dataset was recorded by volunteers using their own device in different locations unlike the at-home collected database which had a fixed setup. This article addressed the specific objectives O1-O4 and O6.

Recording conditions are not the only perturbation for CAD systems though. Biological systems are complex and their behavior is not purely deterministic. Chapter 4 (Naranjo et al., 2021a) studied the variability that occurs when a subject repeats an experiment of vocal production recording and the influence that it makes in the results obtained when training a machine learning model. It takes into consideration regularization methods addressing the intra-subject variability. This article is related to the specific objectives O1, O2, O4, and O5.

Articles in chapters 2 and 3 show that CAD systems are relatively tolerant to environmental noise and changes in recording conditions, whereas article in chapter 4 shows that intra-subject variability can be addressed using the appropriate methodology. Therefore, it might be possible to build systems which are robust against any source of variability in the recording process. This problem is known in machine learning research as domain adaptation (Kouw and Loog, 2021). The circumstances surrounding the sampling process lead to a bias in the model. Therefore, if a model is fed with samples from a different population, i.e. recorded in a different environment, this leads to lower performance than that obtained for the original experimental population.

Multicondition training has been proposed to address the variability problem in the speaker recognition context (Garcia-Romero et al., 2012). By mixing different environmental conditions the system avoids specialization and the resulting model improves in generalization. Chapter 5 (Madruga et al., 2021b) proposes a methodology that takes variability in the recording environment into account, in the context of CAD systems. Three physiological diseases in Reinke's space are considered: Nodules, polyps and Reinke's disease. In this case, four different databases are studied: MEEI database (Massachusetts Eye & Ear Infirmary, 1994), an at-home collected database (in collaboration with *Hospital San Pedro de Alcántara, Cáceres)*, Saarbrücken Voice Database (SVD) (Barry and Pützer, 2016), and *Hospital Universitario Príncipe de Asturias* (HUPA) database (Godino-Llorente et al., 2008). Environmental noise was isolated as a variability source and its influence on model performance measured for each database is evaluated and compared. Specific objectives O1-O4, O6, and O7 are linked to this article.

Environmental noise is not the only source of variability. One often ignored factor in the data acquisition process is the recording equipment. Article in chapter 6 (Madruga et al., 2023) focused on the impact of changing the recording device used in the experiments. Different smartphone models were considered and the performance was compared. The positions of the recording device were also analyzed. Multicondition training was proposed and applied for subjects suffering PD from ARPE. The related specific objectives for this article are O1-O4 and O6-O7.

The work developed for this thesis shows that one of the obstacles CAD systems are facing in their implementation as valid diagnostic tools is the lack of generalization. This research shows that noise is introduced in several stages of the recording process and affects the models created. This leads to a bias that prevents these models from being useful under circumstances other than those used for their training stages. The proposed methodology of multicondition training has proven to be a good solution for this problem as it improves generalization in all the scenarios studied.

Chapter 2

Impact of noise on the performance of automatic systems for vocal fold lesions detection

Title:

Impact of noise on the performance of automatic systems for vocal fold lesions detection

Authors and affiliation:

Mario Madruga[a], Yolanda Campos-Roca[b], Carlos J. Pérez[a]

[a]Universidad de Extremadura, Departamento de Matemáticas, Spain
[a]Universidad de Extremadura, Departamento de Tecnología de los Computadores y las Comunicaciones, Spain

Journal:

Biocybernetics and Biomedical Engineering

DOI:

10.1016/j.bbe.2021.07.001

Abstract: Automatic voice condition analysis systems have been developed to automatically discriminate pathological voices from healthy ones in the context of two disorders related to exudative lesions of Reinke's space: nodules and Reinke's edema. The systems are based on acoustic features, extracted from sustained vowel recordings. Reduced subsets of features have been obtained from a larger set by a feature selection algorithm based on Whale Optimization in combination with Support Vector Machine classification. Robustness of the proposed systems is assessed by adding noise of two different types (synthetic white noise and actual noise recorded in a clinical environment) to corrupt the speech signals. Two speech databases were used for this investigation: the Massachusetts Eye and Ear Infirmary (MEEI) database and a second one specifically collected in Hospital San Pedro de Alcántara (Cáceres, Spain) for the scope of this work (UEX-Voice database). The results show that the prediction performance of the detection systems appreciably decrease when moving from MEEI to a database recorded in more realistic conditions. For both pathologies, the prediction performance declines under noisy conditions, being the effect of white noise more pronounced than the effect of noise recorded in the clinical environment.

Keywords: Acoustic features, Computer aided diagnosis, Reinke's edema, Nodules, Noise robustness, Voice disorders.

BIOCYBERNETICS AND BIOMEDICAL ENGINEERING 41 (2021) 1039–1056

Available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/bbe

Original Research Article

Impact of noise on the performance of automatic systems for vocal fold lesions detection

Mario Madruga [a,b,*], *Yolanda Campos-Roca* [c], *Carlos J. Pérez* [a]

[a] *Departamento de Matemáticas, Universidad de Extremadura, Spain*
[b] *Facultad de Veterinaria, Avenida de la Universidad S/N, 10003 Cáceres, Cáceres, Spain*
[c] *Departamento de Tecnología de los Computadores y las Comunicaciones, Universidad de Extremadura, Spain*

ARTICLE INFO

Article history:
Received 22 December 2020
Received in revised form
25 June 2021
Accepted 4 July 2021
Available online 16 July 2021

Keywords:
Acoustic features
Computer aided diagnosis
Reinke's edema
Nodules
Noise robustness
Voice disorders

ABSTRACT

Automatic voice condition analysis systems have been developed to automatically discriminate pathological voices from healthy ones in the context of two disorders related to exudative lesions of Reinke's space: nodules and Reinke's edema. The systems are based on acoustic features, extracted from sustained vowel recordings. Reduced subsets of features have been obtained from a larger set by a feature selection algorithm based on Whale Optimization in combination with Support Vector Machine classification. Robustness of the proposed systems is assessed by adding noise of two different types (synthetic white noise and actual noise recorded in a clinical environment) to corrupt the speech signals. Two speech databases were used for this investigation: the Massachusetts Eye and Ear Infirmary (MEEI) database and a second one specifically collected in Hospital San Pedro de Alcántara (Cáceres, Spain) for the scope of this work (UEX-Voice database). The results show that the prediction performance of the detection systems appreciably decrease when moving from MEEI to a database recorded in more realistic conditions. For both pathologies, the prediction performance declines under noisy conditions, being the effect of white noise more pronounced than the effect of noise recorded in the clinical environment.

1. Introduction

Voice is a person's main communication tool and, therefore, the impact of voice disorders on quality of life can be substantial. For people involved in certain professions, such as teachers, singers, and many others, voice is also the main working tool and, as an immediate consequence, they are in a high risk of developing voice disorders due to excessive and/or incorrect use of their voices. Voice professionals are prone to suffer from organic diseases and will eventually need some kind of medical diagnosis and care [1]. Some of those voice disorders are exudative lesions of Reinke's space and are manifestations of different etiologic factors like voice abuse leading to nodules, or tobacco use linked to Reinke's Edema [2].

* Corresponding author at: Departamento de Matemáticas, Universidad de Extremadura, Spain.
E-mail addresses: mariome@unex.es (M. Madruga), ycampos@unex.es (Y. Campos-Roca), carper@unex.es (C.J. Pérez).
https://doi.org/10.1016/j.bbe.2021.07.001

The main methods used by otolaryngologists to diagnose laryngeal diseases are direct inspection of the larynx through the use of invasive techniques such as laryngoscopy and videostroboscopy [3], and/or evaluation of voice quality by hearing. The first group of diagnosis techniques causes discomfort to the patient and requires sophisticated equipment like endoscopic instruments or specialized video cameras, whereas the second group is subjective and strongly depends on the experience of the specialist [4].

In recent years, computer aided diagnosis (CAD) of voice disorders has attracted considerable scientific interest with the aim of providing an effective screening method for pathologies in an early stage. Using automatic voice condition analysis (AVCA) helps the physicians providing useful information in the differential diagnostic process [5]. These techniques usually consist of an acoustic feature extraction step followed by the application of machine learning algorithms under the assumption that voice quality is correlated with voice pathology [5]. Compared to the previously mentioned diagnosis methods, these techniques show the advantages that they are non-invasive, fast, objective, and low-cost. Also, acoustic analysis have been proved to be a sensitive, objective, and quantitative tool, being more accurate than perceptual assessment [6]. For example, they can be applied in preventive medicine to professionals at high risk of suffering from voice disorders [7]. Other contribution in the field of automatic detection of structural vocal-fold pathologies is [8], which offers experimental results of binary discrimination between normal and pathological voices, where the pathological voice class is composed of a variety of disorders. A recent scientific review on AVCA systems is provided by [9].

This paper focuses on laryngeal diseases, in particular, nodules and Reinke's edema. In this application context, the most usual vocal task is sustained phonation of /a/ vowels, where the speakers are asked to pronounce a vowel sound as steady as possible in terms of amplitude and fundamental frequency [10]. This vocal task has several advantages. First, it requires continuous motion of the vocal folds, which constitute the main structure involved in these pathologies. Also, this vocal task is quick and easy to perform and it is a common sound across different languages and accents. Sustained phonation of other vowel sounds or production of sentences have also been used [8].

Based on sustained vowel recordings, the studies in the literature consider many different characteristics of speech including perturbation measurements (such as jitter or shimmer), noise measures (such as harmonic-to-noise ratio (HNR) or glottal-to-noise excitation (GNE) ratio), Mel frequency cepstral coefficients (MFCCs), among others [5]. More recent studies show that nonlinear time series analysis methods may be more appropriate for pathological voices than classical measurements. Those methods, including Lyapunov exponents and correlation dimension, have been applied to classification of disordered voice samples [5].

There is also a variety of pattern recognition techniques based on supervised learning applied in this context in the scientific literature. Among the many different classification techniques that have been used, [5] highlights Support Vector Machines (SVM) and Gaussian Mixture Models as the most widely employed, although [9] compiles a much wider range of alternatives which have been used in this particular field. In general, when a large feature set size is used, the model becomes less comprehensible and there is a high risk of overfitting [11]. Therefore, for a reliable classification, it is important to use a small number of measurements, containing an optimal amount of information. Feature selection for classification is an active research area on its own, whose main objective is to reduce the dimension of the original feature set. Wrapper algorithms based on meta-heuristic optimization techniques allow to obtain a global optimum of the predictive accuracy achieved for a certain classification algorithm by using a simple and easy to implement concept. Among the different meta-heuristic approaches, the Whale Optimization Algorithm (WOA) is a recently proposed approach which mimics the hunting behavior of humpback whales. It was originally created as an optimization algorithm [12] and later adapted as a feature selection operator [13,14].

An important aspect to take into account regarding AVCA systems for speech disorder detection is robustness. When the recordings have been obtained under a controlled acoustic environment, the performance of these systems in real-life conditions remains unknown. A clear example is the Massachusetts Eye and Ear Infirmary (MEEI) database [15], whose recordings were taken in Kay Elemetrics and MEEI Voice and Speech Lab [16], being these conditions very difficult to reproduce in everyday situations. In order to be useful, it is required that these systems remain robust even when the recordings are captured in a non-controlled environment. Experiments have been carried on in order to assess different channels in remote disease monitoring [17,18]. Even mobile healthcare applications have been tested in controlled acoustical environments, like [19], which mentions that experiments are carried out in an as low as 30 dB background noise room. However, noise robustness is very seldom present in the scientific literature about automatic detection systems of organic voice disorders. [20] presents assumable noise levels of 25 dBA, 36 dBA, 30 dB, 40 dB and 50 dB for different studies, remarking that the maximum acceptable noise level was not investigated. [21] presents a study about the adverse effects of noise on voice quality measurement. This study focuses on fundamental frequency and perturbation measurements with no particular pathology addressed. [22] studies the numerical effects of noise on the computation of different acoustical features, although it does not test their classifying capabilities. [23] performs a preliminary study on the impact of noise on the automatic detection of a particular voice pathology: Reinke's edema. In the context of Parkinson's disease, [24] shows the impact of noise on an automatic detection system based on acoustic features. Finally, [25] proposes a technique to mitigate the possible differences in recording environments, characterized by different noise conditions.

The main goal of the present paper is to assess the negative effects of realistic noisy recording conditions on the outcome of an AVCA system for voice pathology detection. We have focused on two specific related diseases which are common vocal fold lesions, and their etiologies are related. However, we performed independent experiments with each disease in order to minimize the number of variables present in the study since the main goal is not building an automated

diagnostic system, but to check the potential effects of environmental noise on the outcomes of AVCA systems for vocal fold lesions detection.

We have built AVCA systems to discriminate pathological voices from healthy ones in the context of two structural organic speech pathologies: nodules and Reinke's edema. This work is a significant extension of the conference paper [23]. It introduces new case studies in a different pathology (vocal fold nodules) that allow to improve the generalization capability of the conclusions and to make a disease comparison. Also, it exposes a feature selection algorithm which has been designed, implemented, and tested for these applications. Specifically, the systems are built on reduced acoustic feature subsets, obtained by a feature selection algorithm based on Whale Optimization in combination with SVM classification. This algorithm has been implemented using parallel computing libraries and executed on a Beowulf cluster system. Two voice recording databases are employed: The first one is MEEI, recorded in the most favorable acoustic conditions; the second one is an own database, recorded in a more usual clinical environment. Also, system robustness is evaluated by adding two different types of noise (white Gaussian noise and actual clinical environment noise) to both databases and studying the impact on the discrimination capacity of each system.

2. Materials and methods

This section provides the main information on participants, collection and pre-processing of voice samples and noise recording. Also, the proposed feature extraction approach is summarized and the feature selection algorithm is explained.

2.1. *Participants*

MEEI database, commercialized by KayPentax Corp. [15], is one of the voice databases used for this work. This database, widely used for research in pathological voice classification, has been recorded under very strict acoustical and technical conditions (sound-proof booth, high-quality recording equipment, type of microphone, distance to the source...) [16]. It includes sustained /a/ recordings of 53 healthy and 657 pathological subjects, 19 and 25 of them suffering from vocal-fold nodules and Reinke's edemas, respectively.

Not all the voice samples in the MEEI database were recorded using the same technical parameters, being the healthy voices recorded at a sampling rate of 50 kHz, with a total length of 3 s, whereas the pathological voices were recorded at 25 kHz for one second. For the purpose of our experiments, all the waveforms were resampled when needed and trimmed so the whole database complies with the specifications of a sampling rate of 25 kHz and one second length.

An experiment has been conducted to collect a voice recording database (UEX-Voice) also based on sustained /a/ phonations. This database has been recorded in Hospital San Pedro de Alcántara (HSPdA), Cáceres, specifically, in an ordinary diagnostic room, with its door closed, providing only a certain isolation from the noisy aisles and waiting halls surrounding it.

All the recordings were taken using the same equipment: an AKG 520 head-worn condenser cardioid microphone attached to a TASCAM US322 sound card, being the recording software Audacity 2.0.5. The sampling rate was 44.1 kHz. Four phonations were recorded for each participant, of variable lengths depending on the capacity of each individual, so they were trimmed both at the beginning, ensuring no silence, and at the end, to obtain a uniform duration of one second. All the waveforms were downsampled to 25 kHz in order to match the sampling rate of MEEI database.

Fig. 1 shows the age distribution of the considered subjects with nodules and Reinke's edema from the MEEI and UEX-Voice databases. Summary statistics are provided in Table 1.

2.2. *Noise database*

A noise database has been specifically collected from the room where the research study took place. This room was placed in the external consultation area, on the second floor, of a hospital in a small town (population $<$ 100.000). Background noise was recorded using the same equipment previously defined. The length of the recording was 11 min 50 s and included noise from different sources: multitalker babble, cell phone sounds, fluorescent lighting, door closing, and footsteps, among others. Since post-processing is made altering recording level, we are more interested in the nature of sound than in its power. The recordings made include a realistic representation of the variety of indoor noise sources that are present in the outpatient clinic area of any hospital during consultation hours. Furthermore, national and regional environmental noise laws are very strict in hospital surroundings. Anyway, the impact of external sources on the final recordings is negligible, as in free space the received noise power is inversely proportional to the square of the distance to the source, given that external sources are farther away than internal ones, and accounting for the attenuation due to building walls. For those reasons, considering that the voice samples are at most 3 s long, and that they are trimmed down to one second, these noise recordings provided enough variability to perform all the desired experiments.

The noise waveforms were recorded inside the empty diagnosis room with door and windows closed while noise level was being measured using a certified Brüel & Kjaer 2260 sound level meter, what allows us to assert the acoustical environment recreated when using these recordings. Three one-minute measurements showed an A-weighted mean L_{eq} of 34.17 dBA.

2.3. *Feature extraction*

A total of 94 features were extracted from each voice sample. These features have been previously used in scientific literature, either for voice disease detection, Parkinson's disease detection, or other biomedical signal analysis [5,9]. The extraction methods were coded in Python by direct implementation of the formal mathematical definition, by translating existing code from other authors, or by using available libraries of proven reliability from Python repositories. A comprehensible list is provided in Table 2 including short name,

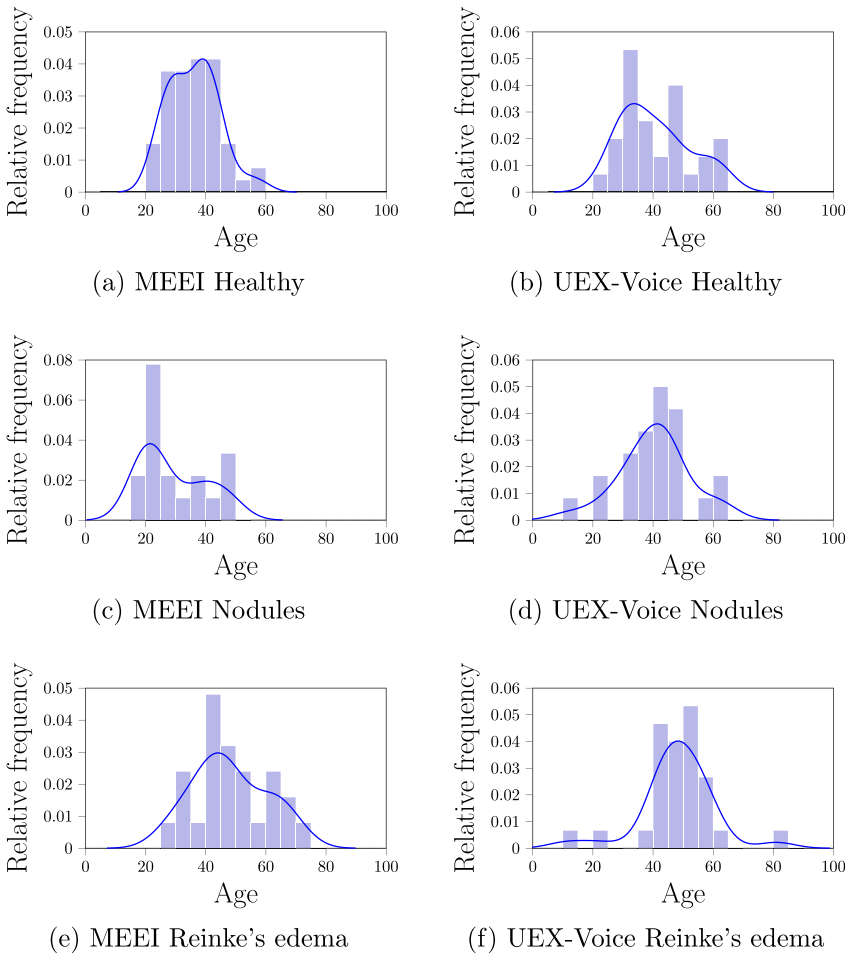


Fig. 1 – Age distribution of the subjects from MEEI and UEX-Voice databases.

Table 1 – Distribution of subjects by health status, sex and age.

		Sex		Age	
Database	Health status	Male	Female	Mean	Std. Dev.
MEEI	Healthy	21	32	36.00	8.29
	Nodules	1	17	29.11	10.45
	Reinke's edema	5	20	48.04	11.97
UEX-Voice	Healthy	4	26	40.76	11.18
	Nodules	1	23	40.41	11.33
	Reinke's edema	3	27	47.96	11.76

references to previous work, and variants taken into consideration.

Age and sex are two features inherent to each subject. Humans undergo several changes with aging that affect the voice production system. For example, changes in the larynx tend to alter the average fundamental frequency and to produce instability of vocal fold vibrations [34]. The impact is different for men than for women; in particular, fundamental frequency tends to increase in men and decrease in women due to some aging effects [35]. Also, women are more prone to suffer from organic voice diseases than men [36]. These and several other aspects related to the impact of age

Table 2 – Features extracted.

Short name	References	Full name and variants
		Linear
CPP	[26]	Cepstral peak prominence
GNE_X	[27]	Glottal-to-noise excitation ratio. Four different statistical features: mean, std, SNR_TKEO, SNR_SEO
GQ	[27]	Glottal quotient. Three statistics used: prc5_95, std cycle open, std cycle closed
HNR	[27]	Harmonic-to-noise ratio
JITTER_X	[27]	Jitter. Twenty-two different statistics used: abs_dif, diff_percent, PQ3_classical_Schoentgen, PQ3_classical_Baken, PQ3_generalised_Schoentgen, PQ5_classical_Schoentgen, PQ5_classical_Baken, PQ5_generalised_Schoentgen, PQ11_classical_Schoentgen, PQ11_classical_Baken, PQ11_generalised_Schoentgen, abs0th_perturb, DB, CV, TKEO_mean, TKEO_std, TKEO_prc5, TKEO_prc25, TKEO_prc75, TKEO_prc95, FM, range_5_95_perc
SHIMMER_X	[27]	Shimmer. Twenty-two different statistics used: abs_dif, diff_percent, PQ3_classical_Schoentgen, PQ3_classical_Baken, PQ3_generalised_Schoentgen, PQ5_classical_Schoentgen, PQ5_classical_Baken, PQ5_generalised_Schoentgen, PQ11_classical_Schoentgen, PQ11_classical_Baken, PQ11_generalised_Schoentgen, abs0th_perturb, DB, CV, TKEO_mean, TKEO_std, TKEO_prc5, TKEO_prc25, TKEO_prc75, TKEO_prc95, FM, range_5_95_perc
MFCC-X	[28]	Mel Frequency Cepstral Coefficient, 13 first coefficients MFCC0 - MFCC12
		Non-linear
D2	[10]	Correlation dimension
FMMI	[29]	First minimum in mutual information
FZCF	[29]	First zero of autocorrelation function
HURST	[10]	Hurst Exponent
MFSW	[30]	Multifractal spectrum width
ZCR	[31]	Zero crossing rate
		Entropies and complexities
PERMUTATION	[32]	Permutation entropy
PPE	[27]	Pitch period entropy
RPDE	[18]	Recurrence Period Density Entropy
SHANNON	[29]	Shannon entropy
LZ-X	[33]	Lempel–Ziv complexity. 16 features quantifying signal 2^1 to 2^{16} steps

and gender on speech have motivated the inclusion of these two features.

Many diseases affecting vocal production cause pitch-related alterations, specifically frequency or amplitude modulation, being sustained vocal analysis the most useful technique to apply [37]. Most studies till recent years focused their attention on acoustical features such as jitter or shimmer, which assume that voice production is a linear system. Though the definition of jitter seems very simple, i.e., the mean variation in the fundamental frequency of the phonation process, there is no method considered as standard for calculating such variation, mainly because the fundamental frequency calculation is not a trivial task. Most usual methods are provided by Multi-Dimensional Voice Program [15], the software tool provided by KayPentax with their database; and Praat suite. Other algorithms have been proposed, such as Sun's algorithm or SWIPE alternatives [38]. In our implementation jitter and shimmer were translated from MATLAB code given by [27]. We obtained 22 different measurements for both jitter and shimmer, each one corresponding to a different mathematical formulation.

Besides jitter and shimmer, other spectrum and fundamental frequency related linear features have been studied. GQ was originally used to monitor Parkinson's disease [39], and shortly after for early diagnosis of pathological voice [40]. GQ takes into account the lengths of time the glottis is open and closed. CPP was proposed as a measure of breathiness and our version was coded following the definition given by [26]. HNR is intended to assess voice hoarseness and tries to estimate the relationship between purely harmonic to turbulent noise in voice production. MFCCs try to describe the spectral components and do not require a previous pitch estimation [28].

Nonlinear behaviors have been shown to play a role in the voice production process and, particularly, in the case of voice pathologies [5]. Therefore, assuming that voice diseases may induce a chaotic behavior in human voice production, nonlinear analysis has also been taken into consideration in the search for new accurate features [7]. RPDE considers the uncertainty in signal cycle estimates using both an embedded space and entropy, being related to fundamental frequency, nonlinear, and entropy measurements [18]. ZCR is not properly a nonlinear measurement, but it is useful in time series analysis [31], measuring the number of times the signal crosses zero level. D2 is an estimator of the correlation dimension, a measure of self-similarity of chaotic systems [10]. HURST and MFSW are closely related: HURST, also known as detrended fluctuation analysis, used in [10], measures a monofractal local fluctuation of the root-mean-square in a time series, whereas MFSW [30] analyzes the q-order Hurst

exponent, or multifractal fluctuation analysis, capable of distinguish fast and slow fluctuations. FMMI measures the time lag for which the signal adds a maximum of information about itself, or for which the information redundancy is minimal [29]. FZCF gives the input lag for which the autocorrelation function is minimal [29].

Another aspect that has been considered is the signal entropy, or the amount of information carried by the signal. Different approaches can be found in the scientific literature: SHANNON is a classical communication theory measurement of the information a signal carries [29]; PERMUTATION adds a perspective of symbolic dynamics, or the temporal order of the values in a series [32]; PPE quantifies the lack of control over pitch beyond natural vibrato and microtremor [27,40]. Finally, LZ measures the regularity or repetitiveness of a sequence [33].

2.4. *Feature selection and classification*

We built different systems for each database-disease combination. Those systems were created using clean samples from the databases, and their ability to handle additive noise was checked by inducing different types of noise at different SNR levels. The systems creation comprised two steps: feature selection and recording classification.

Given the number of features considered and datasets sizes, the risk of overfitting is a relevant issue, whichever classifier is used. To avoid this inconvenience, the following feature selection approach has been designed and implemented.

In general, features belonging to the same family, that is, those which share a common base algorithm, are highly correlated within the group [41]. This is shown in Fig. 2, which represents a heat map of the Pearson correlation coefficient for each pair of features. It can be observed that jitter, shimmer, and LZ features are highly correlated within their groups. Therefore, prior to any WOA related computation, the feature set was reduced to keep only one feature per group in the case of these three families.

The number of features considered after discarding the highly correlated ones is still high compared to the number of individuals included in each database, so further feature selection is performed. We used WOA [12], a bio-inspired evolutionary algorithm properly modified as a wrapper feature selection operator [14], which has recently started to be tested as a feature selection method [13]. It mimics the bubble-net feeding in the hunting behavior of the humpback whales. These whales hunt close to the surface by creating a net of bubbles where the prey is trapped. The algorithm mimics this behavior in two phases: one of them is called exploitation, when a whale herd tries to encircle a prey (solution or, in this case, set of features) in a spiral bubble-net attack; the other phase, called exploration, searches randomly for a new prey.

In each iteration, the algorithm selects a prey, a local optimum point. WOA selection algorithm relies on the fitness function from Eq. (1)

$$f = \alpha \times (1 - accuracy) + \beta \times \frac{number\ of\ selected\ features}{number\ of\ features} \quad (1)$$

based on the accuracy of a given classifier and the number of features selected to train such classifier. In this case, the objective is to maximize the accuracy, that is, minimize the error rate through the α parameter while minimizing the relative number of features using the β parameter, thus decreasing the risk of overfitting due to an excessive number of features involved. Both accuracy and relative number of features are in the range $[0, 1]$, and α and β also range $[0, 1]$ being $\beta = 1 - \alpha$.

Exploitation or prey encircling is done by taking the local optimum point obtained in the previous iteration, or a random point at the beginning of the execution, and then each search agent or "whale" describes a spiral around that point. To create such spiral the whale alters the optimum point, consisting of a feature set, and modifies it by adding or removing features, ensuring a lower euclidean distance to the optimal point in each iteration, so the new candidate obtained by each search agent is always closer to the local optimum point at each iteration.

At this point, as suggested by [42], we changed the updating mechanisms. Feature addition or subtraction in the solutions is performed using Eq. (2),

$$\vec{X}(t+1) = D' \cdot e^{bl} \cdot \cos(2\pi l) + \vec{X}^*(t), \quad (2)$$

where b defines the spiral shape, l is a random number in $[-1, 1]$, D' is the euclidean distance to the best available solution and $\vec{X}^*$ is the best solution so far, as depicted in [12]. In this case, since the solutions space is discrete (a feature is either present or not) the updated leading position is transformed into a binary vector whose positions indicate whether the whale position in a given dimension or feature is above 0.5 or not (e.g. 4-dimensional solution [0.7, 0.3, 0.8, 0.9] would turn into [1, 0, 1, 1]).

In order to extend search to a wider portion of the solutions space, some of the whales will move randomly to another unrelated point in the space, what constitutes the exploration mechanism. Eventually, one or more whales will find a better solution than the temporary optimal one given by the last iteration, and then, all the search agents will turn to the best solution in terms of accuracy and feature number, and will start encircling it. The algorithm ends when it finds a solution with a fitness function lower than a given threshold, or when it has computed a maximum number of iterations.

The algorithm can be fine tuned by using tournament and roulette wheel selection mechanisms instead of a random operator to enhance the exploration phase, as well as crossover and mutation to optimize the exploitation phase [14]. In this case we have implemented the algorithm based on tournament selection as selection mechanism, and mutation as subset search.

Tournament selection randomly chooses two challengers within the search agents population and, according to a random number being greater than a given threshold, selects either the best or worst fitted candidate as new individual.

Mutation provides a tool for generating new possible solutions from actual solutions being considered in the current state. It randomly changes the state of some features from selected to un-selected or vice versa. The number of altered selections decreases as the algorithm reaches the hard limit of iterations, making it more prone to mutations at the beginning of the execution and more unlikely to mutate towards the end.

DIABETES RESEARCH AND CLINICAL PRACTICE 41 (2021) 1039–1056

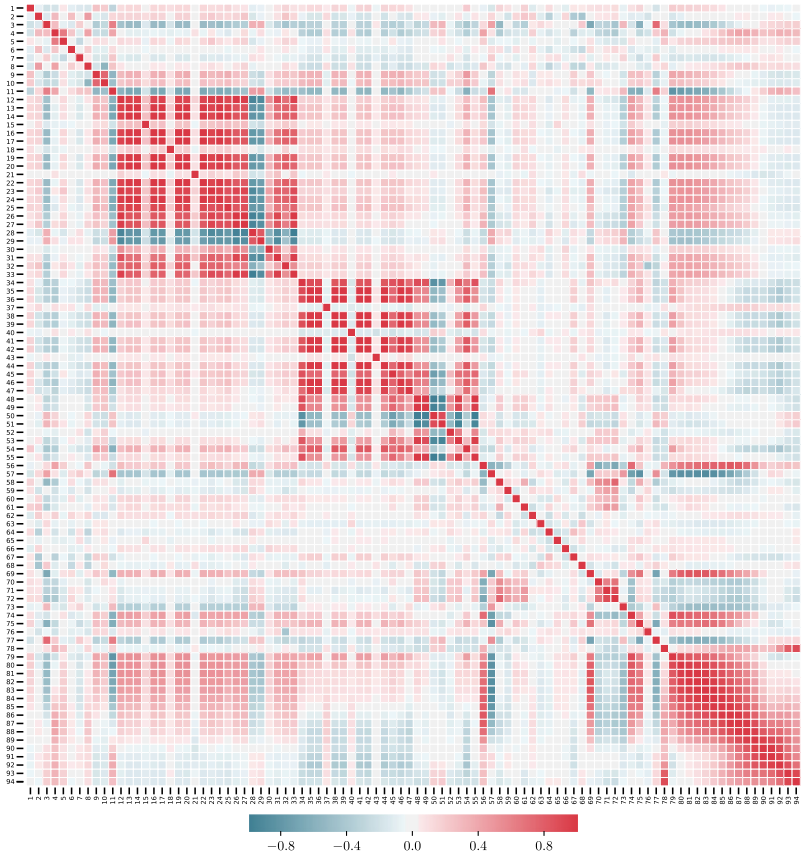


Fig. 2 – Correlation heat map for all the extracted features.

The algorithm also makes use of crossover, where two candidate solutions are mixed, or "bred", in order to create a new candidate solution with mixed characteristics of the two original ones.

The overall procedure is shown as Algorithm 1. It begins initializing candidates in the search field. In every iteration, until it reaches the hard limit imposed or achieves a fitness function above a desired threshold, it performs the following actions. First, an update of the a, A, C, l, p parameters is performed for each whale. a decreases linearly from 2 to 0 as the number of iterations get closer to the hard limit; A and C define the whale position update along with a: A is a coefficients vector built using the value of a and a random vector in [0, 1] and C is built using the same random vector; l is a random number in $[-1, 1]$ which defines the spiral shape as seen in Eq. 2; and p is a random number in [0, 1] whose value determines whether the whale is going to encircle the best solution (exploit) or it is going to explore, and how. Then, if it chooses to explore, it either explores the solutions space by performing a tournament selection or mutation of the best solution, creating a new candidate by crossover. If it chooses to exploit the current best solution, the process is completed by encircling in a spiral shaped curve the best solution.

Algorithm 1. Whale Optimization Algorithm

```
leaderScore = ∞
candidates = random(searchagents, features)
while leaderScore < threshold doanditerations < maxIterations
do
  for all candidates do
    Update a, A, C, l, p following [14]
    if p > 0.5 then
      if |A| > 1 then
        Xrand ⇐ tournament selection
        RA ⇐ mutate(Xrand)
        RE ⇐ mutate(candidate)
        candidateLeftarrowcrossover(RA, RE)
      else
        D ⇐ mutate(leaderPosition)
        candidate ⇐ crossover(D, candidate)
      end if
    else
      Encircle LeaderPosition using Eq. (2)
    end if
  end for
  for all candidates do
    if fitness(candidate) < leaderScore then
      leaderScore ⇐ fitness(candidate)
      leaderPosition ⇐ candidate
    end if
  end for
  iterations ⇐ iterations + 1
end while
```

For the classifier, SVM is considered. Prior to any computation a grid search is performed to find the best parameters for each database, and only for the case without additional noise (called "clean" case) as we intend to show the effects of noise on classification accuracy. The search space includes the kernel function used, among the four implemented ones in Python scikit library (linear, poly, rbf, and sigmoid) as well as their specific parameters.

Given the database sizes, one single run of WOA algorithm could yield a feature set fitted to the training set and the initial random conditions used for that particular run, reaching a local optimal point not suitable for most work settings, thus the need of multiple runs in order to generalize performance. Stratified shuffle and split was performed, all the selected feature sets were collected, and the most repeated features were compiled as the optimal feature set for each database and condition.

3. Results

Experiments were carried out to check the performance and robustness of different detection systems for two databases of voice disorders: nodules and Reinke's edema. This section describes the experimental setting and the main results obtained.

3.1. *Experimental setting*

The experiments consisted of two steps: first, classification systems were built minimizing the number of features used in each case (each database and each disease); then, those features were used to classify the same subjects from the databases they were created from, with and without added noise in the voice recordings, under a stratified repeated random subsampling validation framework.

As there are two heterogeneous databases to work with, some previous steps were taken in order to ensure a reliable results comparison. Most of them, concerning technical recording characteristics like sampling rate or recording length are summarized in Section 2.1. However, UEX-Voice database consists in four recordings per participant in the experiment. In this case, considered features were extracted for each recording, and mean value for each one was used in the following experiments.

In order to minimize the feature set, for each voice database (without artificially-added noise) the collection of results was obtained as follows. 640 instances of the WOA algorithm were launched, each one consisting of 640 whales, and a maximum of 25000 iterations to find the optimum features set. Preliminary studies were carried out to get values for α and β that both yield good accuracy and low feature set size. The values chosen for this specific problem were $\alpha = 0.99, \beta = 0.01$. The execution provided 640 different sets, represented by binary vectors of length 35, where 1 represents the presence of a feature, and 0 the absence of the feature in the set. By adding all the vectors as if they were natural vectors we end up with a total appearances vector.

The most useful features were used to train a set of classifiers, one set per disease, using an increasing number of features. They were incrementally added in the most repeated

order obtained by the WOA algorithm, and classifier performance was computed until no improvement was found for at least three feature additions. At this point, the feature set yielding the local maximum was chosen, so it was possible to check the evolution of the classifier F1 score with respect to the number of features used. Validation of results was performed by stratified repeated random subsampling, by repeating this procedure 1000 times and averaging the results. Training and test sets were selected using a stratified shuffle and split schema, so in each repetition the ratio of healthy and pathological individuals remained constant and identical to the ratio present for the database and disease being considered in each experiment. 2/3 of randomly chosen subjects from the database were used as training set and 1/3 as testing set.

In order to check the robustness of these systems, two different sources of additive noise were used: artificially generated Gaussian white noise and an actual recording of noise taken inside HSPdA. Two different scenarios were considered within each case: taking a random sample within noise recording by randomly selecting a starting point from the noise vector for every single voice recording in each database, and adding both noise sample and voice recording, what inherently introduces more variability in the process; and taking a random sample by randomly selecting a starting point within the noise vector and adding this unique sample to every recording in the database.

This process was repeated from a signal to noise power ratio (SNR) ranging from 0 dB to 30 dB in steps of 6 dB. Since the UEX-Voice database recording conditions are known, we can assume that the noise level present in the recording session is proportional to the value provided in Section 2.2, although we have no means to quantify the voice signal power. On the other hand, we are considering that the MEEI database was recorded in such good acoustical and technical conditions that the noise contained in the recordings is negligible, and as such will not alter significantly the induced SNR. For each SNR level considered, the same training and test sets for each run of the classifier were considered, so we can avoid the variability that random sets would induce in the different classifiers, so differences in the results obtained are a consequence only of the induced noise in each case.

3.2. *Experimental results*

The next subsections summarize the results obtained for each disease, once applied the different levels of noise and trained the set of classifiers. For each database and disease, four experiments were considered, which relate to a particular combination of noise nature (white synthetic noise, or actual recorded noise) and randomness (same noise clip added to every sample in the database, or one randomly generated or selected clip per sample).

We have used confusion matrix analytics to measure the performance of the final classifiers. True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) were computed for each iteration, and accuracy, precision, recall, and F1 score (Eqs. (3)–(6)) were derived from them, and were later averaged:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3}$$

$$precision = \frac{TP}{TP + FP}, \tag{4}$$

$$recall = \frac{TP}{TP + FN}, \tag{5}$$

$$F1\ score = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \tag{6}$$

In each of the four plots in Figs. 3–6, the X-axis shows which features have been selected and, as we move to the right, we add the features to the subset being considered, so the curve represents at a given point the mean F1 score on Y-axis, obtained after stratified repeated random subsampling validation. The upper limit for the number of features has been selected taking in consideration the shape of the clean curve in each case as stated in Section 3.1.

Each plot shows seven curves, one obtained after training the classifiers using the original recordings, no noise added, called *clean*, and six graphs labeled after SNR levels ranging from 0 dB to 30 dB in steps of 6 dB.

3.2.1. *Nodules*

Figs. 3, 4 and Tables 3, 4 show the mean F1 scores using an increasing number of features for voices affected with nodules in MEEI and UEX-Voice, respectively. In the case without noise addition, the results show that the classification F1 scores decrease from 0.91 to 0.61 when moving from MEEI database (Fig. 3)) to UEX-Voice database (Fig. 4). In this case, the implemented procedure has allowed to identify a reduced feature subset (4 or 5 features, depending on the database) that allows to achieve a saturation behavior in the prediction performance. In the case of UEX-Voice database, these features are CPP and three MFCCs, that is, cepstral and spectral features. For MEEI the most useful features resulted to be: MFCC1, CPP, HURST, AGE, and MFSW, which is a mixture of features based on linear and non-linear analysis, and the age.

F1 scores for MEEI database when adding 0 dB SNR noise are not even computable as we can not compute precision as well. This shows that the classifier marks every subject as healthy: Eq. (4) shows that, if there are neither TP nor FP (all the subjects classified as pathological), precision is not computable; also, by Eq. (5), recall equals 0. Looking at recall, which for binary classification shows the ability to detect pathological voices, for MEEI database we get values over 0.9 only for SNRs above 24 dB, and even then precision does not get over 0.9. In the case of UEX-Voice database F1 score, precision, and recall are lower, specially the latter.

The overall behavior results as expected, with higher SNR levels yielding better results, closer to the clean samples classifications. However, that behavior varies as we change the nature of the noise: Actual noise (subplots (a) and (b) in Figs. 3 and 4) tends to be less problematic when taking into consideration a few features, staying closer to the clean samples classifiers than synthetic noise (subplots (c) and (d) in Figs. 3 and 4).

3.2.2. *Reinke's edema*

Figs. 5, 6 and Tables 5, 6 show the discrimination results obtained in the case of Reinke's edema. The comparison between both databases in the case without additional noise

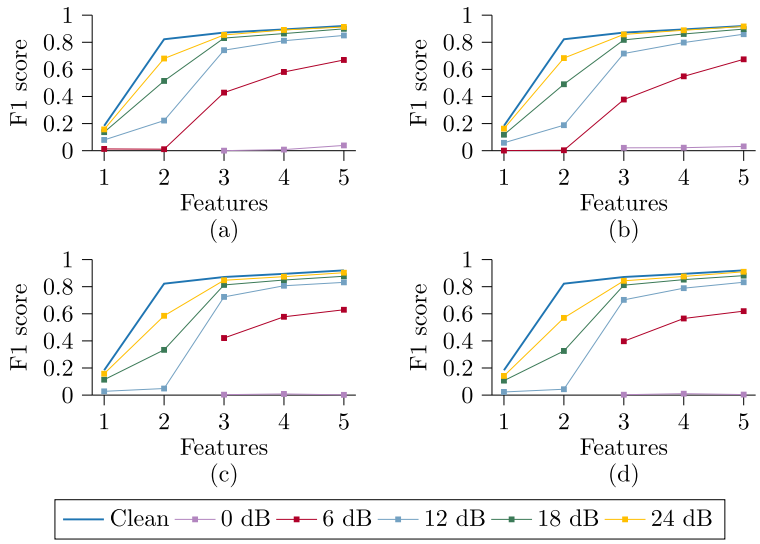


Fig. 3 – Mean F1 scores using cumulative features for nodules disease, MEEI database. SNRs ranging from 0 dB to 30 dB in steps of 6 dB. The features are: 1-MFCC1, 2-CPP, 3-HURST, 4-AGE, 5-MFSW. Noise characteristics: (a) realistic noise, fixed sample, (b) realistic noise, random sample, (c) white noise, fixed sample, (d) white noise, random sample.

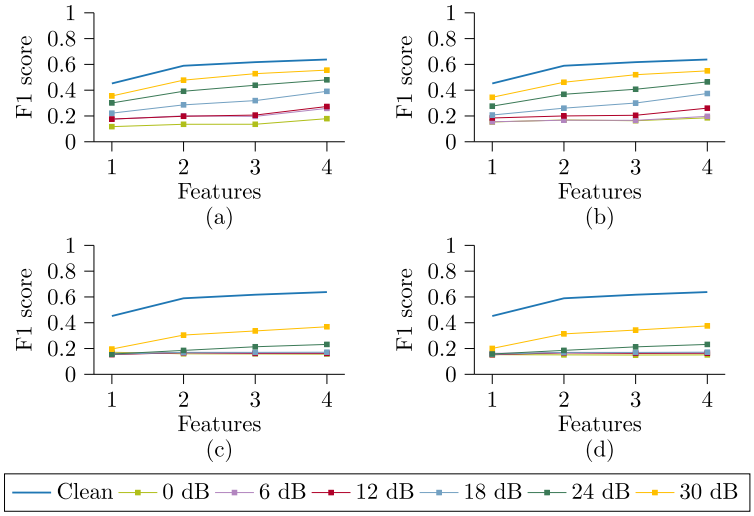


Fig. 4 – Mean F1 scores using cumulative features for nodules disease, UEX-Voice database. SNRs ranging from 0 dB to 30 dB in steps of 6 dB. The features are: 1-CPP, 2-MFCC7, 3-MFCC3, 4-MFCC2. Noise characteristics: (a) realistic noise, fixed sample, (b) realistic noise, random sample, (c) white noise, fixed sample, (d) white noise, random sample.

allows to extract similar conclusions than in the case of nodules. Again, the detection F1 score obtained with MEEI database is higher than in the case of UEX-Voice (0.98 versus 0.83). The number of features needed to reach a saturation behavior is 5 or 6, depending on the database. Cepstral and spectral features play again a relevant role, however an entropy feature is required in both feature subsets. In the case of MEEI, shimmer is also selected.

In the presence of additive noise, the detection performance decreases, and the impact is again higher in the case

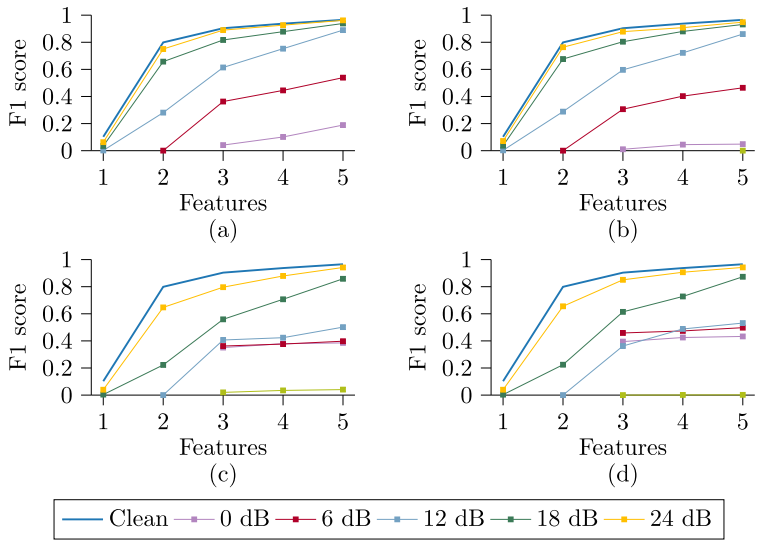


Fig. 5 – Mean F1 scores using cumulative features for Reinke's edema, MEEI database. SNRs ranging from 0 dB to 30 dB in steps of 6 dB. The features are: 1-MFCC1, 2-PERMUTATION, 3-Shimmer, 4-MFCC3, 5-CPP. Noise characteristics: (a) realistic noise, fixed sample, (b) realistic noise, random sample, (c) white noise, fixed sample, (d) white noise, random sample.

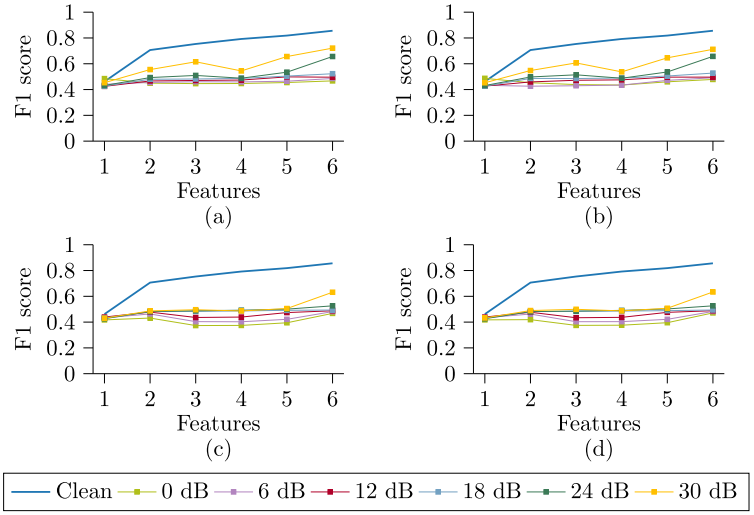


Fig. 6 – Mean F1 scores using cumulative features for Reinke's edema, UEX-Voice database. SNRs ranging from 0 dB to 30 dB in steps of 6 dB. The features are: 1-MFCC7, 2-CPP, 3-MFCC2, 4-SHANNON, 5-MFCC10, 6-MFCC4. Noise characteristics: (a) realistic noise, fixed sample, (b) realistic noise, random sample, (c) white noise, fixed sample, (d) white noise, random sample.

of synthetic white Gaussian noise than in the case of realistic noise. Also, as it happens in the experiment about nodules, the effect of noise addition is more pronounced on UEX-Voice database than in the case of MEEI.

F1 scores for MEEI database along with accuracy show that the classifier is reliable for SNRs as low as 18 dB, where both values reach over 0.9 in the case of realistic noise. For UEX-Voice database, the minimum SNR to show acceptable perfor-

Table 3 – Accuracy, precision, recall and, F1 score values computed for MEEI database, nodules disease, SNRs ranging from 0 dB to 30 dB in 6 dB steps, using 5 features: MFCC1, CPP, HURST, AGE, and MFSW.

		Real fixed	Real random	White fixed	White random
Clean	Accuracy	0.95	0.95	0.95	0.95
	Precision	0.88	0.88	0.88	0.88
	Recall	0.95	0.95	0.95	0.95
	F1 score	0.91	0.91	0.91	0.91
30 dB	Accuracy	0.95	0.95	0.95	0.95
	Precision	0.87	0.88	0.86	0.87
	Recall	0.95	0.95	0.94	0.95
	F1 score	0.91	0.91	0.90	0.90
24 dB	Accuracy	0.95	0.94	0.94	0.94
	Precision	0.86	0.85	0.85	0.86
	Recall	0.94	0.92	0.91	0.91
	F1 score	0.90	0.88	0.88	0.89
18 dB	Accuracy	0.93	0.92	0.92	0.93
	Precision	0.85	0.81	0.87	0.88
	Recall	0.89	0.88	0.83	0.83
	F1 score	0.87	0.85	0.85	0.85
12 dB	Accuracy	0.88	0.87	0.88	0.88
	Precision	0.84	0.83	0.91	0.90
	Recall	0.68	0.64	0.59	0.58
	F1 score	0.75	0.72	0.71	0.70
6 dB	Accuracy	0.75	0.76	0.75	0.75
	Precision	0.85	0.94	0.98	0.98
	Recall	0.04	0.05	0.01	0.01
	F1 score	0.07	0.10	0.01	0.02
0 dB	Accuracy	0.74	0.74	0.74	0.74
	Precision	–	–	–	–
	Recall	0.00	0.00	0.00	0.00
	F1 score	–	–	–	–

mance is 24 dB, where F1 score drops from 0.83 to 0.71 and accuracy, precision and recall score show a similar degradation.

4. Discussion

The results involve two diseases, two different databases and four kinds of noise. This allows to perform a comparative analysis from different perspectives. In spite of the fact that we have studied the effects of noise addition in the performance of a classifier using F1 score as the reference metric, most studies use accuracy as the main performance indicator [9]. However, since we have also computed accuracy, and the best results are obtained using all the features selected in each case, we can compare our system performance with prior research in the field.

Comparing clean case performance allows us to analyze the differences from a database point of view, with MEEI database the detection accuracies reach 0.95, while the systems reach 0.71 for nodules and 0.84 for Reinke's edema with UEX-Voice database. This difference in performance between MEEI and a database obtained in more realistic conditions is in line with the scientific literature. Whereas most reported detection accuracies for MEEI data are in excess of 0.9, in [8] best accuracies of 0.784 and 0.762 were achieved after carrying out voice pathology detection experiments using the Hospital Universitario Príncipe de Asturias database (HUPA) and the Saarbrücken Voice Disorder database (SVD), respectively. [43] computed accuracy, recall, and other metrics when classifying recordings from MEEI (0.91–0.97 accuracy, 0.93–0.98 recall) and HUPA databases (0.68–0.82 accuracy, 0.77–0.85 recall). Moreover, [10] achieves 0.95/0.97 accuracy/recall for MEEI database using spectral-cepstral features, while the results with HUPA database using the same features only reach 0.78/0.74.

Some studies have taken into consideration noise corruption. For example, [24] studies environmental noise and white Gaussian noise effects on Parkinson's disease detection using a variety of vocal tasks including a phonation model based on sustained vowels. Both disease and noise are not directly comparable since Parkinson's disease is a neurological disease and different diseases require different analysis techniques which depend on the specific effects on voice [10]. Moreover, noise was recorded in 8 different scenarios. However, it shows that with clean training the accuracy for the phonation model drops from about 0.7 to 0.5 when SNR is equal to 0 dB, much like the results obtained here. Furthermore, some research has been made in order to alleviate the effects of different recording conditions on disease detection performance [25].

Table 4 – Accuracy, precision, recall and, F1 score values computed for UEX-Voice database, nodules disease, SNRs ranging from 0 dB to 30 dB in 6 dB steps, using 4 features: CPP, MFCC7, MFCC3, and MFCC2.

		Real fixed	Real random	White fixed	White random
Clean	Accuracy	0.71	0.71	0.71	0.71
	Precision	0.74	0.74	0.74	0.74
	Recall	0.52	0.52	0.52	0.52
	F1 score	0.61	0.61	0.61	0.61
30 dB	Accuracy	0.67	0.67	0.59	0.59
	Precision	0.74	0.73	0.59	0.59
	Recall	0.41	0.40	0.24	0.25
	F1 score	0.53	0.52	0.34	0.35
24 dB	Accuracy	0.64	0.63	0.53	0.53
	Precision	0.70	0.68	0.43	0.43
	Recall	0.33	0.32	0.14	0.14
	F1 score	0.45	0.44	0.21	0.21
18 dB	Accuracy	0.60	0.59	0.51	0.51
	Precision	0.63	0.59	0.35	0.35
	Recall	0.26	0.25	0.11	0.11
	F1 score	0.37	0.36	0.16	0.16
12 dB	Accuracy	0.55	0.54	0.51	0.50
	Precision	0.48	0.45	0.32	0.32
	Recall	0.16	0.17	0.10	0.10
	F1 score	0.24	0.24	0.15	0.15
6 dB	Accuracy	0.54	0.53	0.50	0.50
	Precision	0.46	0.40	0.30	0.30
	Recall	0.16	0.12	0.09	0.09
	F1 score	0.24	0.19	0.14	0.14
0 dB	Accuracy	0.54	0.53	0.49	0.49
	Precision	0.43	0.41	0.28	0.27
	Recall	0.11	0.12	0.08	0.08
	F1 score	0.17	0.18	0.13	0.13

On the other side, when noise is added with a low SNR, MEEI database gets much higher results for all metrics than UEX-Voice. Apart from the fact that MEEI database was collected in a more controlled acoustic environment, some authors have pointed out that this database contains no lightly pathological speakers [29], and that the normal and dysphonic voices present in the database are easily separable [44], which makes the classification task easy.

Given the proportions of healthy and pathological samples present in the databases, shown in Table 1, MEEI test set contains roughly 70% of healthy patients whereas in UEX-Voice 50% of test samples are healthy. Those ratios match the accuracies obtained for the worst SNR ratios for all the classifiers for both databases. Precision and recall values support the fact that the classifier is unable to distinguish pathological subjects and marks most of them as healthy. That explains the differences in the lower accuracy levels shown between Tables 3, 5 and 4, 6.

Considering the different kinds of noise it seems that realistic noise is less intrusive than white synthetic noise. This trend is specially pronounced for UEX-Voice database. A possible explanation for this is that the spectral compositions of both types of noise are different. White noise (Fig. 7a) is characterized by an even spectral power density, thus all the frequencies in the full bandwidth are interfered in the same way. However, realistic noise coming from several sources in the hospital environment concentrates most energy in a lower part of the spectrum. A spectrogram of an example of realistic noise segment is shown in Fig. 7b, where it is easy to see the spectral contributions of the noise sources taken in consideration, and how realistic noise most prominent frequencies lie in the lower half of spectrum, and even considering that bandwidth, frequencies below 4 kHz stand out.

Regarding noise randomness, the variability introduced by random noise sampling in all cases has little impact in the overall capacity of the resulting classifiers. Although some differences exist in the results, there is no consistency in any advantage of fixed over random sampling or vice versa, as we can see, for example, in Fig. 5, subplots (a) versus (b) or Fig. 5, subplots (c) versus (d).

The comparative analysis of the results from a disease perspective is more challenging. Vocal fold nodules are smooth, benign masses involving anterior or middle vocal folds and located superficially to the free edge of the fold. Reinke's edema (also known as polypoid degeneration) is characterized by an accumulation of fluid, usually in both vocal folds [45]. Since both pathologies share some histological characteristics, [2] even proposed to use the term "exudative lesions on the Reinke's space" to refer to nodules, polyps, and Reinke's edema. These histological characteristics affect the vibra-

Table 5 – Accuracy, precision, recall and, F1 score values computed for MEEI database, Reinke's edema disease, SNRs ranging from 0 dB to 30 dB in 6 dB steps, using 5 features: MFCC1, PERMUTATION, Shimmer, MFCC3, and CPP.

		Real fixed	Real random	White fixed	White random
Clean	Accuracy	0.99	0.99	0.99	0.99
	Precision	1.00	1.00	1.00	1.00
	Recall	0.96	0.96	0.96	0.96
	F1 score	0.98	0.98	0.98	0.98
30 dB	Accuracy	0.98	0.98	0.98	0.97
	Precision	1.00	1.00	1.00	1.00
	Recall	0.94	0.93	0.92	0.91
	F1 score	0.97	0.96	0.96	0.96
24 dB	Accuracy	0.98	0.97	0.93	0.94
	Precision	1.00	1.00	1.00	1.00
	Recall	0.93	0.91	0.80	0.81
	F1 score	0.96	0.96	0.89	0.90
18 dB	Accuracy	0.94	0.95	0.84	0.84
	Precision	1.00	1.00	0.99	0.97
	Recall	0.83	0.84	0.50	0.52
	F1 score	0.91	0.91	0.66	0.67
12 dB	Accuracy	0.84	0.81	0.81	0.82
	Precision	1.00	1.00	1.00	1.00
	Recall	0.50	0.41	0.39	0.43
	F1 score	0.67	0.58	0.56	0.61
6 dB	Accuracy	0.78	0.76	0.77	0.81
	Precision	0.94	0.93	0.82	0.88
	Recall	0.32	0.26	0.37	0.46
	F1 score	0.47	0.41	0.51	0.60
0 dB	Accuracy	0.69	0.69	0.71	0.69
	Precision	0.87	0.72	0.88	0.77
	Recall	0.03	0.03	0.12	0.06
	F1 score	0.06	0.06	0.20	0.11

tory patterns of the vocal folds (increase in mass of the folds, reduction in the pliability of the overlying cover...), and may produce some common perceptual consequences, such as hoareseness and breathiness. Nevertheless, it can be observed that Reinke's edema is detected with higher accuracy and F1 score (0.99 and 0.98 respectively in the case of MEEI; 0.84 and 0.83 respectively in the case of UEX-Voice) than nodules (0.95 and 0.91 respectively in the case of MEEI; 0.71 and 0.61 respectively in the case of UEX-Voice), which may be the consequence of its inflammatory character producing a more severe impact on voice quality in comparison to a simple mass lesion.

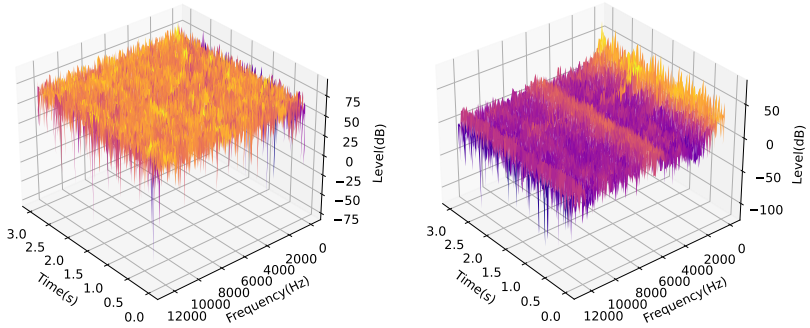
The overall structure of the system is in line with most of previous work, with the common steps of preprocessing, feature extraction, dimensionality reduction, machine learning training, and system evaluation [5]. Regarding dimensionality reduction, prior work in the field include techniques such principal component analysis, linear discriminant analysis, or minimum redundancy maximum relevance among others. The four experimental settings have led to four different feature subsets. The composition of these feature subsets is important as they may provide some clues not only on which features are more important when building a new detection system, but also which ones show a certain noise robustness.

Although the feature selection process is applied on the original databases, without noise addition, UEX-Voice is recorded in a more realistic acoustic environment, so it is possible to conclude that, if there are features that are selected using both databases, they may have a reliable discrimination potential across different databases under moderately controlled acoustic conditions. This is the case of CPP and MFCCs. They play a very important role, as CPP and at least one MFCC have been selected within the most useful features in the four cases, to the point that for nodules disease in UEX-Voice database all the selected features are MFCCs and CPP. Both share the advantage that, unlike traditional acoustic measures such as jitter or shimmer, they do not require a pitch estimation which may be difficult due to the absence of periodicity in severely dysphonic voices. This is in line with results obtained by [28], where it is shown that advanced multiband cepstral analysis might be useful in disease detection and even in disease discrimination, and [10] which shows the ability of spectral-cepstral features to classify disphonic voices based on a sustained vowel analysis.

The rest of selected features is heterogeneous among the four studied cases: Non-linear analysis features are found in Fig. 3 with HURST and MFSW, but no other case shows nonlinear features. Entropies make their appearance in both Reinke's edema cases, Figs. 5 and 6, with permutation and Shannon entropies, but not in nodules cases. From the classical perturbation measurements only shimmer is selected for MEEI Reinke's edema case, but not for UEX-Voice. The reason

Table 6 – Accuracy, precision, recall and, F1 score values computed for UEX-Voice database, Reinke's edema disease, SNRs ranging from 0 dB to 30 dB in 6 dB steps, using 6 features: MFCC7, CPP, MFCC2, SHANNON, MFCC10, and MFCC4.

		Real fixed	Real random	White fixed	White random
Clean	Accuracy	0.84	0.84	0.84	0.84
	Precision	0.84	0.84	0.84	0.84
	Recall	0.83	0.83	0.83	0.83
	F1 score	0.83	0.83	0.83	0.83
30 dB	Accuracy	0.76	0.76	0.68	0.69
	Precision	0.79	0.79	0.68	0.69
	Recall	0.71	0.70	0.68	0.68
	F1 score	0.75	0.74	0.68	0.69
24 dB	Accuracy	0.71	0.71	0.55	0.55
	Precision	0.72	0.71	0.55	0.55
	Recall	0.70	0.70	0.57	0.57
	F1 score	0.71	0.71	0.56	0.56
18 dB	Accuracy	0.56	0.56	0.47	0.47
	Precision	0.56	0.56	0.48	0.48
	Recall	0.58	0.57	0.51	0.51
	F1 score	0.57	0.57	0.49	0.49
12 dB	Accuracy	0.48	0.48	0.47	0.47
	Precision	0.48	0.48	0.47	0.47
	Recall	0.51	0.51	0.52	0.52
	F1 score	0.50	0.50	0.49	0.49
6 dB	Accuracy	0.48	0.48	0.48	0.48
	Precision	0.48	0.48	0.48	0.48
	Recall	0.51	0.51	0.52	0.52
	F1 score	0.50	0.50	0.50	0.50
0 dB	Accuracy	0.47	0.47	0.48	0.48
	Precision	0.47	0.47	0.48	0.48
	Recall	0.47	0.49	0.49	0.49
	F1 score	0.47	0.48	0.48	0.49



(a) White noise. (b) Realistic noise.

Fig. 7 – Spectrograms for white noise and an example of realistic noise segment from the hospital environment. Both noise recordings were used on the same voice recording.

might be that, as an amplitude perturbation measure, shimmer is very sensitive to noise, performing better in a more controlled acoustic environment.

On the classifier side, we chose SVM for its simplicity and execution speed since WOA feature selection is computationally expensive. Many alternatives have been used, line Hidden Markov Models, Gaussian Mixture Models, K-nearest neighbors or decision trees to name a few of them [9]. Most of the alternatives found in previous work uses that kind of algorithms, although in recent years artificial neural networks have gained popularity and we start to see studies using such techniques.

Deep learning methods have seldom been used in this specific application until recent times. [9] mentions artificial neural networks but only shows multilayer perceptron, which barely can be classified as a deep learning method. [46] presents 2 out of 45 studies using deep learning techniques, which date from 2019. The most plausible reason is database size. Looking at the numbers shown in Table 1, the number of samples is very low, and a small multilayer neural network comprises thousands of coefficients. DenseNet has been used on cepstrum features [47] with good results although it cites the low number of pathological samples as a limitation. Other classical deep learning approaches like VGG16 and CaffeNet have been used [48], with the particularity that those algorithms are used for image recognition and classification. Consequently, raw waveforms are feed into the network (in the form of spectrograms) since it will infer features, and transfer learning techniques (neural network partially trained with examples from other fields) are used to overcome the long training times and small dataset size limitations.

Research on robust pathology detectors has not been addressed until recent times. [49] performs experiments using four different databases, aiming at robustness against different recording conditions, but does not focus on specific differences between them. Little work has been done around noise robustness in voice quality assessment, so thorough comparisons can not be made, although this research is necessary. For example, [20,47] point towards differences in recording environment (e.g. background noise) as a limitation for different studies results comparison. [50] points out the difficulties to extrapolate the results obtained with different databases due their recording differences. However, [21] proposes a SNR level of 42 dB for perturbation measurements (jitter and shimmer) to be reliable, and estimates 30 dB as the lowest limit of SNR level for reliable usage of classifiers. This seems to match the results for MEEI database in Figs. 3 and 5, where the F1 score is almost identical for the clean and the 30 dB SNR cases, for all the numbers of features considered, specially when realistic noise is added.

Considering the impact of noise can benefit other research work focused on mobile health tools to detect vocal fold disorders. There is currently a high interest in the development of mobile-aided systems to manage a wide variety of diseases and, in particular, disorders affecting voice [17–19]. A critical aspect is to check if the approaches proposed for controlled conditions are robust or have to be modified when used in increasingly realistic environments.

5. Conclusion

The results of this paper highlight the importance of performing experiments on more realistic voice pathology databases, alternative to MEEI, since the achievable prediction accuracies are not expected to be comparable. The feature subsets obtained by feature selection with MEEI and with a more realistic database collected in the scope of this work emphasize the role of CPP and MFCCs as useful and robust features to discriminate pathological from healthy voices.

Also, the degrading impact of additive noise on AVCA systems based on acoustic features for detection of nodules and Reinke's edema has been demonstrated and quantified. Although the effect of real-world noise recorded in a clinical environment has been shown to be lower than that of white noise, the effect is sufficiently detrimental to motivate further research into noise-robust prediction systems.

In the future, it will be interesting to increase UEX-Voice database by including new organic pathologies. Also, exploring new techniques in the field like deep learning and looking for solutions to overcome the voice databases limitations are of research interest.

CRediT authorship contribution statement

Mario Madruga: Conceptualization, Software, Validation, Data curation, Writing - original draft, Visualization. **Yolanda Campos-Roca:** Conceptualization, Methodology, Resources, Writing - review & editing. **Carlos J. Pérez:** Conceptualization, Methodology, Validation, Formal analysis, Resources, Writing - review & editing, Supervision.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

The authors would like to thank Dr. Moreno for his medical advising, Sandra Paniagua and Esther de la O. for their work recording the UEX-Voice database in HSPdA and all the voluntary individuals, patients and healthy. They also would like to thank Prof. Gómez (from the Acoustics Laboratory) for making the sound level meter available and the Advanced Scientific Computation at the University of Extremadura to provide access to computation facilities.

This research has been supported by project MTM2017-86875-C3-2-R (Ministerio de Ciencia, Innovación y Universidades), projects IB16054, GR18108 and GR18055 (Junta de Extremadura/European Regional Development Funds, EU), and FPU18/03274 grant (Ministerio de Ciencia, Innovación y Universidades).

REFERENCES

[1] Rufo M, Martín J, Pérez C, Paniagua S. A Bayesian decision analysis approach to assess voice disorder risks by using acoustic features. Biometr J 2019;61(3):503–13.
[2] Hantzakos A, Remacle M, Dikkers FG, Degols JC, Delos M, Friedrich G, Giovanni A, Rasmussen N. Exudative lesions of Reinke's space: a terminology proposal. Eur Arch Otorhinolaryngol 2009;266(6):869.
[3] Echternach M, Döllinger M, Köberlein M, Kuranova L, Gellrich D, Kainz M. Vocal fold oscillation pattern changes related to loudness in patients with vocal fold mass lesions. J Otolaryngol Head Neck Surg 2020;49(1):1–9.

[4] Sataloff RT. Clinical assessment of voice. Plural publishing; 2017..

[5] Gómez-García J, Moro-Velázquez L, Godino-Llorente J. On the design of automatic voice condition analysis systems. Part I: Review of concepts and an insight to the state of the art. Biomed Sig Process Control 2019;51:181–99.

[6] Kowalska-Taczanowska R, Friedman A, Koziorowski D. Parkinson's disease or atypical parkinsonism? The importance of acoustic voice analysis in differential diagnosis of speech disorders. Brain Behav 2020;10(8) e01700.

[7] Paniagua M, Pérez C, Calle-Alonso F, Salazar C. An acoustic-signal-based preventive program for university lecturers' vocal health. J Voice 2018;34(1):88–99.

[8] Kadiri SR, Alku P. Analysis and detection of pathological voice using glottal source features. IEEE J Select Top Sig Process 2019;14(2):367–79.

[9] Hegde S, Shetty S, Rai S, Dodderi T. A survey on machine learning approaches for automatic detection of voice disorders. J Voice 2019;33(6):947–58.

[10] Orozco-Arroyave JR, Belalcazar-Bolanos EA, Arias-Londoño JD, Vargas-Bonilla JF, Skodda S, Rusz J, Daqrouq K, Hönig F, Nöth E. Characterization methods for the detection of multiple voice disorders: neurological, functional, and laryngeal diseases. IEEE J Biomed Health Inf 2015;19 (6):1820–8.

[11] J. Tang, S. Alelyani, and H. Liu, Feature selection for classification: A review, Data classification: Algorithms and Applications, 2014:37–64..

[12] Mirjalili S, Lewis A. The whale optimization algorithm. Adv Eng Softw 2016;95:51–67.

[13] Canayaz M, Demir M. Feature selection with the whale optimization algorithm and artificial neural network. In: 2017 International Artificial Intelligence and Data Processing Symposium (IDAP).

[14] Mafarja M, Mirjalili S. Whale optimization approaches for wrapper feature selection. Appl Softw Comput 2018;62:441–53.

[15] Massachusetts Eye and Ear Infirmary, Voice disorders database, Version 1.03 (cd-rom), Lincoln Park, NJ: Kay Elemetrics Corporation; 1994..

[16] Travieso CM, Alonso JB, Orozco-Arroyave JR, Solé-Casals J, Gallego-Jutglà E. Automatic detection of laryngeal pathologies in running speech based on the HMM transformation of the nonlinear dynamics. Int Conf Nonlinear Speech Process 2013.

[17] Arias-Vergara T, Vásquez-Correa J, Orozco-Arroyave J, Nöth E. Speaker models for monitoring Parkinson's disease progression considering different communication channels and acoustic conditions. Speech Commun 2018;101:11–25.

[18] Tsanas A, Little M, Ramig L. Remote assessment of parkinson's disease symptom severity using the simulated cellular mobile telephone Network. IEEE Access. 2021..

[19] Cesari U, De Pietro G, Marciano E, Niri C, Sannino G, Verde L. Voice disorder detection via an m-Health system: Design and results of a clinical study to evaluate Vox4Health. BioMed Res Int 2018;2018.

[20] Saggio G, Costantini G. Worldwide healthy adult voice baseline parameters: a comprehensive review. J Voice 2020.

[21] Deliyski DD, Shaw HS, Evans MK. Adverse effects of environmental noise on acoustic voice quality measurements. J Voice 2005;19(1):15–28.

[22] van der Woerd B, Wu M, Parsa V, Doyle P, Fung K. Evaluation of Acoustic Analyses of Voice in Nonoptimized Conditions. J Speech Language Hearing Res 2020;63(12):3991–9.

[23] Madruga M, Campos-Roca Y, Pérez CJ. Robustness assessment of automatic Reinke's edema diagnosis systems. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[24] Vásquez-Correa JC, Serra J, Orozco-Arroyave JR, Vargas-Bonilla JF, Nöth E. Effect of acoustic conditions on algorithms to detect Parkinson's disease from speech. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2017. .

[25] Madruga M, Campos-Roca Y, Pérez C. Multicondition training for noise-robust detection of benign vocal fold lesions from recorded speech. IEEE Access. 2020..

[26] Fraile R, Godino-Llorente JI. Cepstral peak prominence: A comprehensive analysis. Biomed Signal Process Control 2014;14:42–54.

[27] Tsanas A. Acoustic analysis toolkit for biomedical speech signal processing: concepts and algorithms. Models Anal Vocal Emiss Biomed Appl 2013;2:37–40.

[28] Alves M, Silva G, Bispo B, Dajer M, Rodrigues P. Voice disorders detection through multiband cepstral features of sustained vowel. J Voice 2021.

[29] Henríquez P, Alonso JB, Ferrer MA, Travieso CM, Godino-Llorente JI, Díaz-de María F. Characterization of healthy and pathological voice through measures based on nonlinear dynamics. IEEE Trans Audio Speech Language Process 2009;17 (6):1186–1195..

[30] Ihlen EAF. Introduction to multifractal detrended fluctuation analysis in matlab. Front Physiol 2012;3:141.

[31] Islam R, Tarique M, Abdel-Raheem E. A survey on signal processing based pathological voice detection techniques. IEEE Access 2020;8:66749–76.

[32] Riedl M, Müller A, Wessel N. Practical considerations of permutation entropy. Eur Phys J Spec Top 2013;222(2):249–62.

[33] Orozco JR, Vargas JF, Alonso JB, Ferrer MA, Travieso CM, Henríquez P. Voice pathology detection in continuous speech using nonlinear dynamics. In: 2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA).

[34] Behrman A. Speech and voice science. Plural publishing; 2017..

[35] Hixon TJ, Weismer G, Hoit JD. Preclinical speech science: Anatomy, physiology, acoustics, and perception. Plural Publishing; 2018..

[36] Van Houtte E, Van Lierde K, D'haeseleer E, Claeys S. The prevalence of laryngeal pathology in a treatment-seeking population with dysphonia. Laryngoscope 2010;120(2):306–312..

[37] Brückl M, Ghio A, Alain, Viallet F. Measurement of tremor in the voices of speakers with Parkinson's disease. Proc Comput Sci 2018;128:47–54.

[38] Illner V, Sovka P, Rusz J. Validation of freely-available pitch detection algorithms across various noise levels in assessing speech captured by smartphone in Parkinson's disease. Biomed Signal Process Control 2020;58 101831.

[39] Tsanas A. Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning, Ph.D. dissertation, Oxford University, UK, 2012..

[40] Tsanas A, Gómez-Vilda P. Novel robust decision support tool assisting early diagnosis of pathological voices using acoustic analysis of sustained vowels, in Multidisciplinary Conference of Users of Voice, Speech and Singing (JVHC 13); 2013. .

[41] Despotovic V, Skovranek T, Schommer C. Speech based estimation of Parkinson's disease using Gaussian processes and automatic relevance determination. Neurocomputing 2020;401:173–81.

[42] Luan F, Cai Z, Wu S, Liu S, He Y. Optimizing the low-carbon flexible job shop scheduling problem with discrete whale optimization algorithm. Mathematics 2019;7(8):688.

[43] Arias-Londoño J, Godino-Llorente J, Sáenz-Lechón N, Osma-Ruiz V, Víctor, Castellanos-Domínguez G. An improved method for voice pathology detection by means of a HMM-based feature space transformation. Pattern Recognit 2010;43 (9):3100–12.
[44] Daoudi K, Bertrac B. On classification between normal and pathological voices using the MEEI-KayPentax database: Issues and consequences. In: Fifteenth Annual Conference of the International Speech Communication Association.
[45] Sataloff RT, Chowdhury F, Portnoy JE, Hawkshaw MJ, Joglekar S. Surgical techniques in otolaryngology-head & Neck Surgery: Laryngeal Surgery. JP Medical Ltd 2013.
[46] Syed S, Rashid M, Hussain S. Meta-analysis of voice disorders databases and applied machine learning techniques. Math Biosci Eng: MBE 2020;17(6):7958–79.
[47] Fang S, Tsao Y, Hsiao M, Chen J, Lai Y, Lin F, Wang C. Detection of pathological voice using cepstrum vectors: A deep learning approach. J Voice 2019;33(5):634–41.
[48] Alhussein M, Muhammad G. Voice pathology detection using deep learning on mobile healthcare framework. IEEE Access 2018;6:41034–41.
[49] Harar P, Galaz Z, Alonso-Hernandez Jesus B, Mekyska J, Burget R, Smekal Z. Towards robust voice pathology detection. Neural Comput Appl 2018:1–11.
[50] Karan B, Sahu S, Mahto K. Parkinson disease prediction using intrinsic mode function based features from speech signal. Biocybern Biomed Eng 2020;40(1):249–64.

Chapter 3

A mobile-assisted voice condition analysis system for Parkinson's disease: assessment of usability conditions

Title:

A mobile-assisted voice condition analysis system for Parkinson's disease: assessment of usability conditions

Authors and affiliation:

Javier Carrón[a], Yolanda Campos-Roca[b], Mario Madruga[a], Carlos J. Pérez[a]

[a]Universidad de Extremadura, Departamento de Matemáticas, Spain
[b]Universidad de Extremadura, Departamento de Tecnología de los Computadores y las Comunicaciones, Spain

Journal:

BioMedical Engineering OnLine

DOI:

10.1186/s12938-021-00951-y

Abstract:

Background and objective

Automatic voice condition analysis systems to detect Parkinson's disease (PD) are generally based on speech data recorded under acoustically controlled conditions and professional supervision. The performance of these approaches in a free-living scenario is unknown. The aim of this research is to investigate the impact of uncontrolled conditions (realistic acoustic environment and lack of supervision) on the performance of automatic PD detection systems based on speech.

Methods

A mobile-assisted voice condition analysis system is proposed to aid in the detection of PD using speech. The system is based on a server–client architecture. In the server, feature extraction and machine learning algorithms are designed and implemented to discriminate subjects with PD from healthy ones. The Android app allows patients to submit phonations and physicians to check the complete record of every patient. Six different machine learning classifiers are applied to compare their performance on two different speech databases. One of them is an in-house database (UEX database), collected under professional supervision by using the same Android-based smartphone in the same room, whereas the other one is an age, sex and health-status balanced subset of mPower study for PD, which provides real-world data. By applying identical methodology, single-database experiments have been performed on each database, and also cross-database tests. Cross-validation has been applied to assess generalization performance and hypothesis tests have been used to report statistically significant differences.

Results

In the single-database experiments, a best accuracy rate of 0.92 (AUC = 0.98) has been obtained on UEX database, while a considerably lower best accuracy rate of 0.71 (AUC = 0.76) has been achieved using the mPower-based database. The cross-database tests provided very degraded accuracy metrics.

Conclusion

The results clearly show the potential of the proposed system as an aid for general practitioners to conduct triage or an additional tool for neurologists to perform diagnosis. However, due to the performance degradation observed using data from mPower study, semi-controlled conditions are encouraged, i.e., voices recorded at home by the patients themselves following a strict recording protocol and control of the information about patients by the medical doctor at charge.

Keywords: Acoustic features, Machine learning, mPower database, Parkinson's disease, Speech processing, Voice condition analysis system.

Carrón *et al. BioMedical Engineering OnLine* (2021) 20:114
https://doi.org/10.1186/s12938-021-00951-y

BioMedical Engineering OnLine

RESEARCH Open Access

A mobile-assisted voice condition analysis system for Parkinson's disease: assessment of usability conditions

Javier Carrón[1], Yolanda Campos-Roca[2], Mario Madruga[1] and Carlos J. Pérez[1*]

*Correspondence: carper@unex.es
[1] Departamento de Matemáticas, Universidad de Extremadura, Cáceres, Spain
Full list of author information is available at the end of the article

Abstract

Background and objective: Automatic voice condition analysis systems to detect Parkinson's disease (PD) are generally based on speech data recorded under acoustically controlled conditions and professional supervision. The performance of these approaches in a free-living scenario is unknown. The aim of this research is to investigate the impact of uncontrolled conditions (realistic acoustic environment and lack of supervision) on the performance of automatic PD detection systems based on speech.

Methods: A mobile-assisted voice condition analysis system is proposed to aid in the detection of PD using speech. The system is based on a server–client architecture. In the server, feature extraction and machine learning algorithms are designed and implemented to discriminate subjects with PD from healthy ones. The Android app allows patients to submit phonations and physicians to check the complete record of every patient. Six different machine learning classifiers are applied to compare their performance on two different speech databases. One of them is an in-house database (UEX database), collected under professional supervision by using the same Android-based smartphone in the same room, whereas the other one is an age, sex and health-status balanced subset of mPower study for PD, which provides real-world data. By applying identical methodology, single-database experiments have been performed on each database, and also cross-database tests. Cross-validation has been applied to assess generalization performance and hypothesis tests have been used to report statistically significant differences.

Results: In the single-database experiments, a best accuracy rate of 0.92 (AUC = 0.98) has been obtained on UEX database, while a considerably lower best accuracy rate of 0.71 (AUC = 0.76) has been achieved using the mPower-based database. The cross-database tests provided very degraded accuracy metrics.

Conclusion: The results clearly show the potential of the proposed system as an aid for general practitioners to conduct triage or an additional tool for neurologists to perform diagnosis. However, due to the performance degradation observed using data from mPower study, semi-controlled conditions are encouraged, i.e., voices recorded at home by the patients themselves following a strict recording protocol and control of the information about patients by the medical doctor at charge.

Keywords: Acoustic features, Machine learning, mPower database, Parkinson's disease, Speech processing, Voice condition analysis system

Introduction

Parkinson's disease (PD) is an up-to-now incurable neurodegenerative disorder that mainly, but not exclusively, affects the motor system. It is the most relevant neurodegenerative disorder after Alzheimer's disease, but with a faster growth. The Global Burden of Disease study projects to reach 13 million people affected by PD in 2040 [10].

PD is typically diagnosed by a neurologist when certain motor symptoms become clinically evident, in particular when bradykinesia occurs along with rigidity or tremor. Early diagnosis is key to improve quality of life of people suffering from PD. However, in the European survey presented by Bloem and Stocchi [6], diagnosis time after the first symptoms' onset was above 2 years in 11.8% of the patients. Misdiagnoses are also common and can be as high as 25% when the practitioners have limited clinical experience in PD [26]. The situation is critical in developing countries, where many patients remain undiagnosed [11]. Therefore, new tools seem necessary to obtain an early diagnosis.

Subjects with PD suffer from speech impairment [8]. This leads to consider automatic analysis of voice recordings as a potential tool to aid diagnosis. Different vocal tasks, focused on phonation, articulation, prosody, and cognitive–linguistic aspects have been used for the detection of PD through voice. The most used vocal task is the sustained phonation of the /a/ vowel due to its simplicity and ubiquity in different languages [30, 46]. Previous works have used a wide variety of acoustic features extracted from this type of speech recordings. For example, perturbation measures (such as Jitter or Shimmer [50]), noise measures (for instance, the harmonic-to-noise ratio (HNR) [20]), spectral and cepstral features [37], and several features based on nonlinear analysis [50], among others.

Also, diadochokinesis test recordings studying articulatory tasks [28, 41], prosodic features extracted from reading texts and spontaneous speech [19, 53], and even combinations of different vocal tasks [43] have been proposed. An equally wide range of proposals can be found regarding machine learning techniques. Commonly used classifiers that have been used for this application are: Random Forest, Neural Networks or Support Vector Machines, among others [18, 29, 38].

Those studies were carried out using speech recordings obtained using high-grade equipment like professional microphones and sound cards. Several feature datasets that have been extracted from recordings obtained with this type of equipment are publicly available [24, 30, 44]. Some authors have performed cross-database tests, which involve different microphones, environment, and even languages [35, 54], although always under controlled conditions. In this article, the term "controlled conditions" refers to the fact that there is professional supervision of the recordings and a certain control on the acoustic environment so that at least the noise level is low.

Systems built on recordings based on professional equipment are limited in the range of potential applications. Due to the ever-increasing penetration of smartphones, using these mobile devices would allow for extending the application of automatic PD detection through voice on a larger scale. The use of these devices to record phonations and build databases is an interesting strategy introduced in some recent studies. Almeida

et al. [1] proposed a comparison of two different datasets of sustained vowel phonations. These datasets have been obtained through simultaneous recordings by using a professional microphone and a smartphone. Afterwards, a common methodology, consisting of preprocessing, feature extraction, and classification, was applied to both datasets comparing the results obtained in each case. In a similar way, Rusz et al. [42] simultaneously recorded different vocal tasks with a professional head-mounted condenser microphone and a smartphone, comparing the results. The outcomes point in the direction that detection of speech abnormalities due to PD via a smartphone is possible.

As the use of mobile phones increases the scope of this research line, specialized app development is a natural step. Some reviews have been published on the existing and potentially useful apps for PD patients available in the leading app stores [23, 39]. However, they concluded that, despite the clear potential of this type of technology, further efforts and more improvements are needed for it to be effectively used in a real clinical scenario. In line with this demand, a smartphone app frontend in conjunction with a computing server backend has been designed and implemented as a necessary step to build a mobile-assisted voice condition analysis system. The app allows patients to provide data and physicians to check the complete record of every patient. The system is completed with a machine learning approach to perform PD detection on the server side. This approach is built on top of a feature extraction process that includes some of the most relevant algorithms for PD detection, a recursive feature elimination selection process, and a classifier. To provide robust results cross-validations have been considered. Besides, approaches with six different classifiers have been implemented for comparison purposes. The system also allows its use with future implementations to aid also disease monitoring.

A critical aspect is to check the results obtained in increasingly realistic environments. The works previously mentioned were issued in a controlled environment and under supervision. More concretely, in Rusz et al. [42] the speech recordings were performed in a quiet room with an environmental noise level lower than 50 dB, and with a specialist who guided the participants through the recording protocol. In the case of Almeida et al. [1], the recordings were taken in a sound-proof booth. However, there are also recent studies that use public repositories where participants send their voice recordings and complementary information (age, health status, sex, etc.) without any professional supervision. This is the case of mPower PD database [7]. Some previous contributions using this database show the results of applying different feature extraction and machine learning techniques to perform PD detection based on uncontrolled conditions, that is, unknown acoustic environment and without a professional control to make sure that the recordings strictly follow the protocol [48, 49, 55, 56]. These studies do not ensure age and sex balances in the mPower-based datasets they use. Age and sex balances are necessary to avoid potential biases in the results. Also, to the authors' best knowledge, cross-database studies that use data obtained in a realistic environment have not been presented. Research that considered smartphone recordings has focused on datasets collected either in controlled or uncontrolled conditions. However, both types of scenarios have not been jointly considered under the same methodology.

The research hypothesis is that the accuracy obtained by a mobile-assisted PD detection system based on voice tested on a controlled scenario (in terms of acoustic

environment and professional supervision) is degraded when the scenario is uncontrolled. The aim of this research work is to analyze the impact of uncontrolled acoustic environment and lack of professional supervision during the recordings avoiding the influence of the feature extraction and machine learning algorithms. This requires the application of exactly the same methodology on controlled and uncontrolled databases and the realization of cross-database experiments, in which the training is performed with one database and the test with the other one.

One of the databases is an in-house one (UEX database), collected with professional supervision in a controlled environment. It has been obtained from an experiment specifically conducted to help in the detection of PD. The second one is a subset of the public mPower database, collected in a realistic environment without professional control. This subset has been chosen to ensure age and sex balance as well as comparable disease severity in relation to the in-house database. The concrete voice recordings from mPower study that we have used can be checked in the Appendix, which provides the health codes, unique identifiers provided by mPower. Both databases are also the same size. The comparison allows for evaluation of the performance degradation that might be expected when moving an automatic PD detection system from a controlled mobile scenario to an uncontrolled one. Also, cross-database tests are performed to assess the generalizability of the results.

The novel contributions of this paper can be summarized as follows:

- Performance comparison of a speech-based PD detection approach on two different databases created by using smartphones, one of them recorded under controlled conditions (quiet acoustic environment, professional supervision) and the other one collected without supervision in realistic environments (mPower-based database).
- Cross-database experiments involving the controlled database and the database recorded in realistic environments.
- Methodologically robust analysis based on the following considerations: balanced datasets regarding age and sex, comparable disease stage between datasets, identical methodology (preprocessing, feature extraction, feature selection and six classification algorithms) applied in all the experiments.
- Design and implementation of client–server system architecture: Android-based app and artificial intelligence engine, ready to perform further analysis in semi-controlled clinical trials.

Results

Experimental settings

The methodology proposed in Section is applied to the UEX and mPower-based databases. A total of 100 iterations of stratified 5-fold cross-validations have been used for the feature selection step. For hyperparameter optimization with grid search also a stratified 5-fold cross-validation has been issued. Finally, the classification process consists of 1000 iterations. In each one of them the set is randomly split in training and test subsets with a 75–25% ratio stratified by health status.

Results for UEX database

Table 1 shows the evaluation metrics resulting from applying the machine learning approaches with the considered specifications to the UEX database.

Three out of six approaches (Passive Aggressive, Perceptron, and Support Vector Machine (SVM)) produced accuracy rates greater than 0.9, and Logistic Regression is close to this value. Random Forest and Gradient Boosting showed a downgrade in performance with accuracy rates around 0.75. Sensitivity and specificity are used to analyze how balanced the system is by checking whether PD or healthy subjects are better detected. All of the approaches provided slightly larger sensitivities (right classifications for subjects suffering from PD) than specificities (right classifications for healthy people). However, these differences are small and it can be concluded that all of them are reasonably well balanced.

Figure 1 shows mean receiver operating characteristic (ROC) curves (blue lines) with bands for $\pm$ 1 standard deviation (light gray area) for the six classifiers under consideration. The ROC curve shows the trade-off between false-positive rate (FPR = 1-specificity) in the x-axis and true-positive rate (TPR=sensitivity) in the y-axis. As performance is measured with the area under the curve (AUC) metric, ROC curves closer to the top-left corner indicate a better performance.

Gradient Boosting ROC curve presented in Fig. 1a provides a relatively good AUC mean value of 0.8387 with a standard deviation of 0.0964. Given the shape of the curve the results are far from the optimal classifier (TPR = 1, FPR = 0), and the slow growth indicates that we should face a very high FPR for TPR higher than 0.7. Random Forest ROC in Fig. 1e shows a similar performance, with mean AUC = 0.8787, and the same problem of high FPR for TPR higher than 0.7. On the other side, Logistic Regression (Fig. 1b), Passive Aggressive (Fig. 1c), Perceptron (Fig. 1d) and SVM (Fig. 1f) show a great AUC, well above 0.95 in every case, and a standard deviation that shows a perfect classifier for some of the cross-validation experiments performed. In these cases, the FPR/TPR trade-offs are much more beneficial, with FPR lower than 0.2 for TPR above 0.9 in every case.

Table 2 presents the run times separated by feature selection, grid search and classification. The most time-consuming task for all six classifiers is feature selection, since a very exhaustive recursive feature elimination with cross-validation (RFECV) has been applied, followed by grid search. Finally, classification, applied here with cross-validation, is the least expensive task in terms of computational time. Gradient Boosting and

Table 1 Evaluation metrics (mean $\pm$ standard deviation) obtained with the proposed procedure by using the UEX database

	Accuracy rate	Sensitivity	Specificity	AUC
Gradient Boosting	0.7503 $\pm$ 0.0983	0.7683 $\pm$ 0.1486	0.7331 $\pm$ 0.1697	0.8387 $\pm$ 0.0964
Logistic Regression	0.8897 $\pm$ 0.0820	0.9007 $\pm$ 0.1145	0.8788 $\pm$ 0.1324	0.9627 $\pm$ 0.0522
Passive Aggressive	0.9205 $\pm$ 0.0723	0.9396 $\pm$ 0.1005	0.9018 $\pm$ 0.1108	0.9756 $\pm$ 0.0403
Perceptron	0.9083 $\pm$ 0.0781	0.9284 $\pm$ 0.1030	0.8881 $\pm$ 0.1232	0.9713 $\pm$ 0.0457
Random Forest	0.7631 $\pm$ 0.1024	0.7666 $\pm$ 0.1591	0.7605 $\pm$ 0.1486	0.8787 $\pm$ 0.0821
SVM	0.9148 $\pm$ 0.0853	0.9229 $\pm$ 0.1102	0.9076 $\pm$ 0.1229	0.9749 $\pm$ 0.0483

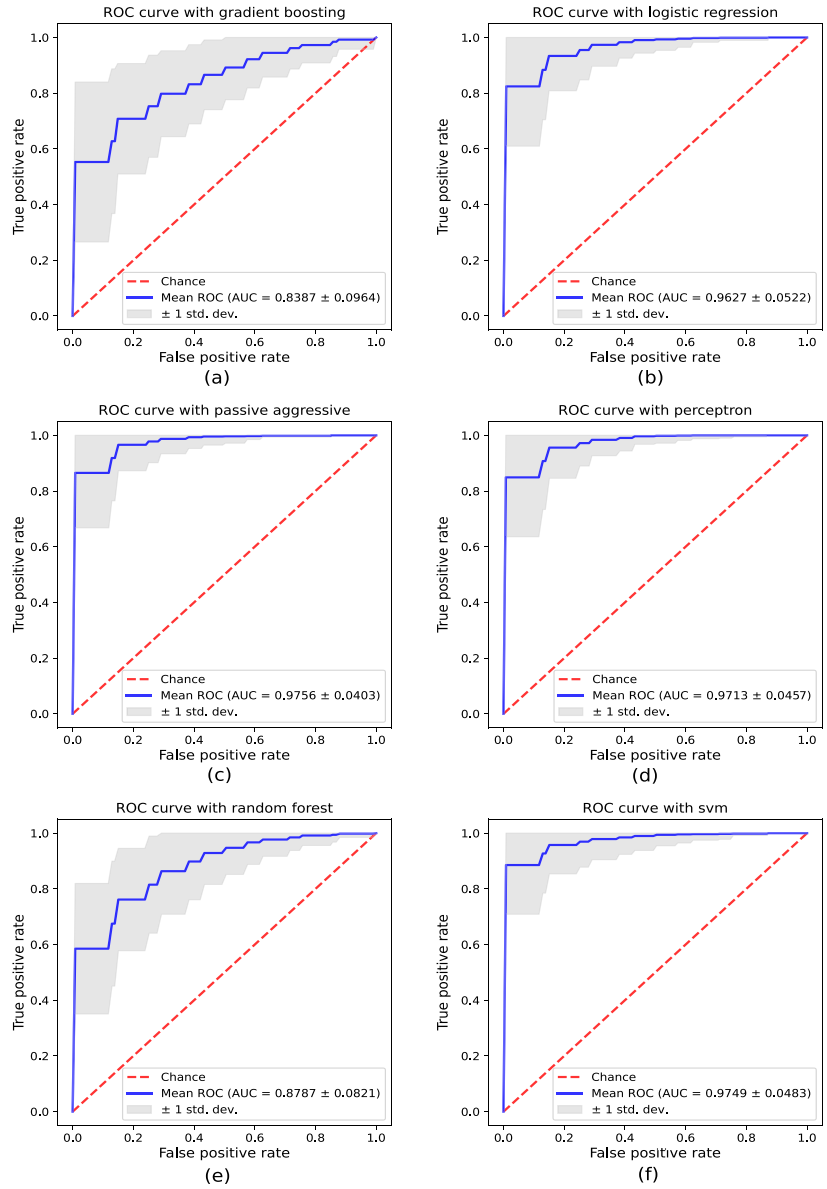


Fig. 1 ROC curves and AUC metric obtained with the proposed procedure by using the UEX database: **a** Gradient Boosting, **b** Logistic Regression, **c** Passive Aggressive, **d** Perceptron, **e** Random forest, **f** Support Vector Machine

Random Forest, which yield the lowest performance, also have the largest execution times. The rest of the classifiers have closer values, all of them with less than one minute for the total run time.

Table 2 Run times in seconds for the different steps of the proposed procedure by using the UEX database

	Feature selection	Grid search	Classification	Total
Gradient Boosting	390.05	318.47	105.69	814.21
Logistic Regression	24.63	21.51	12.28	58.42
Passive Aggressive	23.43	11.97	12.21	47.61
Perceptron	22.17	7.58	12.37	42.13
Random Forest	938.81	286.77	155.31	1380.89
SVM	17.75	14.00	9.16	40.91

Table 3 Selected features for each classifier in the proposed procedure by using the UEX database

	Gradient Boosting	Logistic Regression	Passive Aggressive	Perceptron	Random Forest	SVM	Total
Sex							0
Jitter							0
Shimmer					■		1
LZ-2	■	■	■	■	■	■	6
CPP		■	■	■	■	■	5
Hurst							0
MFS		■			■		2
Shannon							0
Permutation							0
PPE		■				■	2
FMMI							0
FZCF							0
GNE							0
ZCR			■	■	■		3
D2		■	■	■		■	4
HNR					■	■	2
RPDE		■	■	■	■	■	5
GQ_prc5_95							0
GQ_std_cycle_open							0
GQ_std_cycle_closed		■	■	■		■	4
MFCC0		■	■	■		■	4
MFCC1							0
MFCC2	■						1
MFCC3							0
MFCC4		■	■	■	■	■	5
MFCC5	■		■		■		3
MFCC6							0
MFCC7							0
MFCC8		■	■	■	■	■	5
MFCC9		■	■	■		■	4
MFCC10					■		1
MFCC11		■	■	■		■	4
MFCC12			■	■			2
Total	3	12	13	12	11	12	

Table 3 summarizes the results from the feature selection process, providing a global perspective about which features are the most relevant for each approach. Checking the number of times each feature has been selected, it can be determined that Lempel–Ziv complexity (LZ-2), Cepstral Peak Prominence (CPP), Period Density Entropy (RPDE), and 4th and 8th Mel Frequency Cepstral Coefficients (MFCC) are the most selected features. Specifically, the most chosen feature is LZ-2, which is the only one selected by all the approaches. Conventional features like Jitter, Shimmer or HNR are not very relevant. Gradient Boosting only selected three features, but it performs badly in accuracy metrics and run time results. The rest of the classifiers selected a similar number of features and chose the five most relevant ones (LZ-2, CPP, RPDE, MFCC4, and MFCC8).

In summary, the best result is obtained with the Passive Aggressive approach. It produces the largest accuracy rate (0.9205) and AUC (0.9756), with the lowest standard deviations (0.0723 and 0.0403, respectively). Besides, its computing time is low. SVM and Perceptron approaches are also very competitive in accuracy metrics and computing time. Any of these three approaches could be considered for the mobile-assisted system to detect PD.

Results for mPower-based database

The same experimental settings and methodology applied to UEX database is applied to this matched database based on mPower study. Table 4 presents the accuracy metrics. T-tests reported statistically significant differences (p-values < 0.001) for comparisons of each accuracy metric and method between UEX database and mPower-based database.

Accuracy rates are much lower than in the case of UEX database, ranging from 0.6167 to 0.7138. The best approach is based on Gradient Boosting classifier. This means that the accuracy rates have been degraded for all the approaches. In percentage terms, the reductions with respect to UEX database range from 4.9% to 33.0%. Analogously, sensitivities and specificities are also degraded, with reductions ranging from 3.4% to 35.1%, and from 6.3% to 30.7%, respectively. Sensitivities and specificities are close for most of the approaches when applied to mPower dataset.

Figure 2 shows the ROC curves (blue lines) with bands for $\pm$ standard deviation (light gray area). Superiority of the ROC curves in Fig. 1 with respect to Fig. 2 can be seen at a glance. Following the AUC criterion, the best approach is also Gradient Boosting, but its AUC value is only 0.7449. In fact, the AUC values range from 0.6923 to 0.7560, which means reductions of AUC between 9.9% and 28.9% with respect to the UEX database. Every classifier but Gradient Boosting (Fig. 2a) produces an AUC under 0.75, though the latter slightly exceeds that value, making it the best option. In every case, the trade-off between FPR and TPR is quite low. It is worth noting that the curve does not reach TPR $= 1$ in any case, no matter the threshold. Also, Passive Aggressive and Perceptron (Fig. 2c and d) are near random classification, given that standard deviation shows that, in the worst cases, AUC stays as low as 0.5.

It is remarkable that the standard deviations of the metrics are greater in the case of mPower-based database in spite of the fact that the mean values are lower than those of the UEX database. This means that the approaches provide more dispersed values with mPower-based dataset, and therefore the results obtained with UEX dataset are more robust.

With respect to the computing time, the results match those obtained with UEX database. Table 5 shows the computing times separated by tasks. There are two approaches, Gradient Boosting and Random Forest, that have large computing times. The other four approaches keep their execution time below one minute for the whole process.

Table 4 Evaluation metrics (mean $\pm$ standard deviation) obtained with the proposed procedure by using the mPower-based database

	Accuracy	Sensitivity	Specificity	AUC
Gradient Boosting	0.7138 ± 0.1051	0.7419 ± 0.1712	0.6868 ± 0.1665	0.7560 ± 0.1147
Logistic Regression	0.6523 ± 0.1101	0.6530 ± 0.1961	0.6525 ± 0.1910	0.7330 ± 0.1258
Passive Aggressive	0.6167 ± 0.1167	0.6096 ± 0.2168	0.6247 ± 0.2141	0.6935 ± 0.1349
Perceptron	0.6245 ± 0.1179	0.6334 ± 0.2211	0.6164 ± 0.2150	0.6923 ± 0.1411
Random Forest	0.6957 ± 0.1048	0.7123 ± 0.1659	0.6823 ± 0.1664	0.7475 ± 0.1110
SVM	0.6562 ± 0.1122	0.6476 ± 0.2047	0.6657 ± 0.1879	0.7437 ± 0.1240

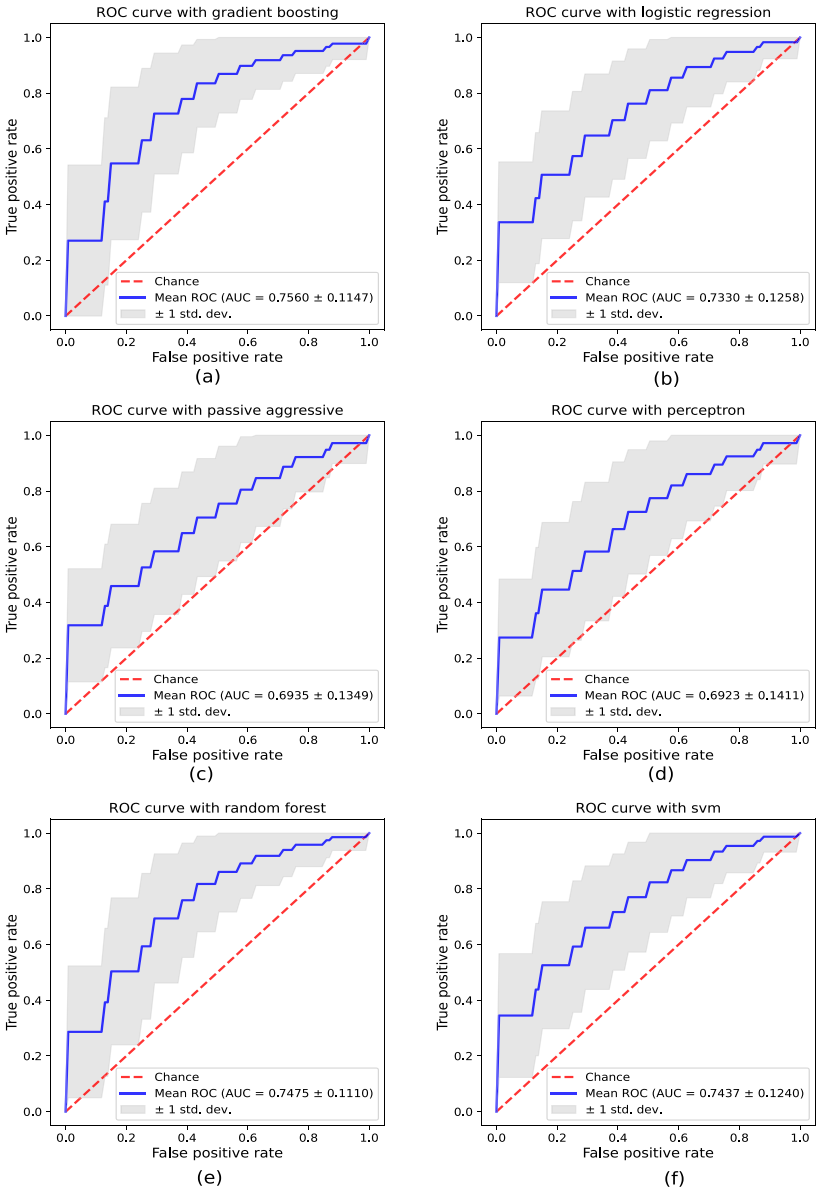


Fig. 2 ROC curves and AUC metric obtained with the proposed procedure by using the mPower-based database: **a** Gradient Boosting, **b** Logistic Regression, **c** Passive Aggressive, **d** Perceptron, **e** Random forest, **f** Support Vector Machine

Finally, it is remarkable that the feature selection processes have provided different results than those of UEX database. Table 6 shows the selected features for each approach. Sex, Shimmer, MultiFractal Spectrum Width (MFSW), Glottal Quotients

Table 5 Run times in seconds for the different steps of the proposed procedure by using the mPower-based database

	Feature selection	Grid search	Classification	Total
Gradient Boosting	392.43	379.88	24.64	796.95
Logistic Regression	19.22	15.81	9.58	44.60
Passive Aggressive	18.27	8.82	8.99	36.09
Perceptron	17.08	5.48	9.62	32.18
Random Forest	939.55	281.49	82.51	1303.55
SVM	18.61	13.59	9.21	41.42

Table 6 Selected features for each classifier in the proposed procedure by using the mPower-based database

	Gradient Boosting	Logistic Regression	Passive Aggressive	Perceptron	Random Forest	SVM	Total
Sex		■	■	■		■	4
Jitter							0
Shimmer	■	■	■	■	■	■	6
LZ-2							0
CPP							0
Hurst							0
MFS		■	■	■		■	4
Shannon	■		■				2
Permutation							0
PPE							0
FMMI							0
FZCF							0
GNE							1
ZCR							0
D2					■		1
HNR							0
RPDE	■				■		2
GQ prc5 95		■	■	■		■	4
GQ std cycle open		■	■	■	■	■	5
GQ std cycle closed	■				■		2
MFCC0	■				■		2
MFCC1							0
MFCC2	■				■		2
MFCC3	■						1
MFCC4							0
MFCC5		■		■		■	3
MFCC6		■	■	■	■	■	5
MFCC7					■		1
MFCC8	■				■		2
MFCC9							0
MFCC10							0
MFCC11							0
MFCC12							0
Total	8	7	7	7	10	7	

(GQ prc5-95 and GQ std cycle open), and MFCC6 have been the most selected features, being Shimmer selected by all the approaches. MFCCs have also been selected with UEX database. The number of selected features range from 7 to 10.

Cross-database tests

In this type of experiments, we use the selected features and hyperparameter values obtained in a single-database experiment and test the performance on the other database.

Table 7 shows the results obtained when the selected features and hyperparameter values obtained from UEX database are applied on mPower-based database. It can be observed that the detection capability has been lost, with a result close to random classification. Specifically, the degradation can be quantified with a reduction percentage with respect to the results obtained with the UEX database in 27.7–45.5% for accuracy, 31.0–49.9% for sensitivity, 30.0–43.1% for specificity, and in 33.7–48.4% for AUC. This indicates that it is not recommendable to train the system with a controlled database if it is going to be applied on an uncontrolled scenario.

The results obtained using the reverse procedure are shown in Table 8. In this case, the selected features and hyperparameter values are obtained from the mPower-based

Table 7 Evaluation metrics (mean ± standard deviation) obtained by selecting features and hyperparameter values from UEX database and testing the performance on mPower-based database

	Accuracy	Sensitivity	Specificity	AUC
Gradient Boosting	0.5234 ± 0.1139	0.5358 ± 0.1827	0.5131 ± 0.1912	0.5377 ± 0.1294
Logistic Regression	0.5380 ± 0.1233	0.5376 ± 0.2024	0.5393 ± 0.2036	0.5569 ± 0.1495
Passive Aggressive	0.5021 ± 0.1243	0.4706 ± 0.2092	0.5357 ± 0.2130	0.5036 ± 0.1548
Perceptron	0.5289 ± 0.1205	0.5267 ± 0.2019	0.5334 ± 0.2027	0.5522 ± 0.1452
Random Forest	0.5519 ± 0.1245	0.5286 ± 0.1956	0.5818 ± 0.1980	0.5822 ± 0.1474
SVM	0.5230 ± 0.1209	0.5308 ± 0.2023	0.5166 ± 0.2025	0.5442 ± 0.1432

Table 8 Evaluation metrics (mean ± standard deviation) obtained by selecting features and hyperparameter values from mPower-based database and testing the performance on UEX database

	Accuracy	Sensitivity	Specificity	AUC
Gradient Boosting	0.6165 ± 0.1046	0.6260 ± 0.1786	0.6089 ± 0.1736	0.6664 ± 0.1216
Logistic Regression	0.6022 ± 0.1175	0.5940 ± 0.2138	0.6114 ± 0.1985	0.6495 ± 0.1426
Passive Aggressive	0.5302 ± 0.1262	0.5877 ± 0.2529	0.4738 ± 0.2426	0.5446 ± 0.1625
Perceptron	0.5877 ± 0.1258	0.5925 ± 0.2219	0.5849 ± 0.2142	0.6322 ± 0.1539
Random Forest	0.6421 ± 0.1003	0.6717 ± 0.1749	0.6152 ± 0.1664	0.6851 ± 0.1216
SVM	0.6053 ± 0.1142	0.6033 ± 0.2062	0.6074 ± 0.2024	0.6511 ± 0.1416

database and tested on UEX database. Now, the reduction percentage with respect to the results obtained with the mPower-based database are in 7.7–13.6% for accuracy, 5.7–11.6% for sensitivity, 6.3–24.2% for specificity, and 8.3–21.5% for AUC. In spite of the low performance, the results are better than in the previous experiment. This indicates that system robustness is increased when a variety of acoustic conditions is used to determine the feature set and hyperparameter values, and they are applied to voice recordings fulfilling a very strict recording protocol.

Discussion

In this study, we have proposed a methodology to discriminate PD patients from healthy subjects based on sustained phonations of /a/ vowel recorded by a smartphone. We applied feature extraction, data standardization, feature selection, hyperparameter optimization, and six different classification techniques. The results obtained when applying this methodology to recordings obtained under controlled conditions (protocol supervised by specialized staff, same recording room and same smartphone) have been presented first.

Under these controlled conditions, the procedure has allowed to identify a set of features that provide good performance using accuracy, sensitivity, specificity and AUC metrics. The results demonstrate the relevance of LZ-2 and RPDE. The high ability for PD discrimination of these and other features based on nonlinear dynamics has been noted by other authors (see e.g., Orozco-Arroyave et al. [36]). It is also remarkable the role played by CPP which, as opposed to classic features such as Jitter, can be robustly extracted even from strongly aperiodic signals like those obtained from PD patients with a severely affected voice. It is also known the huge potential of MFCCs for different

classification applications based on speech. They have been previously used for PD detection by Sakar et al. [45]. MFCCs allow for capturing differences in the resonant characteristics of the vocal tract. It has been reported that patients with PD present an asymmetric centralization of tongue position during the phonation of vowels, which produces a decrease in the vowel space area in comparison to healthy speakers [2]. This can explain the high number of MFCCs present in the subsets of selected features that result from our study.

With UEX database, the best results have been achieved using Passive Aggressive classifier: 0.9205 in accuracy rate, 0.9396 in sensitivity, 0.9018 in specificity, and 0.9756 in AUC. Placing these results in the context of the literature is a complex task since a real comparison of methodologies would require working on the same databases, or at least on databases with comparable disease stages which also ensure age and sex balance. To the authors' best knowledge the published scientific work does not allow for a comparison that fulfills these three requirements. However, in the next paragraphs we provide a rough overview of the performance obtained using professional microphones and smartphones.

In the case of professional microphones, in a recent work, Solana-Lavalle et al. [46] compare their accuracy rate (0.94) with other scientific works presenting values between 0.85 and 1. In the case of databases based on smartphone recordings, Almeida et al. [1] use sustained vowel recordings and a similar methodology than ours: feature extraction and classification process with 2/3 training and 1/3 test ratio for cross-validation. They achieve 0.9294 of accuracy rate and 0.9240 of AUC by using 1-nearest neighbor classifier with smartphone recordings. The health status of PD patients was evaluated at stages 1 to 2.5 according to HY (Hoehn and Yahr) scale. The experimental design was not age-balanced, since the mean age of PD patients was 61.5 years, while the mean age of healthy subjects was 41.8 years. Rusz et al. [42] recorded different vocal tasks including sustained vowels with a professional microphone and a smartphone. The experiment was well balanced in terms of age and sex. The mean HY stage was 2.1 (0.4) in comparison to 2.7 (0.53) in this study. Their methodology is based on the extraction of 6 acoustic features and the use of Logistic Regression with Leave-One-Out cross-validation for classification. They achieved an AUC of 0.85 for smartphones. Zhang [57] proposed a smartphone-based PD detection service by using a deep learning methodology based on stacked autoencoders and K-Nearest-Neighbor classifier achieving a maximum accuracy value of 0.9881. However, this can not be considered a complete smartphone-based system since their experimental results were not obtained from recordings made by mobile phones. Instead, they used already extracted features from publicly available datasets.

Once the potential of our methodology to perform automatic detection of PD has been proved on a controlled scenario, the next step is applying the same techniques in an uncontrolled one, therefore, we considered mPower database [7]. It must be pointed out that this database has been massively collected. As a consequence, it contains some faulty recordings that would not pass a simple playback quality check performed by the majority of the users if they were immersed in a real clinical scenario. Also, it includes some inconsistencies in diagnosis, having recordings from the same subject labeled as PD affected and healthy. In order to issue a valid comparison with it, a previous work has been done to select recordings from the database which provided a balanced set by

sex, age, and disease stage. The results show a best accuracy rate of 0.7138 with sensitivity of 0.7419, specificity of 0.6868 and AUC of 0.7560 for Gradient Boosting versus a best accuracy rate of 0.9205 with sensitivity of 0.9396, specificity of 0.9018 and AUC of 0.9756 for Passive Aggressive with the UEX database. This has provided statistically significant differences for the four accuracy metrics (p-values < 0.001). This shows a clear degradation in the accuracy performance in comparison to UEX database that is not only reported for the best methods, but for all ones. In this case, using mPower-based database produces a performance degradation of 22.5% for accuracy rate, 21% for sensitivity, 23.8% for specificity and 22.5% for AUC.

The aforementioned difficulties arise again when these results are intended to be placed in the context of the scientific literature, because previous works based on mPower database do not use exactly the same subset of recordings. Since the database has been massively collected, experiments based on large cohorts have been performed. For example, with a subset of mPower database consisting of 2222 phonation recordings, 933 PD patients and 1289 healthy subjects, Giuliano et al. [14] obtained AUC values over 0.82 in the discrimination of PD subjects from healthy ones. Their methodology was based on Neural Networks and Logistic Regression models. Wroge et al. [55] reached a maximum accuracy rate of 0.86 by using Minimum Redundancy Maximum Relevance for feature selection and Gradient Boosted Decision Tree for classification, with a total of 5826 voice recordings. Tougui et al. [48] achieved an accuracy rate of 0.9578 by using Least Absolute Shrinkage and Selection Operator feature selector, hyperparameter tuning, and Extreme Gradient Boosting classifier with 18210 recordings (9105 PD patients and 9105 healthy subjects). In these works based on large cohorts, sex and age balances between PD and healthy groups are not ensured in the experiments.

The application of an identical methodology to both databases has allowed for checking the differences that can be expected when moving from a controlled scenario to an uncontrolled one. As previously mentioned, a clear degradation in the detection performance can be noted, but there are also differences concerning the selected features and the best classifier. In terms of selected features, with the exception of Gradient Boosting, the results obtained with UEX database show a good stability when varying the classification method. A similar conclusion regarding stability across classifiers can be extracted from the results obtained on mPower-based database, which means that the database plays a more important role than the classification method. On mPower-based database the most relevant features are: Sex, Shimmer, MFSW, GQ std cycle open, GQ prc5 95 and MFCC6. Although the features are different for each database, we can identify some common aspects. For example, if we consider the most repeated features, in both cases the role is shared by features that are able to capture source-related irregularities considering the classical source-filter theory of speech production (CPP in the case of UEX database, GQ std cycle open and GQ prc5 95 in the case of mPower-based database), resonance-related features (MFCCs) and features based on nonlinear analysis (LZ2 and RPDE in the case of UEx database and MFSW in the case of mPower-based database).

A limitation of our work is the size of the databases. The reason is the difficulty in recruiting people suffering from PD in the case of the controlled database (UEX database). Nevertheless, 60 people (30 with PD and 30 healthy controls) is a reasonable size

compared to other studies in the scientific literature. For example, in Benba et al. [4] the number of participants is 40 (20 with PD and 20 healthy); in Little et al. [24], this number was 31 (23 with PD) and in Novotny' et al. [34] the total number was 80 (40 with PD and 40 healthy).

Regarding computation time, the executions on both databases yield similar conclusions in the comparison of classifiers. In a real clinical application, the first two tasks will be only applied from time to time to improve the learning process, so that both the selected features and the searched hyperparameters will be used during a long time. Furthermore, the third task, classification, is applied here with cross-validation, but in real time approaches it will be applied only to the new subject. For all these reasons, computation time is not a critical issue. Anyway, even for model assessment purposes, the experiments have been performed in a very reduced time.

Due to the differences in the selected features found in the single-database experiments, we have performed cross-database tests, in which the feature set obtained for each classifier with one database has been applied to the other one. Although we observe an important degradation in performance in both cases, the results are slightly better when feature selection is performed on mPower-based database and applied to UEX database than when using the reverse procedure. The wide variety of acoustic conditions available in mPower database due to the fact that the recordings were performed by the participants themselves is considered a strength that could be exploited to achieve robustness. However, it must be taken into account that, since this database has been massively collected, some information provided by the participants may be incorrect and some voice recordings may be of bad quality, having an impact on the performance. Some research initiatives point out that personalized medicine and collaboration between patients and health professionals might provide a greater insight in disease impact by allowing patients to provide and self assess their condition outside clinical environment [22]. Therefore, a semi-controlled scenario appears as a very suitable option. This means that the participants would submit their audio files, recorded by following a strict recording protocol in a variety of acoustic conditions, but the clinical information is provided by the physicians. The proposed mobile-assisted system is considered a very useful tool to address this semi-controlled scenario.

Conclusion

Smartphones have a great potential to assist diagnosis and improve patient monitoring of many diseases. In the case of PD, smartphones allow for an easy collection of speech waveforms that can be used with clinical purposes. This can help general practitioners to conduct triage and neurologists or movement disorders specialists to perform diagnosis and tracking. In particular, PD management could be highly benefited by smartphone-based systems, due to different aspects such as increasing incidence, diagnosis prone to errors, difficulty of tracking progression, and the fact that it mostly targets elderly people, which in general have more difficulties to visit a hospital, among others.

We have designed and implemented a mobile-assisted voice condition analysis system based on an Android app frontend in conjunction with a machine learning-based implementation hosted on a computing server backend. Although the machine learning approach is focused on a detection task, the app allows for monitoring PD progression.

The most relevant novel contribution of our work is that we have applied identical methodology to an in-house smartphone-based database recorded under controlled conditions (in a quiet room with low noise level and with professional supervision of the recordings) and to a subset of mPower database (created by collecting data from free-living scenarios). This comparison of results is performed within a methodologically robust framework ensuring age and sex balance and comparable disease stage. The results of this study show the potential of the proposed system under controlled conditions. The performance decreases when testing the methodology with the uncontrolled database and strongly drops in cross-database tests.

These results confirm the research hypothesis and suggest that semi-controlled scenarios have high potential to be useful in real clinical applications. In these semi-controlled scenarios the relevant clinical information is provided by the physicians. Also, general practitioners (in the context of triage for diagnosis) or patient and caregivers (in a PD monitoring application...) should receive some initial training after which a test should be mandatory to ensure that the speech protocol is fully understood and that the user has some control on the acoustic environment regarding noise level. Within this framework, recordings would be submitted via smartphone from different environments.

Future analyses should be performed on new datasets obtained in the described semi-controlled clinical scenarios. The proposed app is a very suitable tool for this task because it allows patients to submit phonations and physicians to check the complete record of every patient. In those semi-controlled conditions, also longitudinal studies would be interesting for PD tracking. This type of studies are difficult to perform because they require larger amounts of time. However, they would be very useful to achieve optimal treatment of PD.

Methods

A mobile-assisted voice condition analysis system for PD detection is proposed. This system is built through the design and implementation of a mobile application that communicates with a server backend to collect and process voices recorded following a protocol. The system extracts acoustic features from the voice recordings and use them to feed machine learning approaches specifically designed for a PD detection task. An experiment has been conducted to test the proposed approaches. Also, the same architecture was used on a different database collected using smartphones and results are compared. In this section, the several parts that compose the system are described.

System architecture and mobile app design

Voice recordings are received and stored in a server where they can be accessed and processed. The server runs Windows 10 and the Windows Internet Information Services (IIS) functionality has been employed to host a web service written in PHP that manages a MySQL database. An Android app exchanges information with the server via an HTTP connection, using JavaScript Object Notation (JSON) format to organize it. Figure 3 shows the system structure in a schematic way.

The Android application has two types of user accounts: patients and doctors. Every user needs to fill a registration form with the most relevant personal information, some of which will be used for the authentication. This form is slightly different for patient

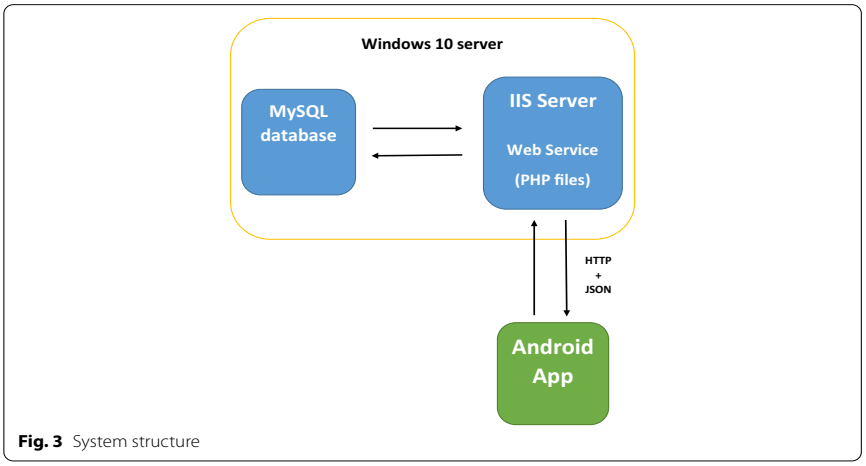


Fig. 3 System structure

and doctor accounts. The user receives a notification in the email account provided after the registration process. It is necessary to give permission for the use of personal data as part of a non-profit study. In the patient case, as part of the registration process, an informed consent document is requested to be signed by accepting participation in the mentioned study. Users can sign the document through the phone's touch screen.

Once completely registered, users can employ their credentials (email and password) to access the functionalities allowed for the type of account created. On the one hand, patient accounts are able to record and send audio files following the given instructions. After each recording, the user can choose between three options: submit to the server, listen or discard and try again. On the other hand, doctor accounts can associate patients with their account to keep track of their cases. Only doctor accounts can access patient data, and only those linked to the doctor account. Figure 4 shows some screenshots extracted from the app: Fig. 4a shows the registration and login screen of the system; Fig. 4b shows the screen that allows to select the type of account (patient or physician) in the registration process; Fig. 4c presents the screen showing the instructions the patient should follow to perform the recordings; after that, three possibilities (listen, send, discard) are offered to the patient, as shown in Fig. 4d.

Participants

Two databases were used in the study. The first one was generated by the University of Extremadura with the collaboration of the Regional Association for Parkinson's Disease of Extremadura (UEX database). A total of 60 participants with ages between 51 and 87 years old were recruited, 30 of whom were affected by PD (PD subjects) and 30 were healthy. Patients suffering from PD were recruited among the voluntary members of the Regional Association for Parkinson's Disease of Extremadura that meet the following inclusion criteria: (1) have a definitive diagnosis of PD; (2) medical reports available. After the voluntary PD patients were recruited, then the healthy group was selected to approximately match sex and age. Healthy subjects were selected with the requirement

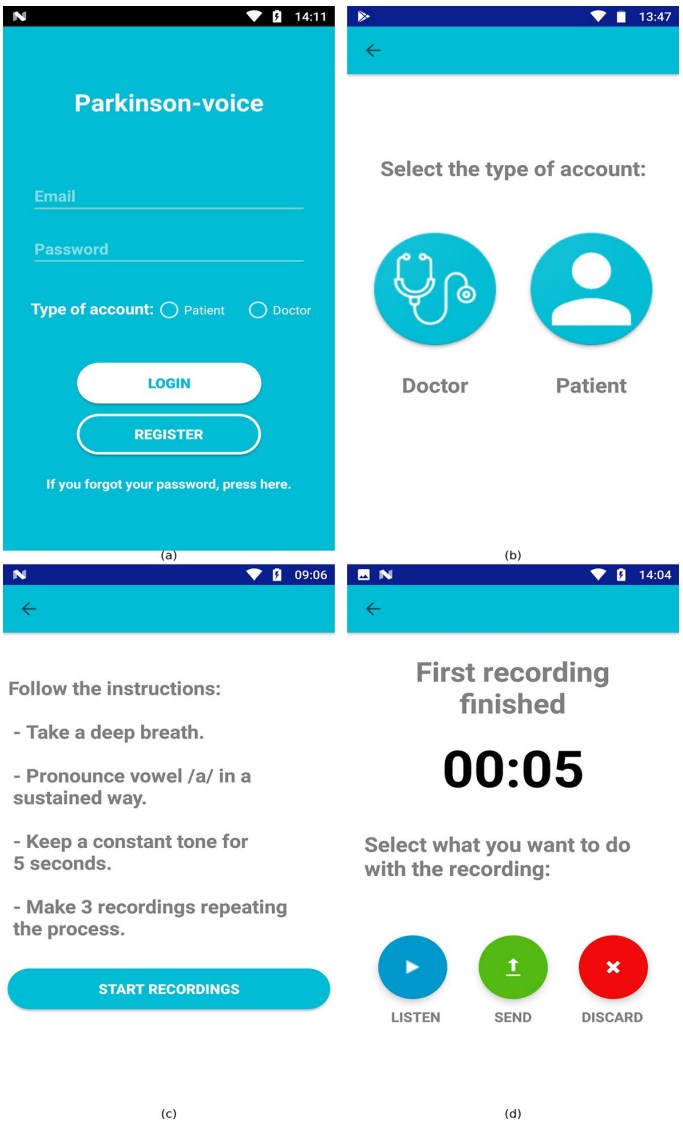


Fig. 4 Screenshots obtained from the Android mobile application: **a** first screen, **b** types of accounts, **c** instructions, **d** recording process

of neither having been diagnosed with PD nor having any symptom related to PD. Those not meeting the inclusion criteria were not eligible for participation. There were 24 men and 6 women in the PD group and 26 men and 4 women in the healthy group. The mean (standard deviation) of the age was 70.27 (9.54) for the PD group and 67.33 (8.57) for the healthy group. The mean time in years since diagnosis was 9.93 (6.16), and the mean

time in hours since the last medication dose was 2.21 (1.32). The mean HY stage was 2.6 (0.4). The research protocol was approved by the Bioethics Committee of the University of Extremadura. All of them signed an informed consent.

The second database (mPower-based database) is a subset extracted from the mPower Public Researcher Portal, a mobile PD study [7]. The goal of this initiative is to collect information of patients suffering from PD. The objective is to describe more precisely the experience, habits, lifestyle, drawbacks, and interactions with medication of those patients. By using a mobile application, each volunteer records different aspects of the impairment caused by the disease and tracks their evolution. The study is open to anyone who wants to participate, and the only requirement is having a personal iPhone for PD patients, and also not having been diagnosed for the control group subjects. These requirements are not checked.

The subjects selected to build the mPower-based database were matched with the ones from the UEX database by keeping exactly the same proportion of health status and sex, and approximately the same age, so the results can be compared. Specifically, the mean of the age was 68.36 (8.14) for the PD group and 65.23 (7.76) for the healthy group. The mean time in years since diagnosis was 7.83 (4.54), whereas the estimated mean HY stage was 2.7 (0.53). The mean time since the last medication dose was not available. The voice recordings were stored for posterior use. Table 9 shows the codes of these voice recordings extracted from mPower.

Recording task and equipment

The selected vocal task was sustained phonation of /a/ vowel due to several advantages, such as its wide spread use in the scientific literature; simplicity to realize by the participants, which avoids fatiguing them, especially in the case of patients with more advanced PD stages; ease of analysis and control; ubiquity in different languages; and the fact that it is unaffected by phonetic context or intonation [12].

The recording task for UEX database consists of performing three 5-seconds voice phonations, pronouncing the /a/ vowel in a continuous and uninterrupted way holding pitch and loudness as constant as possible.

Due to the biological variability, voice recordings from a particular subject result in similar but not identical waveforms. The consequence is that the features are also not identical when extracted from different recordings from the same individual. To obtain more stable predictors, it was decided to record three utterances per subject so that the feature values can be later averaged to produce an only feature vector per subject.

All the voice recordings were made using the same smartphone (model BQ Aquaris V) at a sample frequency of 44.1 kHz. The recordings were taken at the facilities of the Regional Association for Parkinson's Disease of Extremadura (Spain), always in the same room, that was relatively quiet but did not have any special acoustical isolation. A specialized person was present to ensure that all the participants properly followed the voice recording protocol and registered the complementary information based on medical reports.

Voice recordings from mPower were performed on participants' iPhones (4th generation or a more advanced version) or iPods (5th generation or newer) by using the /a/ vowel phonation protocol. A sample frequency of 44.1 kHz was used. Since participants

record themselves without supervision, this database includes a variety of acoustic environments. They were also responsible to fill in the form including the complementary information, which makes the obtained data somehow unreliable.

Before applying feature extraction, all the recordings from both databases were trimmed down to one second discarding any leading or trailing silence. This length has been considered sufficient to extract speech features from sustained vowel phonations by other authors [40]. Voice recordings were edited using Audacity software (release 2.0.5).

Feature extraction

The same feature extraction algorithms are applied to both databases. A total of 33 features have been considered to measure different aspects related to speech production: Sex (male, female), Jitter, Shimmer [51], CPP [34], HNR, glottal-to-noise excitation ratio, zero crossing rate [3], 3 GQ features [45], MFCCs (13 features) [52], correlation dimension, RPDE, pitch period entropy [51], Hurst's exponent, LZ-2 [36], permutation entropy, Shannon's entropy, first minimum in mutual information [25], MFSW [17], first zero in correlation function [16]. The methods have been coded in Python.

Considering these feature extraction algorithms, 180 vectors (60 subjects $\times$ 3 audio recordings/subject) of 34 feature components (health status plus extracted features) were initially stored in a spreadsheet for UEX database. This spreadsheet was reduced to 60 vectors of 34 features by aggregating every 3 vectors corresponding to the same subject through a component-wise average. This ensures that each subject is represented by only one feature vector and no artificial increase of the dataset is considered. In the case of the mPower-based database, 60 vectors of 34 feature components were stored in another spreadsheet. These datasets were used to feed the machine learning approaches.

Statistical methods

Due to the amount of features, many of them measured in different scales, a preprocessing step is required. A standardization was applied based on the mean and standard deviation of each feature.

Several classifier methods have been considered to test their performance in this context. They cover a wide range of techniques commonly used in machine learning applications such as linear methods (Logistic Regression [33]), ensemble decision trees (Random Forest [9]), neural networks (Perceptron [31]), online learning (Passive Aggressive [32]), additive models (Gradient Boosting [21]) and separating data models (SVM [13]).

In order to compare the performance of the procedure with each classifier, and based on the confusion matrix, the following metrics have been considered: accuracy, sensitivity, specificity, and AUC. Student's t-test for independent samples were applied to report statistically significant differences between mean values of accuracy metrics. P-values smaller than 0.05 were considered statistically significant.

Figure 5 represents the whole procedure. After preprocessing, the machine learning approaches contain 3 steps: Feature selection, hyperparameter optimization, and classification process (steps 5, 6 and 7). The involved techniques have been coded in

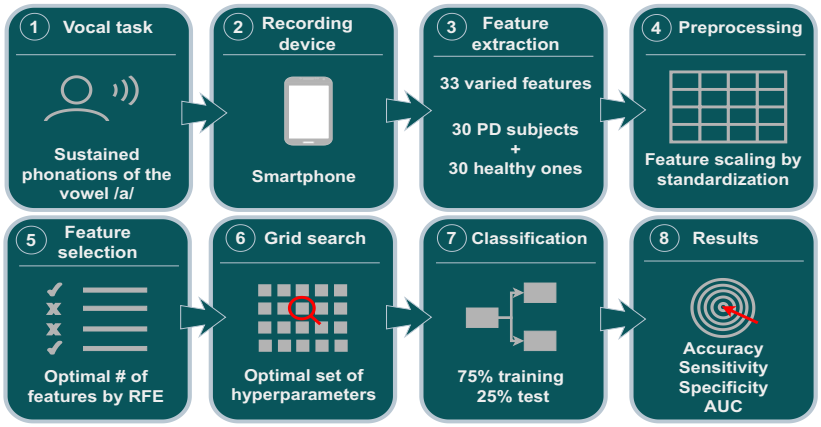


Fig. 5 Methodology followed in the procedure

Python based on the scikit-learn package [15]. Next paragraphs provide a detailed description of these three steps.

Feature selection Once having a standardized dataset, a feature selection process is applied. RFECV [27] is used to eliminate redundant features while keeping a good classification performance. The algorithm trains the chosen classifier and removes the feature with the weakest effect on the classification process, providing a feature top-ranked list based on the order of removal. It provides the optimal number of features by selecting the top-ranked features of the mentioned ranking. The process is repeated several times in order to achieve a representative value. Since the number of optimal features can vary in each iteration, the result of each iteration is stored in a vector and the value of the first quartile after all iterations is chosen as the final number of selected features. A stratified k-fold cross-validation [5] is used in the RFECV algorithm, which consists in splitting the complete dataset in k groups but maintaining the same ratio between PD subjects and healthy ones in each group.

Hyperparameter optimization Each classifier has its own parameters that can be adjusted, these are called hyperparameters. Once the most relevant features are known for the chosen classifier, a hyperparameter tuning has been issued in order to know which is the best configuration for the classifier. The method selected for this step is Grid Search [47]. It optimizes the chosen hyperparameters using stratified k-fold cross-validation again. Accuracy is calculated for each combination of classifier parameter values, selecting the set that provides the best result. These values are used in the classification process.

Classification process With the selected features and the optimal hyperparameter configuration for each classifier, a stratified cross-validation is issued. The dataset is randomly split into only a training and a test subset, maintaining the ratio between the number of PD and healthy subjects in each set. In order to maintain training and test sets independent from each other, the scaling is applied after this splitting with

respect to the training set values. With this splitted data, the classifier is fitted with the training data and after that, it makes a prediction of the PD-healthy state given the test subset. Finally, its predictions are compared with the correct labels. Based on this comparison, the considered metrics are extracted for each iteration. In order to obtain global results, this classification process is repeated several times and the resulting accuracy metrics are averaged after all iterations are finished.

At the end, for each approach, the following are available: the selected features, the optimal hyperparameter configuration, the averaged accuracy metrics, and run times.

Appendix: Codes of voice recordings from mPower

In this appendix, the codes of the voice recordings that have been considered from mPower database are presented in Table 9.

Table 9 Codes of considered voice recordings from mPower

Recording ID	
Healthy	**PD**
0f81a5ef-14d4-4a19-9d89-deabeb728adb	45155beb-a91f-4bca-8296-7612c6915af8
7c5a339d-35ba-48ec-8447-f51aec949a1e	955aa8c3-9116-43e7-9e4b-d1843be4839a
ebfb61fc-c218-4d3a-a680-eb3b4ce3b91d	4412716d-e1b0-4572-b976-8bcb7669925e
b3c61a60-acff-426b-aaeb-d8b6d4c31cb6	0ce23959-8092-47ce-b394-0f65c951a548
740240f3-6752-456b-9f39-6ede3afb3423	a86b7dee-759d-452c-86b5-4b6a248d7286
be0ecb7f-95a2-468a-a12e-2fb738c9b922	9e03615f-1f52-4a95-94bf-cc5805d0c3b8
18cd4553-1c4f-4f6d-a622-8951eb79e780	e2766ec9-e97d-4224-81a8-35b095ea9fd6
3accca87-eaf1-4219-b0e0-af29eb426093	22ad855e-1c57-4f9b-bf67-2a44f2a3ce41
f908e76b-b4e1-40b6-86a5-b4a0def0e6c0	7eac5187-e241-4f80-b704-0f91b8041dc6
75ad7180-afb1-49ea-b766-221106d32e02	0d1c8246-8e42-45e5-b662-91e26e6cb6d4
a3907344-70e3-410c-a6ac-3ae5e790d3ad	02ed9d30-620f-4c6c-88ce-64a286df79b9
393a367c-9727-4390-96f8-6a7a3c6e2797	90899edf-a289-4557-aff9-a168fd82a92e
6348a018-d039-4c38-8920-66ceba01c8e0	06e8ee83-0e3a-4575-a7e4-0c1c813376b6
2fabaecf-423b-4db1-98e6-54daf6844a2d	2b72e6d8-9963-4edd-a8ca-ae2d4262f640
8fa63734-04cb-4f15-a954-34db4d0c9d2e	eb764994-17ef-4421-b052-9acbb0440a3b
15791b9e-89c9-421b-be3c-c3acf89bd167	a9b6687a-c533-410e-8f87-c319a969b98e
4366e9a8-292c-48a2-afa2-d6cbbbf438a9	b662bb1d-ab78-479e-86c8-7fc1bd1df59d
b3277c31-add4-40ae-8621-54da00f50012	1864ea1c-b861-49c3-85f8-549ba6c04679
a467eb63-7f6b-4dee-800f-ee053f0f5d90	2e4b8613-3bab-4cb7-a569-47b52a45a3f9
dcc7e425-7b58-4a04-998a-34822c68cb81	e8a9288b-bdc1-4f09-9f7c-3937d56a4d7f
2df7b01c-d48f-404a-9c09-acdce4cab75c	303e5481-66af-4ba8-bb7e-3dfef44b588b
7c1728ca-408f-4c6c-9d28-94dd61313c65	af9163dc-93e1-4b57-9195-86f6b8ff6725
59ee208f-181e-4d67-9b1f-888cd5036e87	f20af903-16e2-413d-8826-26fc7b51ef38
a17e3358-441b-484a-bac7-868a82784cf6	c79e662a-493c-4d56-9216-b7edd9b4e682
b35db6a8-cff0-4755-9969-3a34a3fc46c7	992b993e-7de6-473a-ae08-d5048a8fb143
54d0e506-71bd-4d27-bc5a-9a360e5b1048	13ded0a1-ea81-4a5c-8895-bc442f79c3f6
5e764adf-411c-42d2-ad2c-a2ddee58abfa	52b32a74-7a52-450c-b8ad-b06020549a98
5fa385c3-e977-45df-8a26-1a41e1086c24	ded9a617-1b5f-4f55-b36c-b89aaa20c08e
8011c74a-aa69-46b3-af41-f09705dd3010	f9bf9e84-39a2-45b4-b9af-5d6e6256b4ad
4a9c103f-6e69-4b1a-a82d-7fc30dd0c488	31d0f0f6-511a-44ac-b69f-fd4b6f278502

Abbreviations
AUC: Area under the receiver operating characteristic curve; CPP: Cepstral peak prominence; FPR: False-positive rate; FZCF: First zero in correlation function; GQ: Glottal quotient; HNR: Harmonic-to-noise ratio; HY: Hoehn and Yahr; IIS: Windows Internet Information Services; JSON: JavaScript Object Notation; LZ-2: Lempel–Ziv complexity; MFCC: Mel frequency cepstral coefficients; MFSW: MultiFractal spectrum width; PD: Parkinson's disease; RFECV: Recursive feature elimination with cross-validation; RPDE: Recurrence period density entropy; ROC: Receiver operating characteristic curve; SVM: Support vector machine; TPR: True-positive rate; UEX: Universidad de Extremadura.

Acknowledgements
The authors would like to thank Rosa Muñoz for registering the complementary information about the participants as well as providing her neurological advising, and Diego Santiago for recording the speech database. It is also acknowledged the Regional Association for Parkinson's Disease of Extremadura and the patients and healthy people who voluntarily participated in this study.

Authors' contributions
YC and CJP conceived the content of the study. JC, YC, and CJP designed this study. All reviewed literature. JC, MM and CJP conducted experiments and collected the data. JC and MM programmed the app and server. JC, YC, and MM implemented the machine learning approaches. All analyzed and interpreted the results. YC wrote the original draft. All reviewed and edited the writing. All authors read and approved the final manuscript

Funding
This research has been funded by ***Agencia Estatal de Investigación***, Spain (Project MTM2017-86875-C3-2-R), ***Junta de Extremadura***, Spain (Projects GR18108 and GR18055), and the ***European Union*** (European Regional Development Funds). The work of Mario Madruga was funded by the ***Ministerio de Ciencia, Innovación y Universidades)*** through the Ph.D. Grant number FPU18/03274.

Availability of data and materials
The dataset from the in-house voice database (UEX database) is available from the corresponding author on reasonable request. The availability of data contributed by users of the Parkinson mPower mobile application are part of the mPower study developed by Sage Bionetworks and can be accessed through Synapse at https://www.synapse.org/mPower doi: [https://doi.org/10.7303/syn4993293.

Declarations

Ethics approval and consent to participate
This study was approved by the Bioethics Committee of the University of Extremadura under the reference number 802020.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Departamento de Matemáticas, Universidad de Extremadura, Cáceres, Spain. [2]Departamento de Tecnología de los Computadores y las Comunicaciones, Universidad de Extremadura, Cáceres, Spain.

Received: 17 September 2021 Accepted: 4 November 2021
Published online: 21 November 2021

References
1. Almeida JS, Rebouças Filho PP, Carneiro T, Wei W, Damaševičius R, Maskeliūnas R, de Albuquerque VHC. Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques. Pattern Recognit Lett. 2019;125:55–62. https://doi.org/10.1016/j.patrec.2019.04.005.
2. Bang Y-I, Min K, Sohn YH, Cho S-R. Acoustic characteristics of vowel sounds in patients with Parkinson disease. NeuroRehabilitation. 2013;32(3):649–54. https://doi.org/10.3233/NRE-130887.
3. Belalcazar-Bolanos E, Orozco-Arroyave J, Arias-Londono J, Vargas-Bonilla J, Nöth E. Automatic detection of Parkinson's disease using noise measures of speech. In: Symposium of Signals, Images and Artificial Vision-2013: STSIVA-2013, pp. 1–5 (2013). IEEE.
4. Benba A, Jilbab A, Hammouch A. Using human factor cepstral coefficient on multiple types of voice recordings for detecting patients with parkinson's disease. Irbm. 2017;38(6):346–51. https://doi.org/10.1016/j.irbm.2017.10.002.
5. Berrar D. Cross-validation. Encycl Bioinform Comput Biol. 2019;1:542–5. https://doi.org/10.1016/B978-0-12-809633-8.20349-X.
6. Bloem B, Stocchi F. Move for change Part I: A European survey evaluating the impact of the EPDA charter for people with Parkinson's disease. Eur J Neurol. 2012;19(3):402–10. https://doi.org/10.1111/j.1468-1331.2011.03532.x.
7. Bot BM, Suver C, Neto EC, Kellen M, Klein A, Bare C, Doerr M, Pratap A, Wilbanks J, Dorsey ER, Friend SH, Trister AD. The mPower study, Parkinson disease mobile data collected using researchkit. Sci Data. 2016;3:160011. https://doi.org/10.1038/sdata.2016.11.

8. Dashtipour K, Tafreshi A, Lee J, Crawley B. Speech disorders in Parkinson's disease: pathophysiology, medical management and surgical approaches. Neurodegener Dis Manag. 2018;8(5):337–48. https://doi.org/10.2217/nmt-2018-0021.
9. Dhakal P, Damacharla P, Javaid AY, Devabhaktuni V. Detection and identification of background sounds to improvise voice interface in critical environments. In: 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 078–083 (2018). https://doi.org/10.1109/ISSPIT.2018.8642755
10. Dorsey ER, Bloem BR. The Parkinson pandemic—a call to action. JAMA Neurol. 2018;75(1):9–10. https://doi.org/10.1001/jamaneurol.2017.3299.
11. Dotchin C, Walker R. The management of Parkinson's disease in sub-Saharan Africa. Expert Rev Neurother. 2012;12(6):661–6. https://doi.org/10.1586/ern.12.52.
12. Gerratt BR, Kreiman J, Garellek M. Comparing measures of voice quality from sustained phonation and continuous speech. J Speech Lang Hear Res. 2016;59(5):994–1001.
13. Gidaye G, Nirmal J, Ezzine K, Frikha M. Wavelet sub-band features for voice disorder detection and classification. Multimedia Tools Appl. 2020;79(39):28499–523. https://doi.org/10.1007/s11042-020-09424-1.
14. Giuliano M, García-López A, Pérez S, Pérez FD, Spositto O, Bossero J. Selection of voice parameters for Parkinson's disease prediction from collected mobile data. In: 2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA), pp. 1–3 (2019). https://doi.org/10.1109/STSIVA.2019.8730219.
15. Hao J, Ho TK. Machine learning made easy: A review of Scikit-learn package in Python programming language. J Educ Behav Stat. 2019;44(3):348–61. https://doi.org/10.3102/1076998619832248.
16. Henríquez P, Alonso JB, Ferrer MA, Travieso CM, Godino-Llorente JI, Díaz-de-María F. Characterization of healthy and pathological voice through measures based on nonlinear dynamics. IEEE Trans Audio Speech Lang Process. 2009;17(6):1186–95. https://doi.org/10.1109/TASL.2009.2016734.
17. Ihlen EAF. Introduction to multifractal detrended fluctuation analysis in Matlab. Front Physiol. 2012;3:141. https://doi.org/10.3389/fphys.2012.00141.
18. Jain D, Mishra AK, Das SK. Machine learning based automatic prediction of Parkinson's disease using speech features. In: Bansal, P., Tushir, M., Balas, V.E., Srivastava, R. (eds.) Proceedings of International Conference on Artificial Intelligence and Applications, pp. 351–362. Springer, (2021)
19. Jeancolas L, Benali H, Benkelfat B-E, Mangone G, Corvol J-C, Vidailhet M, Lehericy S, Petrovska-Delacrétaz D. Automatic detection of early stages of Parkinson's disease through acoustic voice analysis with mel-frequency cepstral coefficients. In: 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), pp. 1–6 (2017). https://doi.org/10.1109/ATSIP.2017.8075567.
20. Kacha A, Mertens C, Grenez F, Skodda S, Schoentgen J. On the harmonic-to-noise ratio as an acoustic cue of vocal timbre of Parkinson speakers. Biomed Signal Process Control. 2017;37:32–8. https://doi.org/10.1016/j.bspc.2016.09.004.
21. Karabayir I, Goldman SM, Pappu S, Akbilgic O. Gradient boosting for Parkinson's disease diagnosis from voice recordings. BMC Med Inf Decis Mak. 2020;20(1):1–7. https://doi.org/10.1186/s12911-020-01250-7.
22. Lalo E, Riff J, Parry R, Jabloun M, Roussel J, Chen C-C, Welter M-L, Buttelli O. Design of technology and technology of design. activity analysis as a resource for a personalised approach for patients with parkinson disease. IRBM. 2016;37(2):90–7. https://doi.org/10.1016/j.irbm.2016.02.010.
23. Linares-Del Rey M, Vela-Desojo L, Cano-de La Cuerda R. Mobile phone applications in Parkinson's disease: A systematic review. Neurología. 2019;34(1), 38–54. https://doi.org/10.1016/j.nrleng.2018.12.002.
24. Little MA, McSharry PE, Hunter EJ, Spielman J, Ramig LO. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Trans Biomed Eng. 2009;56(4):1015.
25. Mekyska J, Galaz Z, Mzourek Z, Smekal Z, Rektorova I, Eliasova I, Kostalova M, Mrackova M, Berankova D, Faundez-Zanuy M et al. Assessing progress of Parkinson's disease using acoustic analysis of phonation. In: 2015 4th International Work Conference on Bioinspired Intelligence (IWOBI), pp. 111–118 (2015). IEEE.
26. Miller DB, O'Callaghan JP. Biomarkers of Parkinson's disease: present and future. Metabolism. 2015;64(3):40–6. https://doi.org/10.1016/j.metabol.2014.10.030.
27. Misra P, Singh A. Improving the classification accuracy using recursive feature elimination with cross-validation. Int J Emerg Technol. 2020;11:659–65.
28. Montaña D, Campos-Roca Y, Pérez CJ. A diadochokinesis-based expert system considering articulatory features of plosive consonants for early detection of Parkinson's disease. Comput Methods Program Biomed. 2018;154:89–97. https://doi.org/10.1016/j.cmpb.2017.11.010.
29. Moro-Velazquez L, Gomez-Garcia JA, Arias-Londoño JD, Dehak N, Godino-Llorente JI. 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. 2021;66: 102418. https://doi.org/10.1016/j.bspc.2021.102418.
30. Naranjo L, Perez CJ, Campos-Roca Y, Martin J. Addressing voice recording replications for Parkinson's disease detection. Expert Syst Appl. 2016;46:286–92. https://doi.org/10.1016/j.eswa.2015.10.034.
31. Nguyen VN, Holone H. N-best list re-ranking using syntactic score: A solution for improving speech recognition accuracy in air traffic control. In: 2016 16th International Conference on Control, Automation and Systems (ICCAS), pp. 1309–1314 (2016). https://doi.org/10.1109/ICCAS.2016.7832482.
32. Nikam SS, Dalvi R. Machine learning algorithm based model for classification of fake news on twitter. In: 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp. 1–4 (2020). https://doi.org/10.1109/I-SMAC49090.2020.9243385.
33. Nilanon T, Yao J, Hao J, Purushotham S, Liu Y. Normal/abnormal heart sound recordings classification using convolutional neural network. In: 2016 Computing in Cardiology Conference (CinC), pp. 585–588 (2016). IEEE.
34. Novotnỳ M, Dušek P, Daly I, Ružička E, Rusz J. Glottal source analysis of voice deficits in newly diagnosed drug-naïve patients with Parkinson's disease: correlation between acoustic speech characteristics and non-speech motor performance. Biomed Signal Process Control. 2020;57: 101818. https://doi.org/10.1016/j.bspc.2019.101818.

35. Orozco-Arroyave J, Hönig F, Arias-Londoño J, Vargas-Bonilla J, Daqrouq K, Skodda S, Rusz J, Nöth E. Automatic detection of Parkinson's disease in running speech spoken in three different languages. J Acoust Soc Am. 2016;139(1):481–500. https://doi.org/10.1121/1.4939739.
36. Orozco-Arroyave JR, Arias-Londono JD, Vargas-Bonilla JF, Nöth E. Analysis of speech from people with Parkinson's disease through nonlinear dynamics. In: International Conference on Nonlinear Speech Processing, 2013;pp. 112–119. https://doi.org/10.1007/978-3-642-38847-7_15 Springer.
37. Orozco-Arroyave JR, Hönig F, Arias-Londoño JD, Vargas-Bonilla JF, Nöth E. Spectral and cepstral analyses for Parkinson's disease detection in spanish vowels and words. Expert Syst. 2015;32(6):688–97. https://doi.org/10.1111/exsy.12106.
38. Pahuja G, Nagabhushan T. A comparative study of existing machine learning approaches for parkinson's disease detection. IETE J Res. 2021;67(1):4–14. https://doi.org/10.1080/03772063.2018.1531730.
39. Petrizzo D, Popolo PS. Smartphone use in clinical voice recording and acoustic analysis: a literature review. J Voice. 2020. https://doi.org/10.1016/j.jvoice.2019.10.006.
40. Romann AJ, Beber BC, Cielo CA, Rieder CRdM. Acoustic voice modifications in individuals with parkinson disease submitted to deep brain stimulation. Int Arch Otorhinolaryngol. 2019;23:203–8.
41. Rusz J, Tykalová T, Krupička R, Zárubová K, Novotnỳ M, Jech R, Szabó Z, Ružička E. Comparative analysis of speech impairment and upper limb motor dysfunction in Parkinson's disease. J Neural Transm. 2017;124(4):463–70. https://doi.org/10.1007/s00702-016-1662-y.
42. Rusz J, Hlavnička J, Tykalová T, Novotnỳ M, Dušek P, Šonka K, Ružička E. Smartphone allows capture of speech abnormalities associated with high risk of developing Parkinson's disease. IEEE Trans Neural Syst Rehab Eng. 2018;26(8):1495–507. https://doi.org/10.1109/TNSRE.2018.2851787.
43. Sakar BE, Isenkul ME, Sakar CO, Sertbas A, Gurgen F, Delil S, Apaydin H, Kursun O. Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. IEEE J Biomedical Health Inform. 2013;17(4):828–34. https://doi.org/10.1109/JBHI.2013.2245674.
44. Sakar CO, Serbes G, Gunduz A, Tunc HC, Nizam H, Sakar BE, Tutuncu M, Aydin T, Isenkul ME, Apaydin H. A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform. Appl Soft Comput. 2019;74:255–63. https://doi.org/10.1016/j.asoc.2018.10.022.
45. Sakar CO, Serbes G, Gunduz A, Tunc HC, Nizam H, Sakar BE, Tutuncu M, Aydin T, Isenkul ME, Apaydin H. A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform. Appl Soft Comput. 2019;74:255–63.
46. Solana-Lavalle G, Galán-Hernández J-C, Rosas-Romero R. Automatic Parkinson disease detection at early stages as a pre-diagnosis tool by using classifiers and a small set of vocal features. Biocybern Biomed Eng. 2020;40(1):505–16. https://doi.org/10.1016/j.bbe.2020.01.003.
47. Syarif I, Prugel-Bennett A, Wills G. SVM parameter optimization using grid search and genetic algorithm to improve classification performance. TELKOMNIKA. 2016;14:1502. https://doi.org/10.12928/telkomnika.v14i4.3956.
48. Tougui I, Jilbab A, El Mhamdi J. Analysis of smartphone recordings in time, frequency, and cepstral domains to classify Parkinson's disease. Healthc Inform Res. 2020;26(4):274. https://doi.org/10.4258/hir.2020.26.4.274.
49. Tracy JM, Özkanca Y, Atkins DC, Ghomi RH. Investigating voice as a biomarker: deep phenotyping methods for early detection of Parkinson's disease. J Biomed Inform. 2020;104: 103362. https://doi.org/10.1016/j.jbi.2019.103362.
50. Tsanas A, Little MA, McSharry PE, Spielman J, Ramig LO. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Trans Biomed Eng. 2012;59(5):1264–71. https://doi.org/10.1109/TBME.2012.2183367.
51. Tsanas A. Acoustic analysis toolkit for biomedical speech signal processing: concepts and algorithms. Models Anal Vocal Emiss Biomed Appl. 2013;2:37–40.
52. Tsanas A, Gómez-Vilda P. Novel robust decision support tool assisting early diagnosis of pathological voices using acoustic analysis of sustained vowels. In: Multidisciplinary Conf. Users of Voice, Speech Sing.(JVHC 13), 2013;pp. 3–12.
53. Vásquez-Correa J, Orozco-Arroyave J, Bocklet T, Nöth E. Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease. J Commun Disord. 2018;76:21–36. https://doi.org/10.1016/j.jcomdis.2018.08.002.
54. Vásquez-Correa JC, Rios-Urrego CD, Arias-Vergara T, Schuster M, Rusz J, Nöth E, Orozco-Arroyave JR. Transfer learning helps to improve the accuracy to classify patients with different speech disorders in different languages. Pattern Recognition Letters. 2021. https://doi.org/10.1016/j.patrec.2021.04.011.
55. Wroge TJ, Özkanca Y, Demiroglu C, Si D, Atkins DC, Ghomi RH. Parkinson's disease diagnosis using machine learning and voice. In: 2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), pp. 1–7 (2018). https://doi.org/10.1109/SPMB.2018.8615607.
56. Zhang H, Wang A, Li D, X, W. Deepvoice: A voiceprint-based mobile health framework for Parkinson's disease identification. In: 2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI), 2018;pp. 214–217. https://doi.org/10.1109/BHI.2018.8333407
57. Zhang Y. Can a smartphone diagnose Parkinson disease? A deep neural network method and telediagnosis system implementation. Parkinson's Dis 2017; https://doi.org/10.1155/2017/6209703.

Publisher's Note

Chapter 4

Replication-based regularization approaches to diagnose Reinke's edema by using voice recordings

Title:

Replication-based regularization approaches to diagnose Reinke's edema by using voice recordings

Authors and affiliation:

Lizbeth Naranjo[a], Carlos J. Pérez[b], Yolanda Campos-Roca[c], Mario Madruga[b]

[b]Departamento de Matemáticas, Facultad de Ciencias, Universidad Nacional Autónoma de México
[b]Universidad de Extremadura, Departamento de Matemáticas, Spain
[c]Universidad de Extremadura, Departamento de Tecnología de los Computadores y las Comunicaciones, Spain

Journal:

Artificial Intelligence in Medicine

DOI:

10.1016/j.artmed.2021.102162

Abstract: Reinke's edema is one of the most prevalent laryngeal pathologies. Its detection can be addressed by using computer-aided diagnosis systems based on features extracted from speech recordings. When extracting acoustic features from different voice recordings of a particular subject at a concrete moment, imperfections in technology and the very biological variability result in values that are close, but they are not identical. This suggests that the within-subject variability must be properly addressed in the statistical methodology. Regularization-based regression approaches can be used to reduce the classification errors by favoring the best predictors and penalizing the worst ones. Three replication-based regularization approaches for variable selection and classification have been specifically designed and implemented to take into account the underlying within-subject variability. In order to illustrate the applicability of these approaches, an experiment has been specifically conducted to discriminate Reinke's edema patients (30 subjects) from healthy people (30 subjects) in a hospital environment. The features have been extracted from four phonations of the sustained vowel /a/ recorded for each subject, leading to a database that has fed the proposed machine learning approaches. The proposed replication-based approaches have been proved to be reliable in terms of selected features and predictive ability, leading to a stable accuracy rate of 0.89 under a cross-validation framework. Also, a comparison with traditional independence-based regularization methods reports a great variability of the latter in terms of selected features and accuracy metrics. Therefore, the proposed approaches contribute to fill a gap in the scientific literature on statistical approaches considering within-subject variability and can be used to build a robust expert system.

Keywords: Acoustic features, Classification, Reinke's edema, Regularization, Replicated measurements, Variable selection.

Artificial Intelligence In Medicine 120 (2021) 102162

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Replication-based regularization approaches to diagnose Reinke's edema by using voice recordings

Lizbeth Naranjo [a], Carlos J. Pérez [b,*], Yolanda Campos-Roca [c], Mario Madruga [b]

[a] Departamento de Matemáticas, Facultad de Ciencias, Universidad Nacional Autónoma de México, 04510 Ciudad de México, Mexico
[b] Departamento de Matemáticas, Facultad de Veterinaria, Universidad de Extremadura, 10003 Cáceres, Spain
[c] Departamento de Tecnologías de los Computadores y de las Comunicaciones, Escuela Politécnica, Universidad de Extremadura, 10003 Cáceres, Spain

ARTICLE INFO

Keywords:
Acoustic features
Classification
Reinke's edema
Regularization
Replicated measurements
Variable selection

ABSTRACT

Reinke's edema is one of the most prevalent laryngeal pathologies. Its detection can be addressed by using computer-aided diagnosis systems based on features extracted from speech recordings. When extracting acoustic features from different voice recordings of a particular subject at a concrete moment, imperfections in technology and the very biological variability result in values that are close, but they are not identical. This suggests that the within-subject variability must be properly addressed in the statistical methodology. Regularization-based regression approaches can be used to reduce the classification errors by favoring the best predictors and penalizing the worst ones. Three replication-based regularization approaches for variable selection and classification have been specifically designed and implemented to take into account the underlying within-subject variability. In order to illustrate the applicability of these approaches, an experiment has been specifically conducted to discriminate Reinke's edema patients (30 subjects) from healthy people (30 subjects) in a hospital environment. The features have been extracted from four phonations of the sustained vowel /a/ recorded for each subject, leading to a database that has fed the proposed machine learning approaches. The proposed replication-based approaches have been proved to be reliable in terms of selected features and predictive ability, leading to a stable accuracy rate of 0.89 under a cross-validation framework. Also, a comparison with traditional independence-based regularization methods reports a great variability of the latter in terms of selected features and accuracy metrics. Therefore, the proposed approaches contribute to fill a gap in the scientific literature on statistical approaches considering within-subject variability and can be used to build a robust expert system.

1. Introduction

Voice is the main communication tool that human beings have. Misuse or overuse of the vocal folds can damage the vocal function. Voice disorders may affect anyone, but they are especially relevant for voice professionals such as teachers, singers, actors, anchors, coaches, lawyers… Voice professionals are prone to suffer from organic voice disorders and, because of that, they need to avoid potential risks and, eventually, ask for medical care [41].

Reinke's edema is one of the most prevalent laryngeal pathologies [33]. It is the result of the gelatinous fluid accumulation in the Reinke's space, mainly due to vocal abuse and/or heavy tobacco use. It mainly affects women, causing progressive hoarse voice with a lower pitch, less vocal power and a tendency to fatigue in more intense cases [5]. Direct inspection of the larynx through laryngoscopy and videostroboscopy (specialized invasive equipment) and/or subjective listening tests to evaluate voice quality are two common diagnostic tools used by otolaryngologists [45].

In the last years, acoustic features extracted from voice recordings have been considered as a potential biomarker (non-invasive, fast, objective, and low cost) to assist in the diagnosis and tracking of voice-related diseases. Computer-Aided Diagnosis (CAD) systems have been built with this purpose, consisting of an acoustic feature extraction step followed by the use of machine learning algorithms. A perspective on automatic speech signal analysis for clinical diagnosis and assessment of speech disorders is provided by Baghai-Ravary and Beet [1] and Gómez-García et al. [13]. These systems have been developed for several diseases affecting the voice such as, e.g., vocal fold nodules, vocal fold polyps, Reinke's edema, or even neurodegenerative disorders such as Parkinson's disease.

* Corresponding author at: Avda. de las Ciencias, s/n, 10003, Cáceres, Spain.
E-mail address: carper@unex.es (C.J. Pérez).

https://doi.org/10.1016/j.artmed.2021.102162
Received 15 February 2021; Received in revised form 21 August 2021; Accepted 31 August 2021
Available online 8 September 2021

Diagnosis of Reinke's edema can be addressed by using CAD systems based on features extracted from speech recordings. Some authors have considered a mix of different pathologies, including Reinke's edema, to build a unique pathological class to discriminate diseased subjects from healthy ones [6,23,31,58]. Verde et al. [56] focused on Reinke's edema and their results were based on a personalized fundamental frequency estimation and no other acoustic feature was considered. Features based on nonlinear dynamics analysis have not been thoroughly used for Reinke's edema diagnosis in the scientific literature. Tavares et al. [51] combined entropy measures and cepstral analysis to discriminate healthy subjects from people suffering from Reinke's edema. Based on energy, zero-crossing rate and signal entropy, Silva Fonseca et al. [48] presented a speech disorder classification method that handles coexisting pathologies (Reinke's edema and laryngitis) that share the main phonic symptom. Phonation of sustained vowels was used in the previous works because they constitute easy to produce tasks, involving vocal fold vibration [39].

MEEI database, commercialized by Kay Elemetrics, is one of the most used voice database for automatic diagnosis research and covers several voice pathologies, including Reinke's edema [34]. However, it suffers from some disadvantages: the recorded phonations have been performed with high quality equipment in an acoustically controlled environment, and normal and pathological voices were recorded in different locations. Besides, the voice recordings have been selected by experts, which allowed for obtaining the best examples of each disease [44]. This has provided high accuracies when applying machine learning methods, but the results are not transferable to realistic situations where the phonations are recorded in medical centers or occupational health and safety services.

Voice databases used for organic disease diagnosis are generally based on one single utterance per subject, i.e., acoustic features extracted from only one voice recording per subject. However, there exists variability between two or more voice recordings from the same subject at a particular time, so using only one utterance per subject may provide different results depending on the voice recording that has been selected. The imperfections in technology and the very biological variability result in values that are similar (but not identical) for recordings from a particular subject, rather than for recordings from different individuals. For Parkinson's disease diagnosis, many authors considered several replicated voice recordings for each subject, so a collection of related features based on consecutive voice recordings for each subject are used (see, e.g., Little et al. [29]). Although the existing variability among the extracted features from the several voice recordings of each subject has been recognized and the experimental design is based on the within-subject dependence of the recordings of each individual, traditional machine learning techniques based on independence have been usually applied to all the utterances as if they were independent [8,29,54]. This means that the considered experimental unit is the utterance, and not the subject, so a voting system is used to decide if a subject is classified as healthy or having the disease by taking into account the larger number of utterances classified as healthy or diseased for each subject. This leads to an artificial increase of the sample size, a diffuse criterion to make decisions since one subject can have utterances classified as healthy and diseased, and the application of independence-based methods to dependent data.

The replicated measurements must be treated with specifically designed methods that address the existing within-subject variability. Pérez et al. [43] developed a logistic regression-based classification approach that takes into account the underlying within-subject dependence based on 6/7 utterances per subject. Later, Naranjo et al. [37] addressed this problem with a probit regression based on 3 utterances per subject, whereas Naranjo et al. [38] proposed a variable selection and classification approach for the same data. All these three approaches have been developed in the context of Parkinson's disease diagnosis with features extracted from voice recordings.

In this paper, replication-based Bayesian regularization approaches for Reinke's edema diagnosis using acoustic features extracted from speech recordings have been developed and implemented. Variable selection and classification approaches have been widely addressed by Bayesian regularization regression with independent instances (see, e.g., van Erp et al. [55]), which aim to shrink small effects to zero while maintaining true large effects. However, there is a lack of regularization methods able to address within-subject variability. To the best of the authors' knowledge, up to now, it has never been demonstrated that having into account the within-subject variability provides more stable results than the approaches based on independent instances at the same time that relevant features are selected and accuracy metrics keep at good values. This study contributes to fill a gap in the scientific literature on statistical approaches considering replicated data and they can be used to build robust CAD systems. The main contributions of this article are:

- Designing and implementing three Bayesian regularization approaches based on replicated measurements.
- Using Markov Chain Monte Carlo (MCMC) methods to solve the increasingly complex models.
- Conducting an experiment to discriminate subjects suffering from Reinke's edema (30 subjects) from healthy people (30 subjects) in a hospital environment.
- Extracting a variety of relevant features based on perturbation, cepstral analysis, noise, nonlinear dynamics, and entropies.
- Proposing and integrating a 95% Bayesian credible interval-based technique to determine the most relevant acoustic features.
- Reporting a robust performance in terms of feature selection and predictive capability, leading to an accuracy of 0.89 by using cross-validation and 0.93 without it.
- Reporting the outperformance of the replication-based approach based on Ridge regression with respect to the traditional regularization methods based on independent instances, which provide a great variability in terms of selected features and accuracy metrics.

The rest of this paper is structured as follows. Section 2 shows the necessary information to collect the dataset, i.e., participants, equipment, speech recordings, and feature extraction procedures. In Section 3, the general Bayesian approach is presented, including the binary response model, the way the replications are addressed in the model, the prior distributions for the different approaches, the Bayesian analysis, and the variable selection method. Section 4 shows the experimental settings and results. In Section 5, a discussion is presented, and the conclusions can be found in Section 6.

2. Data collection

This section provides details on the different aspects related to the generation of the acoustic feature database, i.e., the participants, protocol, recording equipment, vocal task, and feature extraction process.

2.1. Participants

A total of 60 people participated in the study. Half of them were diagnosed as suffering from Reinke's edema and the other half were healthy control subjects. The general eligibility criteria for participation were to be volunteers, native Spanish speakers, aged from 18 to 65, and to properly perform the phonation task in the research protocol.

The group of people suffering from Reinke's edema comprised 27 women and 3 men, with mean (standard deviation) age of 47.9 (11.8) years. They were recruited among the volunteers who attended the voice disorder program at the San Pedro de Alcántara Hospital. Note that there is a gender imbalance due to the fact that women are more affected by organic vocal-fold pathologies than men (see, e.g., Hunter et al. [21]). The gender rate in this study is approximately the same as in people attending the voice disorder program at the moment of the recruitment.

On the other hand, the healthy control group was selected among people with good vocal health status, who had never suffered from any voice pathology or used their voices in a professional way. It comprised 26 women and 4 men, with mean (standard deviation) age of 40.8 (11.2) years.

All the subjects were informed and provided their consent by signing an informed consent letter.

2.2. Protocol and equipment

The participants were asked to fill out a questionnaire for assessment of part of the general and specific eligibility criteria. They provided information such as sex, age, smoking habits, use of medication, and previous surgical interventions. They also underwent a medical examination consisting of a laryngological evaluation by videostroboscopy performed by an otorhinolaryngologist. For the subjects suffering from Reinke's edema, it was confirmed that Reinke's edema was the only existing voice pathology.

A portable computer with an external sound card (TASCAM US322) and a headband microphone (AKG 520) featuring a cardioid pattern was used to record the phonations. The digital recording was performed using Audacity software (release 2.0.5). The sampling frequency was 44.1 kHz and the resolution 16 bits/sample.

This research protocol was approved by the bioethics committees of the San Pedro de Alcántara Hospital and the University of Extremadura.

2.3. Speech recordings

The voice recordings were performed in an ordinary diagnostic room at San Pedro de Alcántara Hospital. The room was not sound-proof, but a certain isolation from the aisles and waiting halls was obtained by regular walls and closed doors. No specific measures for acoustic isolation were implemented.

The participants were asked to perform a sustained voicing of the /a/ vowel, at a comfortable pitch and loudness, as constantly as possible. This phonation was kept up as long as they could after a deep breath. A segment of one second was considered for feature extraction. This procedure was repeated four consecutive times per individual to address the within-subject variability after feature extraction.

2.4. Feature extraction

Different types of acoustic features were considered. The idea was to measure different aspects of speech degradation caused by the voice disorder.

Two conventional perturbation measures (jitter and shimmer) were extracted based on the high values observed in patients with Reinke's edema in previous studies [47]. Fundamental frequency and amplitude perturbations also produce an impact on the cepstral peak prominence (CPP). This measure, originally proposed by Hillenbrand et al. [19], is considered more robust than time-domain techniques, since it does not require pitch tracking and can be reliably extracted even from highly aperiodic signals. For this reason, CPP has been included in the list of features.

Voice roughness is a characteristic symptom of Reinke's edema because the swelling alters the elasticity of the vocal folds [7]. Two noise measures have been included in the feature set to assess roughness: glottal-to-noise excitation (GNE) ratio and the harmonic-to-noise ratio (HNR). These noise measures have been considered suitable for the detection of voice pathologies [12].

According to previous scientific studies, vocal fold pathologies lead to changes in vocal tract configuration during phonation. Lee et al. [27] pointed out that the reason is related to physiological or psychological compensations. Mel-frequency cepstral coefficients (MFCCs) have been widely used to characterize the vocal tract configuration in different application areas of speech classification, also for the detection of vocal-fold disorders [10]. A total of 13 MFCCs were calculated and included in the feature set.

Furthermore, it has been emphasized that nonlinear behaviors play a relevant role in the voice production process, especially in the case of disordered voices [12,32,53]. Therefore, the classical source-filter theory is not sufficient to describe all important aspects of speech that can be useful to detect pathologies. Orozco-Arroyave et al. [40] state different reasons which lead to a nonlinear speech behavior: nonlinear pressure-flow in the glottis, nonlinear stress-strain curves of vocal fold tissues, and nonlinearities in vocal fold collisions. These authors also consider the compensatory movements mentioned in the previous paragraph as nonlinear effects. Based on this nonlinear assumption, some authors have proposed acoustic features taken from the field of time-series analysis to predict diseases affecting voice [25,30,40]. The following ones have been used: Hurst exponent (HURST), correlation dimension (D2), permutation (PERMUTATION) and shannon entropy (SHANNON), pitch period entropy (PPE), and recurrence period density entropy (RPDE). Finally, the zero-crossing rate (ZCR) was also included. This adds up to a total of 25 acoustic features extracted from each voice sample. The extraction methods were coded in Python.

Gender is also important in this topic. Yamauchi et al. [59] used glottal area waveform analysis based on high-speed digital imaging to emphasize the relevant role of gender when deciding whether a vocal fold pattern is normal or pathological. Previous studies [26,52] had already identified gender differences in vocal fold configuration during phonation: in glottal flow, glottal area or contact area waveforms. These anatomical and physiological differences have motivated the inclusion of the gender label as an additional feature, giving a total number of 26 features.

The feature extraction procedure provides a dataset with 240 rows (60 subjects $\times$ 4 utterances) and 27 columns (number of features plus health status).

3. Methodology

In the following subsections the methodology is described. Firstly, a hierarchical model to deal with binary responses and replicated covariables is formulated. This provides a general framework for replication-based classifiers. Then, three Bayesian regularization methods are considered through their respective prior distributions. Next, the posterior distribution is estimated and the posterior predictive probabilities are calculated. Finally, a variable selection method based on Bayesian credible intervals is proposed to determine the most relevant features.

3.1. Binary response model

In order to define the hierarchical model, the first level corresponds to the binary response variable. Let $Y_1, \ldots, Y_n$ be the n independent binary random variables:

$$Y_i \sim \text{Bernoulli}(\theta_i)$$

The probabilities $\theta_i = \text{P}(Y_i = 1)$ are related to two sets of covariates, $\boldsymbol{w}_i$ and $\boldsymbol{z}_i$ by:

$$\Psi^{-1}(\theta_i) = w_i'\beta + z_i'\gamma,$$

where $w_i = (w_{i1}, \ldots, w_{iK})'$ and $z_i = (z_{i1}, \ldots, z_{iH})'$ are covariate vectors of dimension K and H, respectively. The parameters β and γ are vectors of unknown parameters, of dimensions K and H, respectively. $\Psi^{-1}(\cdot)$ is the inverse of the cumulative distribution function (cdf) of the normal distribution.

3.2. Introducing replications

Assume that the covariates $\boldsymbol{z}_i$ are exactly known (e.g. sex), but the covariates $\boldsymbol{w}_i$ are not (acoustic features), instead they have been

measured with J replicates. Let $x_{ij} = (x_{i1j}, \ldots, x_{iKj})'$ be the jth replication of the unknown covariate vector $w_i = (w_{i1}, \ldots, w_{iK})'$, $j = 1, \ldots, J$, and assume that they have a linear relationship specified as an additive measurement error model (see, e.g. Buonaccorsi [3]), i.e.:

$$x_{ikj} = w_{ik} + \varepsilon_{ikj},$$
$$\varepsilon_{ikj} \sim \text{Normal}\left(0, \delta_k^2\right),$$

where the errors ε_{ik} are independent of w_{ik}, and x_{ikj} can be considered as surrogates of w_{ik}.

The rationale under this formulation is that the observed replicated features can be considered as measurement with errors of the underlying real acoustic feature, which is unknown for each individual. This latent variable-based structure is the key idea to address the within-subject variability.

3.3. Integrating regularization

Regularization methods simultaneously perform estimation and variable selection. They favor the best predictors and penalize the worst ones through parameter regularization. A wide variety of regularization methods have been developed (see e.g., Hastie et al. [15] and Hastie et al. [16]). The most usual regularization methods are Least Absolute Shrinkage and Selection Operator (LASSO), Ridge, and Elastic Net. They have been widely used for independent instances, but now they are considered for data with dependent nature in a framework that addresses the within-subject variability of replicated measurements, and therefore for a different type of statistical design.

In typical Bayesian regression, the prior distribution for the regression parameters is normal. When regularization methods are considered, different prior distributions are used. LASSO is one of the most commonly used penalized regression methods (see Park and Casella [42]). The prior distribution for the regression parameters β_k is based on the proposal of Genkin et al. [9], i.e., a Laplace distribution is considered, i.e.:

$$\beta_k \sim \text{Laplace}\left(0, \lambda_1^{-1}\right),$$

with mean 0 and variance $2/{\lambda_1}^2$, for k=1,…, K.

The Laplace pdf is proportional to:

$$\text{p}(\beta_k) \propto exp\{-\lambda_1|\beta_k|\},$$

and it can be represented as a scale mixture of normal distributions with independent exponentially distributed variances, i.e.:

$$\text{p}(\beta_k) = \int_0^\infty \text{p}(\beta_k|\tau_k)\text{p}(\tau_k)\text{d}\tau_k,$$

where

$$\beta_k|\tau_k^2 \sim \text{Normal}\left(0, \tau_k^2\right),$$
$$\tau_k^2 \sim \text{Exp}\left(\lambda_1^2/2\right),$$

being the exponential distribution parameterized so the mean is $2/{\lambda_1}^2$.

Ridge regression is another regularization model (Hoerl and Kennard [20]). In this case, the prior distribution for the regression parameters β_k is:

$$\beta_k \sim \text{Normal}\left(0, \lambda_2^{-1}\right),$$

i.e., its pdf is proportional to:

$$\text{p}(\beta_k) \propto exp\left\{\frac{\lambda_2}{2}\beta_k^2\right\}$$

The prior distribution restricts the regression parameters (with high probability) to a sphere of radius determined by λ_2.

Finally, the Elastic Net method combines LASSO and Ridge regularization methods [60]. The prior distribution for the regression parameters β_k is:

$$\text{p}(\beta_k) \propto exp\left\{-\lambda_1|\beta_k| - \frac{\lambda_2}{2}\beta_k^2\right\}$$

By using latent variables, it is possible to obtain a scale mixture of normal distributions representation:

$$\beta_k|\sigma_{\beta_k}^2 \sim \text{Normal}\left(0, \sigma_{\beta_k}^2\right),$$
$$\sigma_{\beta_k}^2 = \left(\tau_k^{-2} + \lambda_2\right)^{-1},$$
$$\tau_k^2 \sim \text{Exp}\left(\lambda_1^2/2\right)$$

3.4. Exploring the posterior distribution

Firstly, the prior distributions are presented. The prior distributions for the regression parameters related to the acoustic features β_k, $k = 1, \ldots, K$, have been defined in Section 3.3. Besides, normal distributions are assumed for the regression parameters related to the exactly known covariates, i.e. $\gamma_h \sim \text{Normal}(c_h, C_h)$, for $h = 1, \ldots, H$, where $c = (c_1, \ldots, c_H)$ and $C = (C_1, \ldots, C_H)$ are fixed values. Inverse Gamma distributions are considered for variances ${\delta_k}^2$, i.e., ${\delta_k}^2 \sim \text{InvGamma}(s_k, r_k)$, where S_k and r_k are the shape and rate parameters, respectively.

Normal distributions are considered for the latent variables, i.e., $w_{ik} \sim \text{Normal}(\mu_k, {\tau_k}^2)$. For the hyperparameters of the latent variables, the prior distributions are defined as $\mu_k \sim \text{Normal}(m_k, {v_k}^2)$ and ${\tau_k}^2 \sim \text{InvGamma}(u_k, t_k)$. The hyperparameters of the regularization methods, ${\lambda_1}^2$ and λ_2, can be fixed values, but they may have hyperprior distributions, e.g., ${\lambda_1}^2 \sim \text{Gamma}(a_1, d_1)$ and $\lambda_2 \sim \text{Gamma}(a_2, d_2)$.

The binary hierarchical model with replications defined in Sections 3.1 and 3.2 results in the likelihood function, considering the observed and the latent variables, given by:

$$\begin{aligned}&\mathscr{L}\left(\beta, \gamma, \delta^2, \mu, \tau^2 | y, x, z, w\right)\\ &= \text{p}(\mathbf{y}|z, w, \beta, \gamma)\text{p}\left(x|w, \delta^2\right)\text{p}\left(w|\mu, \tau^2\right)\\ &= \prod_{i=1}^{n}\left\{\text{p}(y_i|z_i, w_i, \beta, \gamma)\left[\prod_{k=1}^{K}\left\{\prod_{j=1}^{J}\text{p}\left(x_{ikj}|w_{ik}, \delta_k^2\right)\right\}\text{p}\left(w_{ik}|\mu_k, \tau_k^2\right)\right]\right\}\end{aligned} \quad (1)$$

The joint posterior distribution is obtained by using the likelihood function (1) and the prior distributions previously defined, and it is given by:

$$\begin{aligned}&\text{p}\left(\beta, \gamma, \delta^2, \mu, \tau^2 | y, x, z, w\right)\\ &\propto \mathscr{L}\left(\beta, \gamma, \delta^2, \mu, \tau^2 | y, x, z, w\right)\text{p}(\beta)\text{p}(\gamma)\text{p}\left(\delta^2\right)\text{p}(\mu)\text{p}\left(\tau^2\right)\text{p}(\lambda)\end{aligned} \quad (2)$$

A Markov Chain Monte Carlo (MCMC) algorithm has been implemented in JAGS[1] through the R platform[2] to estimate the posterior distribution. The source code and instructions that allow to run the approach for a simulation-based dataset can be found in the GitHub repository through the link https://github.com/lizbethna/ClassificaReplicaRegulariza.git.

Other Monte Carlo approaches could be applied. For instance, particle filtering could be considered [11]. It deals with targets that are influenced by the proximity and/or behavior of other targets. Also, Hamiltonian Monte Carlo methods can be used. They utilize techniques from differential geometry to generate transitions spanning the full marginal variance [2] or the No-U-Turn sampler, which is an adaptive form of Hamiltonian Monte Carlo sampling [4].

3.5. Determining the most relevant features

After the chain has converged, a random sample for each parameter from the posterior distribution is obtained. Based on the estimated

[1] http://mcmc-jags.sourceforge.net/http://mcmc-jags.sourceforge.net/
[2] https://cran.r-project.org/https://cran.r-project.org/

posterior densities for the regression parameters, a variable selection method based on Bayesian credible intervals is proposed here. For the estimated posterior density of each parameter, this method considers a 95% Bayesian credible interval, being the lower interval limit the 2.5% percentile and the upper one the 97.5% percentile (see, e.g., Hespanhol et al. [18]). The features related to the regression parameters that do not contain 0 in the Bayesian credible interval are selected as relevant features, given the important contribution for predicting the response. Then, the approach is applied to these features to provide accuracy rate, sensitivity, specificity, and AUC-ROC (Area Under the Curve Receiver Operating Characteristic).

The details about the concrete practical implementation considering cross-validation frameworks for variable selection and accuracy metrics are provided in the experimental setting subsection of the results section.

4. Results

4.1. Experimental settings

The replication-based Bayesian regularization approaches in Section 3 are applied to the dataset described in Section 2. The response variable Y takes values Y=0 for healthy subjects and Y=1 for people suffering from Reinke's edema, whereas the 25 acoustic variables have been individually normalized to have mean 0 and standard deviation 1, and the variable sex Z takes values $Z = 0$ for men and $Z = 1$ for women.

The MCMC sampling is applied using the following hyperparameters for the prior distributions. For the regression parameters of the covariates exactly known $\gamma_h \sim$ Normal(0,0.01), for $h = 1, \ldots, H$. For the latent variables in the replications, $w_{ik} \sim$ Normal($\mu_k, {\tau_k}^2$), where $\mu_k \sim$ Normal $(0, 1)$, ${\tau_k}^2 \sim$ InverseGamma$(1, 1)$, and ${\delta_k}^2 \sim$ InverseGamma(0.01,0.01), for $k = 1, \ldots, K$. For the parameters in the regularization methods, ${\lambda_1}^2 \sim$ Gamma$(1, 1)$ and $\lambda_2 \sim$ Gamma$(1, 1)$.

A total of 30,000 iterations with a burn-in of 10,000 and a thinning period of 10 generated values are used, providing a sample of length 2000. With these specifications, the chains generated by using the MCMC sampling algorithm seem to have converged. Bayesian Output Analysis (BOA) package was used to perform the convergence analysis [49]. The previous specifications are enough to provide evidence of convergence for all parameters in the three regularization approaches.

Posterior predictive probabilities are obtained for the accuracy metrics. The used metrics are accuracy rate ((TP + TN)/n), sensitivity (TP/(TP + FN)), specificity (TN/(TN + FP)), where TP = True Positive; TN = True Negative; FP = False Positive; FN = False Negative. AUC-ROC is also considered.

A stratified cross-validation framework is considered. Specifically, the dataset is randomly split into a training subset composed of 75% of the control subjects (3 men and 20 women healthy) and 75% of the people with Reinke's edema (2 men and 20 women with Reinke's edema) for each iteration. The remaining individuals constitute the testing subset, 25% of healthy people (1 man and 6 women) and 25% with Reinke's edema (1 man and 7 women). This framework is applied for variable selection and, later, for evaluating accuracy metrics by using the selected variables in an independent way, i.e., in each one of the iterations, the partitions are independent. In the first case, the model parameters are determined using the training subset, and the 95% Bayesian credible intervals (built as specified in the Section 3.5) for the model parameters are computed using the testing subset. This procedure is independently repeated 100 times. Then, the variables associated to parameters having more than one non-null 95% credible intervals out of the 100 iterations are selected. This leads to one only set of selected features for the whole cross-validation process. In the second case, once the variables have been selected, the model parameters are determined using the training subset, and the accuracy metrics are computed using the testing subset. This procedure is repeated 100 times and the accuracy metrics are then averaged. Each regularization approach has been trained independently. Note that the second stage has been introduced to test if the concrete set of acoustic features performs well in an independent cross-validation framework. In practical applications, the first stage is applied to select the features, then the classification of the new subjects is done by applying the proposed approach with the selected features without cross-validation.

Three scenarios have been independently considered for each regularization approach, all of them start with the 25 acoustic features plus gender:

1. All the features were used by training and testing with the whole dataset, and later the previously described cross-validation scheme was performed.
2. Common principal components (CPCs) [22] were used to reduce the dimension of the variable space and, then, the approaches were applied to the selected CPCs under the defined cross-validation scheme.
3. The 95% Bayesian credible interval-based approach defined in Section 3.5 was applied to provide the most relevant features based on the previously defined cross-validation framework for variable selection. Then, the approaches are applied to the selected features under the defined cross-validation framework for accuracy metrics.

Finally, an analogous Bayesian credible interval-based approach is applied for the corresponding Bayesian regularization approaches based on independent instances (LASSO, Ridge, and Elastic Net). These methods are designed to be applied to individual instances, i.e., each subject is represented by a feature vector extracted from a single voice recording. Since the database consists of four replications of the sustained /a/ phonation for each subject, four independent cases are considered. The first one uses the first feature vector of each subject, the second case considers the second feature vector of each subject and so on, i.e., the cases are R_1, R_2, R_3 and R_4, where R_j means that only the Jth replication for each individual is used. This leads to four independent experiments with independence-based regularization approaches. The same cross-validation framework for variable selection and accuracy metrics as those defined for the replication-based approaches are used for comparison purposes. Fig. 1 summarizes the experiment capturing within-subject variability and the four experiments based on independent instances, which do not capture the within-subject variability.

Next subsection shows the experimental results obtained for the three scenarios based on replications, and for the four cases of independent instances as well as the comparison among them.

4.2. Experimental results

4.2.1. Replication-based approaches

Firstly, all the acoustic features plus gender were considered by training and testing with the whole dataset, i.e., all the subjects were considered for training and all of them for testing. No differences were found for accuracy rate, sensitivity and specificity, with the three approaches providing the same value of 0.9333 for these three metrics. AUC-ROC results were very close, larger than 0.98. Specifically, 0.9944 for LASSO, 0.9933 for Ridge, and 0.9844 for Elastic Net.

The approaches were applied to all 26 variables with the defined cross-validation scheme for accuracy metrics, and the results are shown in Table 1. The accuracy rates, sensitivities, and specificities are around 0.79, 0.81, and 0.76, respectively, for the three regularization models. The best result was obtained by Elastic Net with an accuracy rate of 0.7927, a sensitivity of 0.8150, and a specificity of 0.7671. The AUC-ROC measures are very close and around 0.88. In general, the differences are very small, so in this scenario very similar results are obtained for the three regularization methods.

The second scenario considers CPCs. Specifically, 75% of the total variability is obtained with eight CPCs. The three regularization methods with the defined cross-validation scheme were applied to these

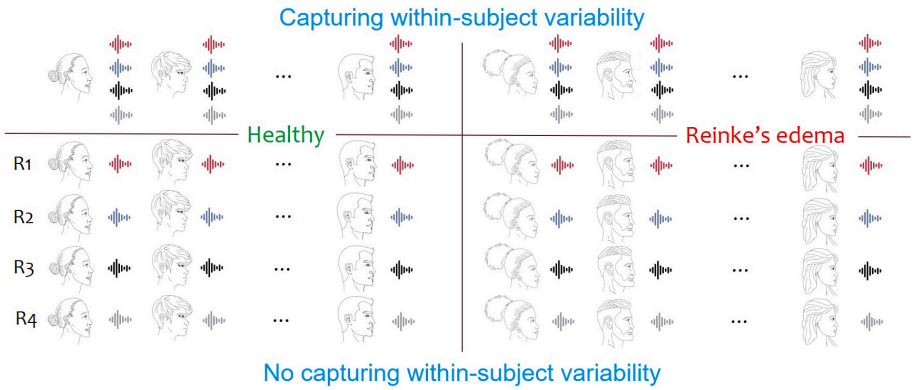


Fig. 1. Graphical scheme of the experiment capturing within-subject variability (top) and the four experiments based on independent instances (bottom).

Table 1
Means and standard deviations of accuracy rate, sensitivity, specificity, and AUC-ROC by using the replication-based regularization models with all the features under the defined cross-validation scheme for accuracy metrics (Scenario 1).

	LASSO	Ridge	Elastic Net
Accuracy rate	0.78733 (0.08352)	0.78533 (0.08128)	0.79267 (0.08254)
Sensitivity	0.80625 (0.13574)	0.80250 (0.13551)	0.81500 (0.13702)
Specificity	0.76571 (0.13548)	0.76571 (0.13548)	0.76714 (0.13568)
AUC-ROC	0.88553 (0.06604)	0.88553 (0.06643)	0.88642 (0.06717)

eight CPCs and the results are shown in Table 2. It can be observed how the loss of information provided lower accuracy rates, being now close to 0.76. The same happens for sensitivity, and specificity, which are around 0.73, and 0.80, respectively. In summary, the accuracy metrics have decreased, but they are still very similar for the three regularization approaches.

The third scenario considers the variable selection based on Bayesian credible intervals that has been previously described. Each replication-based regularization approach selects its own feature set under the defined cross-validation framework for variable selection. Table 3 shows the features selected for the three approaches. Note that LASSO selects 7 features, Ridge 7, and Elastic Net 5. Note that CPP, MFCC4, MFCC7, MFCC10, and SHANNON are selected by the three approaches.

Once the feature sets have been defined for each method, the regularization approaches are applied with the defined cross-validation scheme for evaluating accuracy metrics. The results are presented in Table 4. It can be observed how the best performance is provided by Ridge regression for the four accuracy metrics. The accuracy rate is 0.8893, larger than the ones corresponding to LASSO and Elastic Net, which are 0.8240 and 0.8253, respectively.

Table 5 shows the posterior estimations for the model parameters of the three considered replication-based regularization approaches. These are the mean and standard deviation of the parameter estimates obtained from the 100 iterations in the cross-validation framework. Note that standard deviations for intercept parameters, and parameters associated to GNE and sex are higher than the absolute value of the estimate itself. Therefore, the estimations of these three parameters

Table 2
Means and standard deviations of accuracy rate, sensitivity, specificity, and AUC-ROC by using the replication-based regularization models with eight CPCs under the defined cross-validation scheme for accuracy metrics (Scenario 2).

	LASSO	Ridge	Elastic Net
Accuracy rate	0.76466 (0.09776)	0.76600 (0.09639)	0.76733 (0.10005)
Sensitivity	0.72875 (0.12818)	0.73125 (0.12609)	0.73000 (0.13614)
Specificity	0.80571 (0.14574)	0.80571 (0.14574)	0.81000 (0.14649)
AUC-ROC	0.85410 (0.08973)	0.85464 (0.08977)	0.85625 (0.10005)

Table 3
Acoustic features selected by considering the replication-based regularization approaches under the defined cross-validation framework for variable selection (Scenario 3).

Features	LASSO	Ridge	Elastic Net
GNE	■		
Jitter			
Shimmer			
HNR			
CPP	■	■	■
MFCC1			
MFCC2		■	
MFCC3			
MFCC4	■	■	■
MFCC5			
MFCC6			
MFCC7	■	■	■
MFCC8			
MFCC9			
MFCC10	■	■	■
MFCC11			
MFCC12			
MFCC13			
HURST			
PERMUTATION			
PPE			
RPDE	■	■	
SHANNON	■	■	■
D2			
ZCR			
Total	7	7	5

Table 4
Means and standard deviations of accuracy rate, sensitivity, specificity, and AUC-ROC by using the replication-based regularization models considering the selected features under the defined cross-validation framework for accuracy metrics (Scenario 3).

	LASSO	Ridge	Elastic Net
Accuracy rate	0.82400 (0.09064)	0.88933 (0.07104)	0.82533 (0.07971)
Sensitivity	0.84375 (0.13690)	0.90750 (0.09504)	0.82125 (0.13560)
Specificity	0.80142 (0.14198)	0.86857 (0.12463)	0.83000 (0.13111)
AUC-ROC	0.92232 (0.05568)	0.95500 (0.04450)	0.92160 (0.05930)

Table 5
Means and standard deviations of the parameters for the replication-based regularization models considering the selected features under the defined cross-validation scheme (Scenario 3).

Parameters	LASSO	Ridge	Elastic Net
β_0 Intercept	0.26411 (1.19126)	−0.23393 (0.99422)	−0.46030 (0.62526)
β_1 GNE	1.39850 (1.41908)	–	–
β_5 CPP	8.52840 (5.00575)	9.85189 (5.67987)	2.21863 (0.60491)
β_7 MFCC2	–	8.84516 (5.11999)	–
β_9 MFCC4	−2.72866 (2.37346)	−5.54057 (3.08061)	−1.16507 (0.53925)
β_{12} MFCC7	4.21886 (2.65877)	5.20452 (2.54884)	1.56190 (0.47818)
β_{15} MFCC10	−5.81843 (4.36594)	−6.46859 (4.24449)	−1.94313 (0.70239)
β_{22} RPDE	−3.38629 (2.21472)	−8.18662 (4.04477)	–
β_{23} SHANNON	−2.19405 (1.42865)	−3.67750 (2.58735)	−1.09540 (0.32860)
γ Sex	0.10964 (1.16550)	0.23033 (1.11815)	0.59197 (0.60488)
λ_1	0.59822 (0.13337)	–	0.65412 (0.06362)
λ_2	–	0.18626 (0.07188)	0.39791 (0.09816)

come from dispersed values.

4.2.2. Independence-based approaches

Now the traditional independence-based regularization approaches LASSO, Ridge, and Elastic Net are applied to the four cases R_1, R_2, R_3 and R_4, where R_j means that only the jth replication for each individual is used. Each case is treated independently of the others, so each case contain independent instances.

An analogous Bayesian credible interval-based approach is applied and the features are independently selected for each case. The cross-validation framework defined in Section 4.1 has been also applied in this case. Table 6 shows the selected features for the three traditional regularization-based methods in the four cases. Within each method, the selected features are different for each dataset. There are important differences in the chosen features and in the number of them. The four cases select between 8 and 10 features for LASSO, with only 4 common features. For Ridge, between 7 and 10 features are selected, with 5 common features. Finally, for Elastic Net, there are between 5 and 9 features selected with only 3 of them common. This shows a great variability in number and kind of features within each method for the different cases constituted by the individual replications.

Table 6
Acoustic features selected by considering the traditional independence-based regularization approaches in the four cases under the defined cross-validation framework for variable selection.

Features	LASSO				Ridge				Elastic Net			
	R_1	R_2	R_3	R_4	R_1	R_2	R_3	R_4	R_1	R_2	R_3	R_4
GNE	■	■	■		■	■				■		
Jitter												
Shimmer				■				■				■
HNR												
CPP	■	■	■	■	■	■	■	■		■	■	■
MFCC1												
MFCC2	■	■	■		■	■	■	■	■	■		
MFCC3												
MFCC4	■	■		■		■		■		■		■
MFCC5												
MFCC6												
MFCC7	■	■	■	■	■	■	■	■	■	■	■	■
MFCC8												
MFCC9												
MFCC10	■	■	■	■	■	■	■	■	■	■	■	■
MFCC11	■			■	■			■				■
MFCC12												
MFCC13	■		■				■					
HURST												
PERMUTATION	■	■	■	■	■	■	■	■	■	■	■	■
PPE				■								
RPDE		■	■	■		■	■	■		■	■	
SHANNON	■	■		■	■	■		■	■	■		
D2												
ZCR												
Total	10	9	8	10	8	9	7	10	5	9	5	7

The variability in the feature selection considering the four cases is translated into the accuracy metrics. The defined cross-validation scheme is independently applied to each case with their selected features and the results are shown in Table 7. In LASSO approach, accuracy rates ranging from 0.8100 to 0.8580 are obtained for the different cases. Ridge approach provides accuracy rates ranging from 0.8160 to 0.8720, whereas accuracy rates for Elastic Net approach range from 0.8326 to 0.8560. Different results are also obtained for sensitivities, specificities, and AUC-ROC through the four cases.

With this experiment, it has been shown how different results for the selected variables and the accuracy metrics are obtained, depending on the concrete voice recording for each subject being considered. For the first time, it has been demonstrated that having into account the within-subject variability provides more stable results at the same time that relevant features are selected and accuracy metrics keep at good values.

5. Discussion

Bayesian independence-based regularization regression methods have been widely used in many contexts (see, e.g., Kadoya et al. [24]). These methods are based on independent instances as input data. When there exists a dependent nature among some instances, methods that are able to properly address this dependency are demanded. Imperfections in technology and the very biological variability result in acoustic features that are not identical for one specific individual in a particular recording time. This leads to the concept of replication that tries to address the within-subject variability underlying the experimental design. The recording of only one phonation per individual introduces lack of confidence in the process, because if other phonations had been performed, different feature vectors representing the subject would have been obtained and, therefore, the results would have been different. Using independence-based approaches has been the common way to address automatic detection of laryngeal pathologies from speech recordings in the scientific literature [23,31,56].

Table 7
Means and standard deviations of accuracy rate, sensitivity, specificity, and AUC-ROC by using the traditional independence-based regularization approaches in the four cases under the defined cross-validation scheme.

	LASSO	Ridge	Elastic Net
R_1			
Accuracy rate	0.85800 (0.08614)	0.87200 (0.08022)	0.85600 (0.08135)
Sensitivity	0.83875 (0.13680)	0.85875 (0.12264)	0.83875 (0.12346)
Specificity	0.88000 (0.11614)	0.88714 (0.11175)	0.87571 (0.12786)
AUC-ROC	0.92696 (0.06097)	0.94035 (0.05128)	0.92785 (0.05484)
R_2			
Accuracy rate	0.83333 (0.08658)	0.83600 (0.08602)	0.84066 (0.08781)
Sensitivity	0.85000 (0.13176)	0.85375 (0.13301)	0.86000 (0.13328)
Specificity	0.81428 (0.14285)	0.81571 (0.14683)	0.81857 (0.14900)
AUC-ROC	0.91964 (0.06104)	0.92071 (0.06050)	0.92375 (0.05753)
R_3			
Accuracy rate	0.81000 (0.08958)	0.81600 (0.08995)	0.83266 (0.08554)
Sensitivity	0.83000 (0.12997)	0.81250 (0.13588)	0.85125 (0.12897)
Specificity	0.78714 (0.13993)	0.82000 (0.14303)	0.81142 (0.14050)
AUC-ROC	0.91910 (0.05740)	0.93482 (0.05191)	0.94500 (0.04826)
R_4			
Accuracy rate	0.84800 (0.07875)	0.83533 (0.07896)	0.83866 (0.06974)
Sensitivity	0.87625 (0.11581)	0.86875 (0.11148)	0.85000 (0.10952)
Specificity	0.81571 (0.12728)	0.79714 (0.13187)	0.82571 (0.12934)
AUC-ROC	0.92839 (0.05369)	0.93196 (0.05297)	0.91303 (0.07204)

Three regularization-based approaches have been implemented and applied to detect Reinke's edema based on features extracted from replicated voice recordings. The existing within-subject variability for each subject has been statistically addressed by considering that the replicated observations from a feature are measurements with errors of the real underlying feature, which is unknown. In this way, the observed replicated features act as surrogates. This idea allows to build hierarchical models based on latent variables that are handled with Bayesian methodology. Due to the way that the models have been designed, MCMC methods can be used to generate from the posterior predictive distribution.

The three replication-based regularization approaches (LASSO, Ridge, and Elastic Net) consist of variable selection and classification. A total of 26 variables have been considered (25 acoustic features plus gender). Each one provides information that may be useful for voice disorder detection. However, there are many variables to feed the classifiers, some of them highly correlated. This may produce a multicollinearity problem and overfitting. To avoid this, two variable selection approaches have been considered. The first one uses CPCs [22]. Note that this is not a conventional principal component analysis, since CPC analysis allows to properly consider the replicated measurements, because the extracted features display a correlation structure that is stable throughout the replications. This kind of analysis has been widely used in other contexts (see, e.g., [28]). However, it has the disadvantage that none of the CPCs is a feature itself, so no interpretation can be obtained in terms of the disease's effects. The second variable selection approach has been specifically proposed for this problem and it is based on Bayesian credible intervals. Relevant features are obtained from those whose regression parameter estimations do not contain 0. This variable selection method within a cross-validation scheme has provided the selection of relevant features related to the malfunctioning of the voice production system under Reinke's edema. Note that the second stage under the defined cross-validation framework is independently applied to the selected features from the first stage to test if this concrete set of selected features works well for metric performance. This step is not necessary for realtime applications, once the selected variables have been tested.

An analysis of selected features from the experiments based on Bayesian credible intervals reveals that the following five features are selected in the three considered replication-based approaches: CPP, MFCC4, MFCC7, MFCC10 and SHANNON. In the case of the method providing the best accuracy metric results, Ridge, two additional features (MFCC2 and RPDE) have been also selected to complete a seven-feature set. However, when considering the independence-based counterparts applied to the four datasets (each one composed by only one of the four replicated feature vectors for each individual) a great variability of selected features is obtained depending on the voice recording considered, ranging from 5 to 10 features per experiment and a total of 15 different features out of the 25 available acoustic features. This contrasts with the previously reported results for feature selection with replication-based regularization approaches.

The selected features provide information about how the voice production system is failing under Reinke's edema disease. CPP, obtained from the cepstrum of a sound, has shown promising results as an acoustic biomarker of dysphonia [17]. High CPP values correspond to a well-defined harmonic structure, whereas periodicity perturbations (either in amplitude or frequency) lead to a lower amplitude of the cepstral peak. Reinke's edema produces an alteration of vocal-fold vibration patterns which has been quantified by means of CPP. The important role played by MFCCs (with three coefficients selected in the three cases, or even four in the case of Ridge method) may be related to the fact that Reinke's edema patients may produce compensatory articulatory changes in response to altered vocal-fold vibration. These compensatory movements modify the resonance properties of the vocal tract. The selection of SHANNON feature lines up with previous results in the literature showing that entropy measures produce higher values in people with vocal-fold disorders in comparison to healthy ones Scalassara et al. [46]. Pathological speech is characterized by an increase in the signal unpredictability that can be quantified by the use of entropy measures. Finally, RPDE also uses the concept of entropy, in this case, to measure the uncertainty in pitch period estimation. Some physiological aspects of this pathology, such as vocal-fold asymmetry, make it difficult for these patients to maintain a stable vocal fold oscillation. These physiological aspects of Reinke's edema have been shown through the use of high-speed digital imaging and videostroboscopy by Watanabe et al. [57].

From an accuracy metric perspective, the application of the independence-based regularization approaches has also provided a great variability within each regularization method, attaining the best accuracy rate with Ridge regression for the dataset with the first voice recordings (R1). This has shown that different results can be obtained depending on the voice recording considered for each individual. In contrast, the replication-based regularization approaches have provided a reduced number of features, and greater agreement regarding selected features among the three methods, at the same time that good accuracy metrics have been obtained. The best approach has been obtained with Ridge regression, providing an accuracy rate of 0.8893, sensitivity 0.9075, specificity 0.8686, and AUC-ROC 0.9550 (see Table 4). All the four metrics outperform those obtained with the independence-based regularization approaches (see Table 7). Even more, the other comparable approach for variable selection and classification that considers replications, that was developed for Parkinson's disease detection [38], provides worse results in this context. Specifically, when applying that methodology to this dataset with the same cross-validation scheme, lower accuracy metrics were obtained, specifically, an accuracy rate of 0.8120, sensitivity of 0.80375, specificity of 0.82142, and AUC-ROC of 0.8750. Finally, it is remarkable that the combination of selecting a reduced number of relevant features, good accuracy metrics and a rigorous statistical basis make the replication-based regularization approaches worthwhile.

In certain related contexts such as in Parkinson's disease detection by voice recordings, it has become usual to use features extracted from replicated recordings of each subject as if they were independent (see, e.g., Little et al. [29] and Hariharan et al. [14], and references therein). This means that the experimental unit becomes the phonation and not the subject. Given the fact that each subject has several consecutive feature vectors (each one coming from a phonation), which are dependent, a voting-based system is usually established to decide if a subject is classified as healthy or diseased after applying an independence-based classifier to each phonation. In our case, this increases the sample size from 60 subjects (30 healthy and 30 suffering from Reinke's edema) to 240 feature vectors, which are not all independent. This artificial increase of the sample size may or may not provide better accuracy rates, but it provides incoherent results. Specifically, applying a voting system based on independence-based Ridge regularization regression, it is obtained that, for the 30 healthy subjects, 12 of them (40%) had incoherences in their own voice recording classification (not all the voice recordings were assigned to the healthy group), whereas for the 30 people suffering from Reinke's edema 11 of them (36.67%) had incoherences in a similar way. Regarding the accuracy rate, it was obtained 0.8566, which is lower than the corresponding counterpart based on replications, 0.8893. However, this is not always true, for LASSO, the voting system provides an accuracy rate of 0.8440, which is larger than 0.8240, the one from the corresponding counterpart considering within-subject variability. In order to avoid this conceptual and methodological concern, the methods addressing within-subject variability provide an only response for each subject containing all the information from all voice recordings.

The proposed CAD system relies on a voice recording experiment to detect Reinke's edema based on the phonation of the vowel /a/ in a sustained way, a feature extraction process considering a variety of relevant features and a statistical methodology for variable selection and

classification based on Bayesian regularization for replicated covariates. Any of these components could be modified or replaced to try a better approach in different ways. For example, regarding the phonation protocol, other authors have considered other vowels and their combination for detecting voice disorders (see, e.g., Oliveira et al. [39]). It would be interesting to check if it is possible to further decrease the within-subject variability and improve stability by using recordings of different sustained vowels. Another relevant CAD component is feature extraction, since it provides the main ingredient for the classifiers. We have considered an initial set of features that had shown potential in the scientific literature about vocal-fold pathologies, mixing features based on perturbation, cepstral analysis, noise, nonlinear dynamics, and entropies. The proposed variable selection procedure selected the most relevant ones for Reinke's edema detection. However, classification approaches based on replications could be applied with the same benefits to other feature sets as well. For example, PLP coefficients constitute an interesting option to test. Also filtering as RASTA could be studied for PLP coefficients providing RASTA-PLP features [36] that could be tested on databases recorded under mismatched acoustic conditions for Reinke's edema detection. Robustness on environmental noise and recording channel effects in realistic environments is a research topic of great interest that has not been fully addressed up to now for voice disorder detection. Finally, the third CAD component to discuss is the statistical methodology. The regularization-based approaches considered in this paper can be easily modified to handle other methods different from the most usual ones: LASSO, Ridge and Elastic Net. In this Bayesian context, this is achieved through the use of other shrinkage prior distributions. For example, van Erp et al. [55] provided a theoretical and conceptual comparison of nine different shrinkage prior distributions that included local Student's *t*, group LASSO, hyperLASSO, horseshoe, and discrete normal mixture in addition to LASSO, Ridge and Elastic Net. An approach that would need a different framework to handle within-subject variability would be based on nonlinearity. For example, it would be interesting extending artificial neural networks and support vector machine for replicated covariates. In an independent-based approach, they have been used in the diagnosis of voice diseases by automatic speech recognition [50]. The idea of considering replications in a proper way could also be extended to the construction of kernels, which have been successfully developed for independent instances in the problem of semi-supervised learning using a small number of training samples [35].

There is a scientific and technological challenge to develop robust CAD systems that can be incorporated into medical center protocols in such a way that they provide assistance in the diagnosis and monitoring of voice diseases to the health professionals. The proposed system, including or not modifications of its components, could be integrated into a protocol that could be used in primary care as a triage method. This would enable the family doctor to refer the patient to the appropriate hospital department based on an objective criterion that supports his or her basic knowledge of the symptoms.

6. Conclusion

The proposed CAD system capturing within-subject variability due to the multiple replications of voice recordings for each individual constitutes a robust system to address the detection of voice disorders by using acoustic features. The system relies on a voice recording experiment to detect Reinke's edema, a feature extraction process, and variable selection and classification approaches based on Bayesian regularization considering replications.

The replication-based regularization methods provide a more robust approach to the solution of the current problem than the independence-based methods, at the same time that good accuracy metrics and a relevant set of features are selected, which can be interpreted in relation to the effects of Reinke's edema on the voice production mechanisms. This study constitutes a contribution to fill in the gap provided by the lack of within-subject variability management in the scientific literature. Although the approaches have been applied in the context of an experiment specifically designed for Reinke's edema detection, they can be applied to different contexts where the replications play a key role.

Larger experiments containing different voice recording protocols in mismatched acoustic conditions and the study of other signal processing algorithms for feature extraction are issues of interest to improve the CAD system, as well as trying to explore the possible nonlinearity through the development of new replication-based variable selection and classification approaches based on kernels.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Dr. Moreno for his medical advising, and Sandra Paniagua and Esther de la O. for their work recording part of the speech database. It is also acknowledged the collaboration of the patients and healthy people who voluntarily participated in this study.

This research has been funded by *Agencia Estatal de Investigación*, Spain (Project MTM2017-86875-C3-2-R), *Junta de Extremadura*, Spain (Projects IB16054, GR18108 and GR18055), and the *European Union* (European Regional Development Funds). Lizbeth Naranjo has also been partially funded by *UNAM-DGAPA-PAPIIT* (Project IN118720), Mexico. Mario Madruga has been funded by *Ministerio de Universidades* under the doctoral fellowship FPU18/03274.

References

[1] Baghai-Ravary L, Beet SW. *Automatic speech signal analysis for clinical diagnosis and assessment of speech disorders*. Springer briefs in electrical and computer engineering - speech technology. New York: Springer; 2013.
[2] Betancourt M, Girolami M. Hamiltonian Monte Carlo for hierarchical models. In: Dipak US, Dey K, Loganathan A, editors. Current trends in Bayesian methodology with applications. Chapman & Hall/CRC Press; 2015.
[3] Buonaccorsi JP. Measurement error: models, methods and applications. Boca Raton, FL: Chapman and Hall/CRC; 2010.
[4] Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: a probabilistic programming language. J Stat Softw 2017;76(1):1–32.
[5] Çomunoglu N, Batur S, Önenerk AM. Pathology of nonneoplastic lesions of the vocal folds. In: Ahmed M, editor. Voice and swallowing disorders. IntechOpen; 2019. p. 126–75.
[6] Cordeiro H, Fonseca J, Guimarães I, Meneses C. Hierarchical classification and system combination for automatically identifying physiological and neuromuscular laryngeal pathologies. J Voice 2017;31(3) (384.E9–384.E14).
[7] Cordeiro HT, Fonseca JM, Ribeiro CM. LPC spectrum first peak analysis for voice pathology detection. Proc Technol 2013;9:1104–11.
[8] Das R. A comparison of multiple classification methods for diagnosis of Parkinson's disease. Expert Syst Appl 2010;37(2):1568–72.
[9] Genkin A, Lewis DD, Madigan D. Large-scale Bayesian logistic regression for text categorization. Technometrics 2007;49(3):291–304.
[10] Godino-Llorente JI, Gomez-Vilda P, Blanco-Velasco M. Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short-term cepstral parameters. IEEE Trans Biomed Eng 2006;53(10):1943–53.
[11] Godsill S. Particle filtering: the first 25 years and beyond. In: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2019. p. 7760–4.
[12] Gómez-Garca J, Moro-Velázquez L, Arias-Londoño JD, Godino-Llorente J. On the design of automatic voice condition analysis systems. Part III: Review of acoustic modelling strategies. Biomed Signal Process Control 2021;66:102049.
[13] Gómez-Garca J, Moro-Velázquez L, Godino-Llorente J. On the design of automatic voice condition analysis systems. Part I: review of concepts and an insight to the state of the art. Biomed Signal Process Control 2019;51:181–99.
[14] Hariharan M, Polat K, Sindhu R. A new hybrid intelligent system for accurate detection of Parkinson's disease. Comput Methods Prog Biomed 2014;113(3):904–13.
[15] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Data mining, inference, and prediction. Springer series in statistics. 2nd ed. Springer; 2009.
[16] Hastie T, Tibshirani R, Wainwright M. Statistical learning with sparsity: the lasso and generalizations. In: Chapman & Hall/CRC Monographs on Statistics and Applied Probability. 1st ed. Chapman and Hall/CRC; 2015.

[17] Heman-Ackah YD, Michael DD, Goding Jr GS. The relationship between cepstral peak prominence and selected parameters of dysphonia. J Voice 2002;16(1):20–7.
[18] Hespanhol L, Vallio CS, Menezes Costa L, Saragiotto BT. Understanding and interpreting confidence and credible intervals around effect estimates. Braz J Phys Ther 2019;23(4):290–301.
[19] Hillenbrand J, Cleveland RA, Erickson RL. Acoustic correlates of breathy vocal quality. J Speech Lang Hear Res 1994;37(4):769–78.
[20] Hoerl A, Kennard R. Ridge regression. In: Encyclopedia of statistical sciences. vol. 8. New York: Wiley; 1988. p. 129–36.
[21] Hunter EJ, Tanner K, Smith ME. Gender differences affecting vocal health of women in vocally demanding careers. Logopedics Phoniatrics Vocol 2011;36(3): 128–36.
[22] Jolliffe IT. Principal component analysis. 2nd ed. New York: Springer; 2002.
[23] Kadiri SR, Alku P. Analysis and detection of pathological voice using glottal source features. IEEE J Select Top Signal Process 2019;14(2):367–79.
[24] Kadoya S, Nishimura O, Kato H, Sano D. Regularized regression analysis for the prediction of virus inactivation efficiency by chloramine disinfection. Environ Sci Water Res Technol 2020;6:3341–50.
[25] Kantz H, Schreiber T. Nonlinear time series analysisvol. 7. Cambridge University Press; 2004.
[26] Kob M, Dejonckere P, Calderon E, Kaynar S. Simulation of differences between male and female vocal fold configuration during phonation. In: NAG/DAGA; 2009. p. 1755–6.
[27] Lee J-W, Kang H-G, Choi J-Y, Son Y-I. An Investigation of Vocal Tract Characteristics for Acoustic Discrimination of Pathological Voices. BioMed Research International; 2013 (page ID 758731).
[28] Li H. Accurate and efficient classification based on common principal components analysis for multivariate time series. Neurocomputing 2016;171:744–53.
[29] Little MA, McSharry PE, Hunter EJ, Spielman J, Ramig LO. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Trans Biomed Eng 2009;56(4):1015–22.
[30] Little MA, McSharry PE, Roberts SJ, Costello DA, Moroz IM. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. Biomed Eng Online 2007;6(1):23.
[31] Lopes L, Vieira V, Behlau M. Performance of different acoustic measures to discriminate individuals with and without voice disorders. J Voice 2020 (In press).
[32] Madruga M, Campos-Roca Y, Pérez CJ. Multicondition training for noise-robust detection of benign vocal fold lesions from recorded speech. IEEE Access 2021;9: 1707–22.
[33] Martins RHG, do Amaral HA, Tavares ELM, Martins MG, Gonçalves TM, Dias NH. Voice disorders: etiology and diagnosis. J Voice 2016;30(6) (761.E1–761.E9).
[34] Massachusetts Eye and Ear Infirmary. Voice disorders database, version 1.03 (cd-rom). Lincoln Park, NJ: Kay Elemetrics Corporation; 1994.
[35] Mhaskar H, Pereverzyev SV, Semenov VY, Semenova EV. Data based construction of kernels for semi-supervised learning with less labels. Front Appl Math Stat 2019; 5:21.
[36] Moro-Velazquez L, Gómez-Garca JA, Godino-Llorente JI, Villalba J, Orozco-Arroyave JR, Dehak N. Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's disease. Appl Soft Comput 2018;62:649–66.
[37] Naranjo L, Pérez CJ, Campos-Roca Y, Martín J. Addressing voice recording replications for Parkinson's disease detection. Expert Syst Appl 2016;46:286–92.
[38] Naranjo L, Pérez CJ, Martín J, Campos-Roca Y. A two-stage variable selection and classification approach for Parkinson's disease detection by using voice recording replications. Comput Methods Prog Biomed 2017;142:147–56.
[39] Oliveira BF, Magalhães DM, Ferreira DS, Medeiros FN. Combined sustained vowels improve the performance of the Haar wavelet for pathological voice characterization. In: 2020 International Conference on Systems, Signals and Image Processing (IWSSIP). IEEE; 2020. p. 381–6.
[40] Orozco-Arroyave JR, Belalcazar-Bolaños EA, Arias-Londoño JD, Vargas-Bonilla JF, Skodda S, Rusz J, et al. Characterization methods for the detection of multiple voice disorders: neurological, functional, and laryngeal diseases. IEEE J Biomed Health Inform 2015;19(6):1820–8.
[41] Paniagua MS, Pérez CJ, Calle-Alonso F, Salazar C. An acoustic-signal-based preventive program for university lecturers' vocal health. J Voice 2020;34(1): 88–99.
[42] Park T, Casella G. The Bayesian LASSO. J Am Stat Assoc 2008;103(482):681–6.
[43] Pérez CJ, Naranjo L, Martín J, Campos-Roca Y. A latent variable-based Bayesian regression to address recording replication in Parkinson's disease. In: EURASIP, editor, *proceedings of the 22nd European signal processing conference, EUSIPCO 2014*. Lisbon, Portugal: IEEE; 2014. p. 1447–51.
[44] Sáenz-Lechón N, Godino-Llorente JI, Osma-Ruiz V, Gómez-Vilda P. Methodological issues in the development of automatic systems for voice pathology detection. Biomed Signal Process Control 2006;1:120–8.
[45] Sataloff RT. Clinical assessment of voice. Plural publishing; 2017.
[46] Scalassara PR, Dajer ME, Maciel CD, Guido RC, Pereira JC. Relative entropy measures applied to healthy and pathological voice characterization. Appl Math Comput 2009;207(1):95–108.
[47] Schyberg YM, Bork KH, Sørensen MK, Rasmussen N. Cold-steel phonosurgery of Reinke edema evaluated by the multidimensional voice program. J Voice 2018;32 (2):244–8.
[48] Silva Fonseca E, Capobianco Guido R, Barbon Junior S, Dezani E, Rosseto Gati R, Mosconi Pereira DC. Acoustic investigation of speech pathologies based on the discriminative paraconsistent machine (DPM). Biomed Signal Process Control 2020;55:1–7.
[49] Smith BJ. BOA: an R package for MCMC output convergence assessment and posterior inference. J Stat Softw 2007;21(11):1–37.
[50] Souissi N, Cherif A. Artificial neural networks and support vector machine for voice disorders identification. Int J Adv Comput Sci Appl 2016;7(5):339–44.
[51] Tavares R, Brunet N, Costa SC, Correia S, Neto BGA, Fechine JM. Combining entropy measurements and cepstral analysis for pathological voice assessment. In: ISSNIP biosignals and biorobotics conference 2011; 2011. p. 1–5.
[52] Titze IR. Physiologic and acoustic differences between male and female voices. J Acoust Soc Am 1989;85(4):1699–707.
[53] Travieso CM, Alonso JB, Orozco-Arroyave JR, Vargas-Bonilla J, Nöth E, Ravelo-Garca AG. Detection of different voice diseases based on the nonlinear characterization of speech signals. Expert Syst Appl 2017;82:184–95.
[54] Tsanas A, Little MA, McSharry PE, Spielman J, Ramig LO. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Trans Biomed Eng 2012;59(5):1264–71.
[55] van Erp S, Oberski DL, Mulder J. Shrinkage priors for Bayesian penalized regression. J Math Psychol 2019;89:31–50.
[56] Verde L, De Pietro G, Sannino G. A methodology for voice classification based on the personalized fundamental frequency estimation. Biomed Signal Process Control 2018;42:134–44.
[57] Watanabe T, Kaneko K, Sakaguchi K, Takahashi H. Vocal-fold vibration of patients with Reinke's edema observed using high-speed digital imaging. Auris Nasus Larynx 2016;43(6):654–7.
[58] Wu H, Soraghan J, Lowit A, Di Caterina G. Convolutional neural networks for pathological voice detection. In: 2018 40th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2018. p. 1–4.
[59] Yamauchi A, Yokonishi H, Imagawa H, Sakakibara K-I, Nito T, Tayama N, et al. Age-and gender-related difference of vocal fold vibration and glottal configuration in normal speakers: analysis with glottal area waveform. J Voice 2014;28(5): 525–31.
[60] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B 2005;67(2):301–20.

Chapter 5

Multicondition Training for Noise-Robust Detection of Benign Vocal Fold Lesions From Recorded Speech

Title:

Multicondition Training for Noise-Robust Detection of Benign Vocal Fold Lesions From Recorded Speech

Authors and affiliation:

Mario Madruga[a], Yolanda Campos-Roca[b], Carlos J. Pérez[a]

[a]Universidad de Extremadura, Departamento de Matemáticas, Spain
[a]Universidad de Extremadura, Departamento de Tecnología de los Computadores y las Comunicaciones, Spain

Journal:

IEEE Access

DOI:

10.1109/ACCESS.2020.3046873

Abstract: This study evaluates the effects of Multicondition Training (MCT) on computer aided diagnosis systems for voice quality assessment associated to exudative lesions of Reinke's space. This technique adds various noise conditions to the speech recordings in order to recreate realistic acoustic environments. Four different databases (Massachussets Eye and Ear Infirmary, UEX-Voice, Saarbrücken, and Hospital Universitario Príncipe de Asturias) recorded in very different acoustic environments are used. We compare the outcomes of random forest classifier models comprising feature selection, hyperparameter tuning, and cross-validation attending the specific MCT schema used to separate healthy from pathological subjects for three diseases (nodules, polyps, and Reinke's edema). Apart from the clean case baseline, an asymmetric (one subject recording is affected only by one noise recording) and two symmetric (one subject recording is affected by all the noise recordings) noise-based MCT scenarios are considered. These scenarios are created by adding realistic acoustic noise of different types to the sustained /a/ vowel recordings. The symmetric approaches are affected by methodological concerns and are tested with a comparative purpose, to emphasize these issues. Experimental results highlight the drawbacks of symmetric MCTs and exclude these techniques as a viable option. In contrast, asymmetric MCT is proven to be a suitable noise-robust approach to build a diagnosis system for exudative lesions of Reinke's space, as performance obtained with the resulting classifiers is not far from the performance obtained for clean training.

Keywords: Acoustic features, computer aided diagnosis (CAD), machine learning, multicondition training (MCT), nodules, polyps, Reinke's edema.

SPECIAL SECTION ON BEHAVIORAL BIOMETRICS FOR EHEALTH AND WELL-BEING

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 4, 2020, accepted December 13, 2020, date of publication December 23, 2020, date of current version January 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3046873

Multicondition Training for Noise-Robust Detection of Benign Vocal Fold Lesions From Recorded Speech

MARIO MADRUGA[1], **YOLANDA CAMPOS-ROCA**[2], **AND CARLOS J. PÉREZ**[1]
[1]Departamento de Matemáticas, Universidad de Extremadura, 10003 Cáceres, Spain
[2]Departamento de Tecnología de los Computadores y las Comunicaciones, Universidad de Extremadura, 10003 Cáceres, Spain

Corresponding author: Mario Madruga (mariome@unex.es)

This work was supported in part by the Ministerio de Ciencia, Innovación y Universidades, under Project MTM2017-86875-C3-2-R; in part by the Junta de Extremadura/European Regional Development Funds, EU, under Project IB16054, Project GR18108, and Project GR18055; and in part by the Ministerio de Ciencia, Innovación y Universidades, under Grant FPU18/03274.

ABSTRACT This study evaluates the effects of Multicondition Training (MCT) on computer aided diagnosis systems for voice quality assessment associated to exudative lesions of Reinke's space. This technique adds various noise conditions to the speech recordings in order to recreate realistic acoustic environments. Four different databases (Massachussets Eye and Ear Infirmary, UEX-Voice, Saarbrücken, and Hospital Universitario Príncipe de Asturias) recorded in very different acoustic environments are used. We compare the outcomes of random forest classifier models comprising feature selection, hyperparameter tuning, and cross-validation attending the specific MCT schema used to separate healthy from pathological subjects for three diseases (nodules, polyps, and Reinke's edema). Apart from the clean case baseline, an asymmetric (one subject recording is affected only by one noise recording) and two symmetric (one subject recording is affected by all the noise recordings) noise-based MCT scenarios are considered. These scenarios are created by adding realistic acoustic noise of different types to the sustained /a/ vowel recordings. The symmetric approaches are affected by methodological concerns and are tested with a comparative purpose, to emphasize these issues. Experimental results highlight the drawbacks of symmetric MCTs and exclude these techniques as a viable option. In contrast, asymmetric MCT is proven to be a suitable noise-robust approach to build a diagnosis system for exudative lesions of Reinke's space, as performance obtained with the resulting classifiers is not far from the performance obtained for clean training.

INDEX TERMS Acoustic features, computer aided diagnosis (CAD), machine learning, multicondition training (MCT), nodules, polyps, Reinke's edema.

I. INTRODUCTION

Human voice production can be affected by a wide range of conditions, either vocal specific like nodules, polyps, cleft lip and palate, or by other disorders which affect motor control like neurodegenerative diseases. Either way, voice quality assessment is a reliable source of information for physicians and patients for diagnosis and monitoring of the underlying disease.

Nodules, polyps, and Reinke's edema are the main lesions that occur in Reinke's space [1]. Although their etiologic factors are different, their pathologic features are quite similar and diagnosis usually relies on the clinical description of the patient. Classical voice quality assessment relies on cumbersome techniques such as videostroboscopy or laryngoscopy, procedures which are highly invasive and uncomfortable for patients, and require expensive equipment and expert practitioners. It is for such that Computer Aided Diagnosis (CAD) tools are of great interest since they can help diagnosis procedures by using voice recordings as a non-invasive biomarker. They are non-intrusive as they only perform signal processing of voice samples [2].

Different signal sources have been taken into consideration, being the most usual vocal production recordings and electroglottography (EGG) [3]. Both techniques have their pros and cons: whereas the latter one needs of specific equipment like electrodes and laryngograph, voice analysis only needs common recording equipment like microphones and

The associate editor coordinating the review of this manuscript and approving it for publication was Jiri Mekyska.

sound interfaces, being high quality devices widely available even in portable format like modern smartphones. However, such vocal recording devices are prone to be affected by interferences like environmental and electronic noise or reverberation, whereas EGG, measuring the glottal activity, is affected only by noise induced in the equipment electronics. Furthermore, the need of specific devices makes EGG less common and available. Vocal recordings will be, therefore, the subject of this study.

Research conducted in order to find reliable automatic voice quality assessment systems has considered different approaches [4]. One of them is by looking for new meaningful features, using a well known classifier. In that regard multiple research lines have been proposed, from pitch related features [5], cepstral analysis [6]–[8], non-linear analysis [9], [10] or wavelet transformation [11]. Other common route is researching a good new classifier which improves the already known ones, since new machine learning techniques are being constantly researched, and many of them have been applied to this particular field using already known features [12]. Examples are hidden Markov models (HMM) [13], gaussian mixture models (GMM) [14], support vector machines [15], random forests [16] or more recently artificial neural networks [7] and deep neural networks [17] among others. Even data augmentation techniques have been proposed, creating synthetic feature values in order to supply data for the classifiers due to the lack of pathological recordings [18], or new selection techniques, like paraconsistent machines [19].

Most of these systems are developed on voice databases collected in the best recording conditions available. The most common database is the Massachusets Eye and Ear Infirmary (MEEI) database [20], available since 1994, but nowadays some other databases have been created, like Hospital Universitario Príncipe de Asturias (HUPA), spanish database [21], Saarbrücken Voice Database (SVD), german database [22], or the Arabic Voice Pathology Database (AVPD), arabic database [23]. All of them were recorded in sound proofed rooms and even use KayPENTAX Computerized Speech Lab. However, those controlled acoustical and technical conditions can not be replicated in a real clinical environment, or from the opposite side, realistic noise conditions are not represented in the databases.

Multicondition Training (MCT) alleviates such underrepresentation by artificially adding noise to selected samples from the voices database prior any processing. That technique has been used in other application fields [24], [25] but, to the best of the author's knowledge, it has never been applied to voice quality assessment. The field of pathological voice detection represents a new challenge since the noise components caused by the pathology have to be discriminated within a noisy environment. In the present study we build MCT systems and evaluate their effects on the ability of the resulting classifier to distinguish between healthy and pathological voices affected by Reinke's space diseases such as nodules, polyps, and Reinke's edema.

II. VOICE DATABASES

We use four voice databases recorded in different environments: MEEI, well known and widely used as a research dataset, recorded in the most favorable conditions; a dataset collected at Universidad de Extremadura (UEX-Voice), recorded at a more realistic environment; SVD collected by at Institut für Phonetik, Universität des Saarlandes; and HUPA database, recorded by Universidad Politécnica de Madrid.

A. PARTICIPANTS

Details of the participants taken into consideration can be found below. All of the databases were previously sanitized in order to avoid undesired issues, as some databases lack information like some subjects' age at the time of recording, others include more than one recording for a given subject and health status, and there are even cases where a subject has samples in both healthy and pathological groups in the same database.

MEEI database, commercialized by KayPentax Corp, compiles recordings of voices affected by a wide variety of diseases along with a control group of healthy recordings as well. 53 healthy people are present, and nodules, polyps, and Reinke's edema have a representation of 18, 20, and 25 subjects, respectively.

UEX-Voice database recordings were performed in a diagnosis room at Hospital San Pedro de Alcántara (HSPdA), Cáceres [26], with no special sound isolation from aisles and surroundings (street noise, waiting rooms…). Those recordings include 24 nodules, 30 polyps, and 30 Reinke's edema samples. 30 healthy subjects were recruited among administration staff volunteers from Universidad de Extremadura during an annual health check-up, where an otorhinolaryngologist performed an evaluation and assessed a good vocal health status. All of the volunteers signed an informed consent concerning subsequent studies using the collected information.

SVD database [22] is a vast collection of recordings compiled by Institut für Phonetik at Universität des Saarlandes and the Phoniatry Section of the Caritas Clinit St. Theresia in Saarbrücken. It contains 869 healthy recordings, 17 nodules, 40 polyps, and 51 Reinke's edema samples. This huge imbalance in number had to be addressed by making a selection of healthy subjects: We tried to match the numbers of female and male subjects while keeping the average and standard deviation of the age as even as possible by matching each of the pathological utterances with a healthy one of the same sex and closest age possible, without repetitions.

HUPA database [27] was recorded by Universidad Politécnica de Madrid in Hospital Universitario Príncipe de Asturias. It contains 239 healthy, 29 nodules, 28 polyps, and 28 Reinke's edema utterances. As for SVD database, the imbalance was addressed by picking healthy subjects which matched the sex and age distribution of each of the diseases being considered, again matching healthy sex-age samples with each pathological recording without repetitions.

Table 1 shows sex and age distribution for each combination of database and disease after balancing SVD and HUPA databases.

B. RECORDING EQUIPMENT

MEEI database was recorded in a most optimal environment using KayPENTAX Computerized Speech Lab, a state-of-the-art equipment purposely designed for voice disease research, including features like professional grade audio capture or calibrated input [28]. Although recording conditions were strictly controlled, they vary among pathological and healthy voices, with different sampling rates, 50 kHz for normal vs. 25 kHz for pathological, with normal and pathological voices also recorded in different locations, which are not described but assumed to be acoustically identical [28].

Regarding UEX-Voice database, it was compiled using an AKG 520 head-worn condenser cardioid microphone attached to a TASCAM US322 interface using Audacity 2.0.5 recording software, with no special sound isolation from aisles and surroundings. The sampling rate was 44.1 kHz, and the resolution was 16 bits per sample.

SVD recordings were collected using a headset condenser microphone fed directly into a Kay elemetrics Computerized Speech Lab (CSL) station model 4300B, and recorded at 50 kHz sample rate and a bit depth of 16 bits inside a sound-treated room [29].

Finally, for HUPA database recordings were performed with the CSL 4300B equipment of Kay Elemetrics, using a condender microphone as input device, sampling both signals with a frequency of 50 kHz and 16 bits of quantization. All the recordings were taken under the same conditions and recording parameters, and were collected in a soundproof room [27].

C. VOCAL TASK

In MEEI database each subject was asked to perform a sustained phonation at a comfortable pitch and level for at least 3 seconds of the /a/ vowel, repeating the process 3 times, after which an expert speech pathologist chose the best sample for the database [28]. That sample was also trimmed down to 1 second looking for the stable part of the phonation before including it into the database.

In the case of UEX-Voice, the phonation of the /a/ vowel was kept up for at least 5 seconds in a single breath. Laryngological evaluation was performed by an otorhinolaryngologist using videostroboscopy. The leading and trailing segments of the recording were discarded prior to storing the utterance in the database. The depicted recording and research protocol was approved by the bioethics committees from both UEX and HSPdA.

SVD subjects on their side had to perform a phonation of the /a/ vowel, among other tasks which are not of interest for this study. A mid-section of the phonation was stored in the database, avoiding onset and offset segments.

TABLE 1. Age distribution by database, disease, health status, and sex.

Database	Disease	Health	Sex	Nº	Mean	Std
MEEI	Nodules	Normal	M	21	38.81	8.49
			F	32	34.16	7.87
			T	53	36.00	8.36
		Pathologic	M	1	47.00	0.00
			F	17	28.05	10.08
			T	18	29.11	10.75
	Polyps	Normal	M	21	38.81	8.49
			F	32	34.16	7.87
			T	53	36.00	8.36
		Pathologic	M	12	37.83	15.63
			F	8	55.00	14.91
			T	20	44.7	16.82
	Reinke	Normal	M	21	38.81	8.49
			F	32	34.16	7.87
			T	53	36.00	8.36
		Pathologic	M	5	50.6	14.72
			F	20	47.4	11.87
			T	25	48.04	12.22
UEX-Voice	Nodules	Normal	M	4	39.00	14.17
			F	26	41.04	11.18
			T	30	40.42	11.58
		Pathologic	M	1	64.00	0.00
			F	23	39.39	10.66
			T	24	40.42	11.58
	Polyps	Normal	M	4	39.00	14.17
			F	26	41.04	11.18
			T	30	40.42	11.58
		Pathologic	M	6	43.33	13.26
			F	24	46.21	11.83
			T	30	45.63	11.95
	Reinke	Normal	M	4	39.00	14.17
			F	26	41.04	11.18
			T	30	40.42	11.58
		Pathologic	M	3	35.67	22.19
			F	27	51.29	8.38
			T	30	47.97	11.97
SVD	Nodules	Normal	M	4	41.75	19.63
			F	13	31.92	10.87
			T	17	34.24	13.40
		Pathologic	M	4	40.25	23.10
			F	13	31.92	10.87
			T	17	33.88	14.21
	Polyps	Normal	M	23	52.00	14.49
			F	17	54.35	15.24
			T	40	53.00	14.67
		Pathologic	M	23	51.04	12.93
			F	17	54.64	15.94
			T	40	52.57	14.21
	Reinke	Normal	M	7	60.29	5.77
			F	44	53.57	11.57
			T	51	54.49	11.16
		Pathologic	M	7	60.14	5.08
			F	44	51.5	11.36
			T	51	52.69	11.07
HUPA	Nodules	Normal	M	1	18.00	0.00
			F	28	27.5	9.39
			T	29	27.17	9.39
		Pathologic	M	1	11.00	0.00
			F	28	27.46	9.79
			T	29	26.90	9.82
	Polyps	Normal	M	14	37.28	8.68
			F	14	40.29	8.17
			T	28	38.78	8.41
		Pathologic	M	14	37.07	8.30
			F	14	40.29	8.17
			T	28	38.68	8.24
	Reinke	Normal	M	12	54.00	11.09
			F	16	46.18	8.61
			T	28	49.53	10.33
		Pathologic	M	12	53.83	11.17
			F	16	46.25	8.68
			T	28	49.5	10.36

Patients in HUPA database had to perform a sustained phonation of the /a/ vowel. The resulting recording was later trimmed, discarding the first 500 ms and the last part of the utterances to avoid onset and offset issues, storing a midvowel segment of about 3 seconds length for each utterance.

III. CORRUPTION METHODOLOGY

The main problem we find in moving from a research context to a clinical one is the difference in environmental conditions. Most diagnosis rooms are much more noise affected than the labs where research recordings are usually taken. This is especially true in the case of MEEI database, where not only recording conditions are strictly controlled, but recordings are also screened in order to obtain the best examples of each disease. Therefore we have created a series of noise corruption schemata that try to replicate some of the most usual noises that could happen inside or in the surroundings of a typical diagnosis room.

A. NOISE DATABASE

There are many resources available on the Internet, with repositories containing sound samples from different sources, some of them oriented to other fields such as speech recognition. However, we have not found any published noise database for voice corruption in a CAD setting. Specifically, we were looking for sounds that meet the following requirements:

- The noise source would be common in a clinical environment.
- The recording is clean, containing one kind of noise.
- The noise is recognizable when listening, so the recording contains mostly noise from the source and not static noise.

The most suitable alternative we found is the MUSAN dataset [30], included in the OpenSLR repository.[1] It contains recordings of a variety of sounds, from which we extracted a subgroup which fulfills the aforementioned conditions. We selected 31 noise files which contain 7 different noise types. Table 2 shows the distribution of recordings and noise types present in the database. Noise classes include: indistinct voices, keyboard typing, doors (opening, closing, and squeaking), paper flicking, phone buzzers, meteorological conditions, and people walking around.

TABLE 2. Types of noises considered and number of recordings present in the corruption schemata.

Noise type	# of recordings
Babble	3
Keyboards	7
Doors	3
Paper manipulation	2
Phone buzzes	6
Rain	4
Steps	6

[1] www.openslr.org

MUSAN dataset recording characteristics remain unknown, as it is a compilation of different sources and recording situations. However, all the noise samples contained are available at a sampling rate of 16 kHz and a resolution of 16 bits per sample, with a highly variable recording length.

B. SPEECH CORRUPTION

Voice samples from both databases are intended to be affected by selected noise samples in a realistic way. In this case, noise is added to the recordings making sure that the Signal-to-Noise Ratio (SNR) does not exceed a given threshold to be configured at corruption time. We consider that noise is usually produced at a low enough level to be unnoticed by the patient or the practitioner at recording time, so the maximum noise level should remain below the desired threshold at all times during the voice recording.

In order to mimic such level, we perform the corruption by applying some gain to the additive noise in order to limit the effects of residual noise present in the voice recording, as we only have control over the former. Even though voice samples are intended to be recorded so that their signal power remains constant, one of the effects of voice diseases is the inability to control a steady output level. The same happens for noise recordings since no considered noise is stationary. Therefore, we have to ensure that the minimum signal and noise difference stays in a predefined range. Consequently, a Welch's periodogram is computed both on the voice and noise samples using a sliding window 10 ms long and a stride of 5 ms for power calculation; the window with least difference between signal and noise power is used to calculate the noise gain in order to get the desired SNR. We decided to add noise using SNRs of 20 and 30 dB, as lower levels would probably be noticed at recording time.

Noise recordings usually exceed voice recording length, so a random segment of the noise waveform is selected each time corruption was performed, adding some variability as noise samples will not be repeated in any iteration.

C. MCT APPROACHES

MCT requires a variety of conditions in the development dataset but, from that starting point, there are different ways to confront such task, which are shown in Fig. 1 and explained below.

The first one is asymmetric MCT, where the development dataset equals the size of the original dataset, but noise is added proportionally to the number of noise types present in the noise database, plus clean condition where no noise is added. In our experiments there are 7 types of noise, so for each classifier trained, 1/8 random subset of the original dataset is affected by each type of noise and the rest remain intact. As we have different number of noise recordings for each type of noise, random selection of noise recording is performed prior noise addition and subsequently we pick a random one-second clip for noise addition.

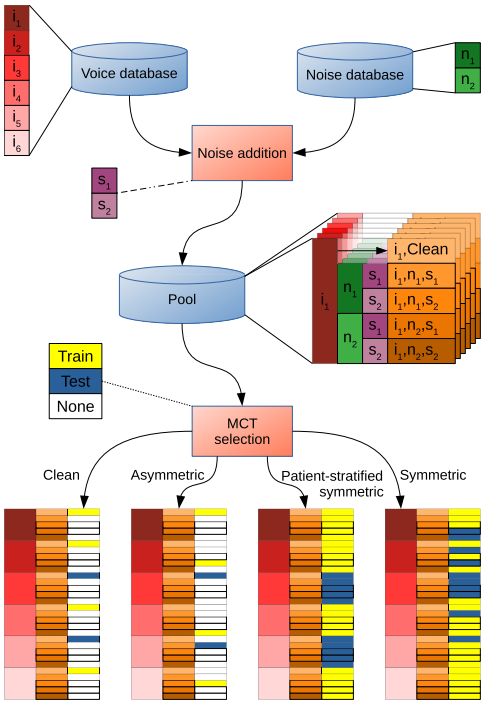


FIGURE 1. MCT selection process. The diagram shows an example: color coded are i=6 subjects, n=2 noises, and s=2 SNR levels in a 2/3 train - 1/3 test split. In our case i is the database size as shown in Table 1, n=31, s=2.

Another approach is symmetric MCT, where data augmentation is performed. This MCT technique takes the original recordings database and increases its size by adding all the different noise conditions being considered. In our case, 31 noise recordings were chosen, so the final dataset is 32 times the size of the original one (31 different noise affected datasets plus clean recordings). It is important to note that with symmetric MCT one subject appears in the dataset as many times as corruption conditions are present. This leads to two different approaches: the first one treats each recording as an independent instance and, when splitting into training and test, recordings from the same subject can lie in both subsets. It is also possible to add a stratification level in which training and test sets are not built with recordings but subjects, thus assigning all the recordings from a subject to the randomly chosen subset, either training or test, so a subject never has representation in both of them.

However, symmetric MCT methodology raises major concerns. Data augmentation can lead to good classification metrics, but constraints the generalization of the system, and performance when assessing new unknown recordings usually suffers. This is especially true in the case of symmetric MCT, since all of the individuals present in the development dataset can have representation in both training and test sets. In any case, although symmetric multicondition appears to be flawed by design, we are including the experiments and results obtained in order to further emphasize the concerns this approach rises.

Figure 1 shows the process followed to implement the aforementioned strategies. We start with one of the voice databases containing recordings for i individuals and the noise database of n noise recordings. We perform noise addition by adding the noises to the voice utterances using s different SNRs and create a pool of recordings available for MCT selection. That pool contains a total of $i+(i\times n\times s)$ utterances, i for the database size, $i \times n \times s$ for all the combinations with noises. For visual simplicity, in Fig. 1 $i = 6$, $n = 2$, $s = 2$.

From that pool, the MCT selection schema can be clean, where only clean utterances are selected; asymmetric, where each noise type is present proportionally, including clean recordings, and only one recording per individual; patient-stratified symmetric, where all of a individual recordings lay either in train or test set; symmetric, where train and test utterances are selected randomly.

IV. CAD SYSTEM

The process followed to build a CAD system for each pathology is described next, specifically, feature extraction, feature selection and classification, and cross-validation methods.

A. FEATURE EXTRACTION

An initial number of 94 features was originally considered, from which 2 are sex and age, and the rest are described next. That set includes linear and non-linear features, all of them used in previous work either for functional voice disease diagnosis or other biomedical signal analysis. Extraction methods are coded in Python either using free implementations available in public repositories or translating code from other implementations. Analysis is performed in a long term basis since all recordings have been pre-processed to match some standard parameters as shown in section II-C.

Linear features include Cepstral Peak Prominence (CPP) [6], [31], Glottal-to-Noise Excitation ratio (GNE, 4 features: mean, standard deviation, Teager Kaiser energy Operator and squared energy operator) [32], [33], Glottal Quotient (GQ, 3 features) [33], [34], Harmonic-to-Noise Ratio (HNR) [33], [35], Jitter (22 features) [33], [36], Shimmer (22 features) [31], [33], Mel Frequency Cepstral Coefficients (MFCC, 13 features) [7].

In the nonlinear subset we consider correlation dimension (D2) [37], [38], First Minimum in Mutual Information (FMMI) [38], [39], First Zero in Correlation Function (FZCF) [38], [39], Hurst's exponent (HURST) [37], [40], MultiFractal Spectrum Width (MFSW) [40], and Zero Crossing Rate [38] (ZCR).

Finally, a set of entropies and complexities was computed, including permutation entropy (PERMUTATION) [41], Pitch Period Entropy (PPE) [33], [34], Recurrence Period Density Entropy (RPDE) [42], Shannon's entropy (SHANNON) [39], [43] and Lempel-Ziv complexity (LZ, 16 features attending to different quantization bin size) [44], [45].

B. FEATURE SELECTION AND CLASSIFICATION

The number of features extracted is very high, and comparable to the development set size for each disease. One desirable characteristic in CAD systems is simplicity, as it would not only solve the problem but also provide some insight in the possible causes of the disease and why the system assigns a label to a given sample. In classification tasks using acoustic features, complexity grows as we increase the number of features considered in the solution: a low number of features can be interpretable as it is possible to discern which conditions cause abnormal values.

Moreover, big feature vectors imply the possibility of overfitting. In our case that risk is evident since the initial number of features being considered outnumbers the size of the databases used as seen in Table 1, where the sum of pathological and normal individuals is lower than the number of features for all but one database-disease combination (SVD-Reinke's edema).

Given that we do not know the optimal number of features, our approach mixes feature selection an classification techniques in order to obtain optimal, small feature subsets: The first step is getting rid of redundant information considering pairwise correlation, reducing all the feature pairs that have a high Pearson coefficient to a single representative, repeating the process for every feature pair until no high correlation pairs are present. This step is performed once and applied for all the experiments proposed, as correlation only depends on feature extraction step.

From the low correlation feature set we select, in each case, a subset making use of Recursive Feature Elimination with cross-validation (RFECV). A significant number of RFECV repetitions with random cross-validation sampling are made and the selected features of each one are collected. Then, we created an optimum subset by counting the number of times each feature is selected and choosing only the ones which exceed the median number of repetitions.

Once we have a unique feature set for each training schema we proceed to apply random forest classifiers. Prior to any training we obtain an idea of the best hyperparameters by means of a grid search over each MCT strategy-dataset combination.

Finally, making use of the selected features and hyperparameters in each combination of database, disease and corruption schema we train a set of classifiers: Starting with the most repeated single feature in the RFECV step, a random forest is trained and its performance measured. The process is repeated adding features following the number of selections order obtained by the RFECV process, until all features are used, collecting the results for every feature set size. These steps are repeated, all classifier outcomes are collected, performance metrics are averaged and accuracy rate is used as performance measurement.

Confusion matrices containing true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) are collected. Average results for accuracy rate $((TP+TN)/(TP+TN+FP+FN))$, specificity $(TN/(TN+FP))$, sensitivity or recall $(TP/(TP+FN))$, precision $(TP/(TP+FP))$, and area under the curve - receiver operating characteristic (AUC-ROC) are collected as well as their coefficient of variation $(s/\bar{x} \times 100)$, where s is the standard deviation, and $\bar{x}$ is the arithmetic mean.

Though the number of features selected by RFECV is much lower than the original feature set size, we consider that it is still high since accuracy usually reaches a plateau or decays due to overfitting, so we chose to set a limit in the number of features used by taking the lowest subset whose mean accuracy reaches a certain threshold with respect to the maximum accuracy obtained.

C. CROSS-VALIDATION

Training a classifier and thus creating a model is a process driven by chance. The outcome is highly dependent on the selection of training and test sets, especially when the development set is small. In an ideal situation any combination of training and test sets would yield equivalent models of nearly identical performance. However, real life systems do not fulfil this requirement, so we need a reliable method to check a system performance. Cross-validation replicates an experiment multiple times with different test-train splits and averages their results, thus obtaining closer to the ideal situation metrics. Two steps in the pipeline require of cross-validation, and each one is performed in a different way: RFECV and classifier training.

In feature selection, RFECV uses K-Fold cross-validation in each step to select the least relevant features and discard them. K-Fold is designed making sure to keep the subject stratification correct, meaning that we take special care in the patient-stratified symmetric case for which, instead of splitting by recording, we pick subgroups by patient, and all recordings from a given patient lay in one of the folds.

To check the possible performance impact of MCT schemata, we perform cross-validation using a stratified shuffle split strategy, where in each iteration we randomly choose a portion of healthy and sick patients for the training set and the rest for test set. In the cases of clean and asymmetric MCT that task is trivial since pathological voice stratification is enough, keeping the normophonic-pathological proportion constant in training and test sets. However, in the case of symmetric corruption the multiplicity of recordings from each patient needs a closer look.

Two options arise, and both of them are tested: firstly, a simple shuffle and splitting technique on the recordings is performed, so we do not care if a patient had recordings in both training and test sets; secondly, a patient-stratified shuffling and splitting is performed along the usual pathological stratification, ensuring that all the recordings from a given individual lay in either training or test sets while maintaining the normophonic-pathological proportion in each one.

V. RESULTS

A. EXPERIMENTAL SETTINGS

We performed the steps detailed in Section IV: feature extraction, feature selection, classification, and cross-validation as follows, repeating the experiment several times and averaging the results. We have taken into consideration all 4 different scenarios depicted:

- Clean recordings: Using the original datasets without further manipulation.
- Asymmetric MCT: Partitioning the datasets into not overlapping equal size subsets and adding one kind of noise to each subset choosing a different noise sample for each recording. We also kept one of the partitions untouched.
- Symmetric MCT: Adding every sample from all of the noise types to the whole recording set of each dataset, thus working with an augmented database. Two different approaches were taken in this case regarding patients:
  - Patient stratified: Data manipulation in CAD training is aware of the patient, and it is taken into consideration when splitting the dataset (patient-stratified symmetric).
  - Raw datasets: Every recording is considered as an independent event (symmetric).

All vocal recordings were processed in the same way: First, all samples were trimmed down to 1 second length in order to ensure homogeneous length across databases; later, all of them were downsampled to 16 kHz prior corruption in order to match noise files sampling rate; after that, noise was added from all sources at all proposed SNRs; preprocessing was applied to the sound files prior feature extraction, normalizing amplitude to range $[-1, 1]$; and lastly, feature extraction was performed for each recording.

Highly correlated features were discarded when the Pearson coefficient exceeded 0.8. After feature discarding, most of the *feature families* such as jitter or shimmer were stripped down to one representative feature. We finally worked with the following 34 features: SEX, CPP, D2, FMMI, FZCF, GNE mean value (GNE_mean), GNE standard devition (GNE_std), GNE Teager Kaiser energy operator (GNE_SNR_TKEO), GNE squared energy operator (GNE_SNR_SEO), GQ pecentiles 5-95 (GQ_prc5-95), HNR, HURST, JITTER absolute difference (JITTER_abs_diff), LZ2, MFCC (MFCC_1-13), MFSW, PERMUTATION, PPE, RPDE, SHANNON, SHIMMER absolute difference (SHIMMER_abs_diff), and ZCR.

RFECV was performed following a 2-Fold cross-validation strategy, which consequently uses a 50/50 training/test splitting, computing 500 iterations during feature selection stage, using a random forest classifier with default parameters. For the classification task, 1000 shuffle and split repetitions were made using a 2/3 to 1/3 train/test proportion, and each train-test pair was used to train classifiers using an increasing number of features following the number of times each feature was selected in the RFECV selection step until all features were used, and their confusion matrices were collected. The threshold in accuracy for the final feature selection step was 0.975 times the maximum mean accuracy rate.

We will now detail the results obtained after training classifiers for the studied diseases: nodules, polyps, and Reinke's edema, and will compare the outcomes of using the original voice recordings, and the noise corruption scenarios proposed. Different scenarios will make use of different feature sets, which will be detailed and compared. Average results for accuracy, specificity, sensitivity, precision, and AUC-ROC will be displayed as well as their coefficient of variation.

B. NODULES

Metrics (Table 4) reveal that classifiers trained using MEEI database recordings are much more capable of a correct classification than the classifiers trained using any other database by a huge margin of more than 25% in accuracy rate: the almost perfect MEEI recordings easily achieve accuracies over 0.9 for all the experiments, no matter the corruption method, whereas the more realistic recordings of UEX-Voice, SVD, and HUPA do not get over 0.71 of accuracy, with the exception of symmetric corruption.

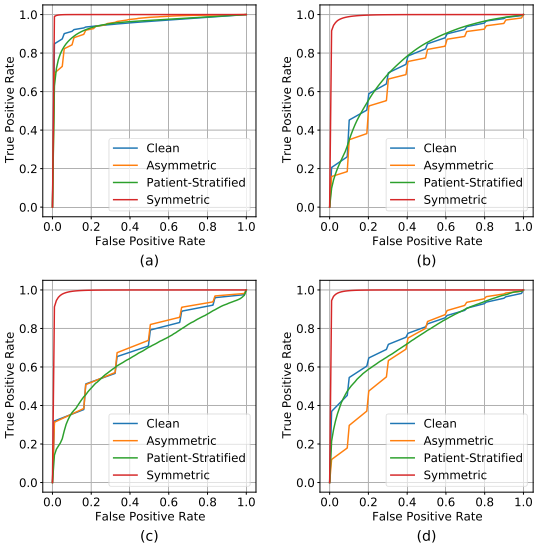


FIGURE 2. Mean ROC curves for nodules disease experiments. (a) MEEI database, (b) UEX-Voice database, (c) SVD database, (d) HUPA database.

Furthermore, the behavior of specificity, sensitivity, precision, and AUC-ROC appears to follow that of accuracy rate as a general rule, decreasing in a similar way as noise is introduced, so the system tends to maintain its ability throughout all the patients for a given database. AUC-ROC (curves on Fig. 2) values under clean conditions indicate a moderate ability to discern healthy from pathological voices for any disease. However, specificity shows a sub-par

TABLE 3. Features selected for nodules disease. Corruption cases are: Clean, Asymmetric, Patient-stratified symmetric, Symmetric.

	MEEI				UEX-Voice				SVD				HUPA			
	C	A	P	S	C	A	P	S	C	A	P	S	C	A	P	S
CPP	■			■	■	■	■						■		■	■
D2	■			■									■	■		■
FMMI																
FZCF			■	■			■									
GNE_SNR_SEO																
GNE_SNR_TKEO																
GNE_mean				■					■	■	■	■				■
GNE_std																
GQ_prc5_95																
HNR																
HURST	■	■	■					■		■	■	■				
JITTER_abs_dif									■			■				
LZ2													■			
MALE																
MFCC01	■	■	■													
MFCC02														■		
MFCC03					■	■	■	■								
MFCC04						■										
MFCC05																■
MFCC06					■		■	■								
MFCC07								■								
MFCC08								■								
MFCC09								■	■		■	■		■		
MFCC10								■								
MFCC11							■		■	■	■	■				
MFCC12													■		■	■
MFCC13															■	■
MFSW		■	■	■												
PERMUTATION						■							■	■	■	■
PPE																
RPDE													■			
SHANNON					■								■			■
SHIMMER_abs_dif		■		■					■	■						
ZCR												■				
TOTAL	4	4	4	6	4	4	5	7	5	4	4	6	7	4	4	8

performance for clean, asymmetric, and patient-stratified symmetric MCTs for all but SVD, showing that the classifier struggles to correctly classify healthy utterances, which is interesting as MEEI and UEX-Voice databases healthy group outnumbers pathological groups.

Coefficient of variation provides a deeper insight in the different performances. In MEEI database, while accuracy, sensitivity, and AUC-ROC variation tend to stay low, specificity and precision variation coefficient is three times as high. UEX-Voice, SVD, and HUPA on the other hand show a lower performance, not only in the mean values, but also in variability, with extreme cases like sensitivity for SVD database, asymmetric case, where we find that the coefficient of variation reaches 34%.

Differences in performance as we change corruption are remarkable: as we introduce noise, in the asymmetric case, performance decays slightly for MEEI and HUPA databases, but for UEX-Voice and SVD database accuracy remains almost equal, and even some variation coefficients are better. Looking at the symmetric corruption schema performance, it is very interesting to compare the results when performing two different data augmentation strategies: not taking care of patients when dividing the dataset, and splitting the training and test sets attending to the patient. In the former case performance levels rise to almost perfect classifiers with accuracy, specificity, sensitivity and precision levels between 0.94 and 0.99. For the latter case results are quite interesting: the levels achieved are generally lower than the clean and asymmetric counterparts.

Table 3 shows features selected for nodules disease when using the different database-MCT schema combinations. When taking apart MEEI database which is not realistic, and both symmetric MCT schemata because of their metholodogical issues, the only feature selected more than once under good methodological and environmental conditions is PERMUTATION.

C. POLYPS

Table 6 shows that for MEEI database, the baseline of clean case is quite good, with high accuracy, specificity, sensitivity, precision, and AUC-ROC mean levels, being specificity the worst and also the most affected by corruption, with a 13.4% performance dropping in the case of asymmetric corruption and even more for patient-stratified symmetric corruption.

Meanwhile, UEX-Voice database shows more homogeneous values: for clean, asymmetric, and patient-stratified

TABLE 4. Mean and coefficient of variation (CV) for accuracy, specificity, and sensitivity obtained for nodules disease.

		Clean		Asymmetric		Patient		Symmetric	
		Mean	CV	Mean	CV	Mean	CV	Mean	CV
MEEI	Accuracy	0,95	4,43	0,92	5,40	0,90	4,15	0,98	0,37
	Specificity	0,87	15,13	0,79	20,78	0,75	18,87	0,94	1,24
	Sensitivity	0,98	4,04	0,96	5,44	0,96	3,14	0,99	0,32
	Precision	0,93	11,01	0,88	15,38	0,84	11,35	0,97	0,92
	AUC-ROC	0,95	4,53	0,95	4,71	0,95	4,28	0,99	0,02
UEX-Voice	Accuracy	0,68	13,02	0,69	11,92	0,67	8,31	0,96	0,31
	Specificity	0,54	32,16	0,54	30,55	0,50	23,03	0,94	0,57
	Sensitivity	0,79	16,26	0,81	16,06	0,81	9,22	0,97	0,39
	Precision	0,70	23,19	0,62	21,64	0,67	15,18	0,96	0,99
	AUC-ROC	0,75	13,39	0,71	13,54	0,76	11,04	0,99	0,13
SVD	Accuracy	0,62	19,68	0,62	19,97	0,57	16,43	0,95	0,86
	Specificity	0,55	37,78	0,62	32,60	0,59	25,40	0,95	1,39
	Sensitivity	0,69	30,52	0,62	34,07	0,56	27,29	0,95	1,32
	Precision	0,67	27,33	0,64	25,02	0,58	17,95	0,95	1,24
	AUC-ROC	0,71	18,37	0,72	17,54	0,67	13,77	0,99	0,15
HUPA	Accuracy	0,71	13,35	0,65	14,74	0,65	10,31	0,95	0,70
	Specificity	0,68	20,40	0,63	27,55	0,54	22,43	0,95	1,07
	Sensitivity	0,75	19,02	0,67	25,04	0,75	13,60	0,95	1,15
	Precision	0,74	16,54	0,67	18,26	0,69	12,79	0,95	1,07
	AUC-ROC	0,77	11,93	0,71	13,77	0,75	10,48	0,99	0,08

symmetric cases, we obtain less than 5% difference in mean accuracy. SVD and HUPA databases yield worse results: whereas in the former asymmetric corruption only drops 3% and patient-stratified MCT drops 10%, the latter decays about 12% for asymmetric MCT and 15% for patient-stratified MCT. It is also remarkable the surprisingly low values obtained in the symmetric case for SVD database, which are good in comparison within the dataset, but quite low for a MCT-based comparison. Apart from that exception, symmetric corruption on its side gets overoptimistic results between 0.95 and 0.98 values for all the databases.

Once again, specificity, sensitivity, precision, and AUC-ROC follow the values obtained for accuracy, although in this case, unlike with nodules disease, specificity does not show the same weakness, with the exception of MEEI database. In this case, area under ROC curves, shown in Fig. 3 is quite good, reaching values over 0.80 for UEX-voice and HUPA databases which makes the system a fairly good detector in both cases. We can see in the flatter curves of subfigure 3(c) the difficulties with SVD database.

However, corruption affects differently all datasets: MEEI and HUPA corruption tends to be more noticeable with worse outcomes as we introduce noise, whereas UEX-Voice and SVD mean levels usually remain closer to clean condition for asymmetric and patient-stratified symmetric corruption schemata. Coefficient of variation follows the same trend: whereas asymmetric corruption in MEEI affects more negatively than in the other three databases, patient-stratified symmetric levels are better and, in some cases, even outperform the clean case with less variation for mean values in the same range.

In Table 5 we can see that in this case CPP stands as a good predictor under all circumstances. Furthermore, if we restrict the selection to realistic conditions (UEX-Voice, SVD, HUPA databases, and clean or asymmetric MCT), CPP is the only common feature selected.

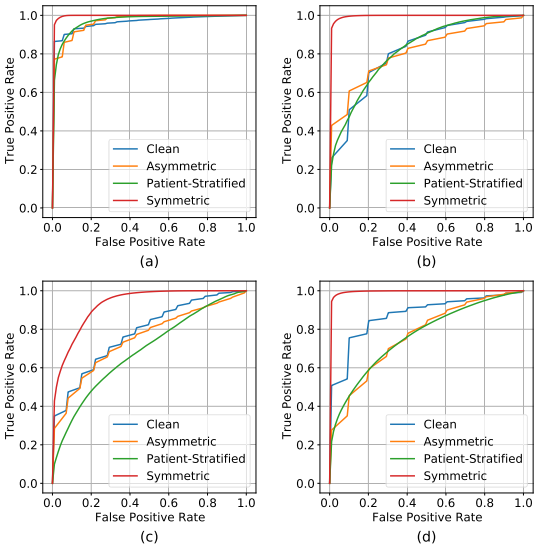


FIGURE 3. Mean ROC curves for polyps disease experiments. (a) MEEI database, (b) UEX-Voice database, (c) SVD database, (d) HUPA database.

D. REINKE's EDEMA

Once again, performances obtained, shown in Table 8, are great for MEEI database, with all metrics over 0.9 under clean training conditions, and accuracy, sensitivity, precision, and AUC-ROC above 0.96. Asymmetric and patient-stratified symmetric accuracy stay in the same range, with a penalty

TABLE 5. Features selected for polyps disease. Corruption cases are: Clean, Asymmetric, Patient-stratified symmetric, Symmetric.

	MEEI				UEX-Voice				SVD				HUPA			
	C	A	P	S	C	A	P	S	C	A	P	S	C	A	P	S
CPP			■		■	■	■	■	■	■	■		■		■	
D2	■												■	■		
FMMI																
FZCF			■					■				■			■	
GNE_SNR_SEO																
GNE_SNR_TKEO												■			■	■
GNE_mean		■	■	■				■								
GNE_std						■										
GQ_prc5_95																
HNR																
HURST	■		■				■	■			■					
JITTER_abs_dif						■				■	■	■				
LZ2									■				■			
MALE																
MFCC01																
MFCC02																
MFCC03	■	■	■	■	■		■		■		■	■				
MFCC04																
MFCC05																■
MFCC06					■	■	■	■								
MFCC07								■								■
MFCC08								■							■	■
MFCC09												■	■	■	■	
MFCC10								■				■				
MFCC11					■							■				■
MFCC12												■		■	■	■
MFCC13																
MFSW		■	■	■						■						
PERMUTATION		■	■							■		■				
PPE													■	■		
RPDE									■							
SHANNON	■															
SHIMMER_abs_dif				■									■			
ZCR								■								
TOTAL	4	4	7	4	4	4	4	9	4	4	4	9	6	4	6	6

of 4-5%, and sensitivity stays above 0.96, while specificity suffers a significant drop of 11% for asymmetric MCT, and 14% for patient-stratified MCT.

UEX-Voice, on its side, reaches good accuracy, specificity, and sensitivity levels, all over 0.72, for clean and asymmetric schemata, and results are also good for patient-stratified symmetric corruption, which gets the best accuracy and sensitivity results within the database. The same is true for HUPA database, with very similar to those of UEX-Voice mean levels for all metrics in all clean, asymmetric, and patient-stratified schemas. SVD on its side yields worse accuracy results. While UEX-Voice and HUPA performance drop with respect to MEEI database is 21%, in the case of SVD it goes further, up to 26%. Once again, symmetric MCT yields almost 1 accuracy values for every database.

Specificity, sensitivity, precision, and AUC-ROC easily follow accuracy in both, values and trend, as we introduce corrupted recordings, which shows the classifiers consistency for both healthy and pathological samples, although it is worth mentioning that for MEEI database, specificity drop is more noticeable than in any other database. AUC-ROC is remarkably good for HUPA and UEX-Voice databases, with values over 0.85. Once again, the flatter curves for SVD database shown in Fig. 4 show the difficulties the system finds in detecting diseases within this dataset.

In this case, Table 7 shows that GNE_mean is a great predictor since it is selected by 10 out of 12 database-MCT schema combinations. If we restrict ourselves to UEX-Voice, SVD, and HUPA databases, and clean and asymmetric MCT, GNE_mean is also the only common selected feature.

VI. DISCUSSION

We have studied the effects of three MCT strategies over three diseases and four databases. Results show a clear influence of the MCT strategy on the outcomes. Symmetric MCT is noteworthy as it gets very good results in every database-disease combination, not only in mean values, but also in relative dispersion. Under this type of corruption method, all considered noises are added to every utterance in the database. The result is striking, especially comparing it with patient-stratified symmetric corruption, for which the performance is assimilable to the one obtained with clean recordings and asymmetric corruption.

Athough addressed for other non physiological diseases, voice replication and data augmentation techniques are a major concern in the field of diagnosis using vocal

TABLE 6. Mean and coefficient of variation (CV) for accuracy, specificity, and sensitivity obtained for polyps disease.

		Clean		Asymmetric		Patient		Symmetric	
		Mean	CV	Mean	CV	Mean	CV	Mean	CV
MEEI	Accuracy	0,93	4,98	0,88	6,45	0,89	4,36	0,98	0,31
	Specificity	0,82	17,35	0,71	24,45	0,66	21,64	0,97	0,92
	Sensitivity	0,97	4,07	0,95	5,67	0,98	1,41	0,99	0,32
	Precision	0,93	10,01	0,88	14,67	0,90	7,50	0,97	0,90
	AUC-ROC	0,96	0,04	0,97	2,94	0,97	2,45	0,99	0,06
UEX-Voice	Accuracy	0,72	13,01	0,69	13,68	0,70	7,25	0,95	0,35
	Specificity	0,67	23,23	0,64	25,15	0,63	16,79	0,95	0,50
	Sensitivity	0,78	16,40	0,74	19,78	0,77	8,83	0,94	0,57
	Precision	0,78	13,82	0,77	14,03	0,72	11,24	0,95	0,88
	AUC-ROC	0,81	9,95	0,81	9,39	0,82	9,36	0,99	0,08
SVD	Accuracy	0,70	9,44	0,68	10,39	0,63	8,90	0,77	1,83
	Specificity	0,67	17,92	0,67	20,77	0,60	17,87	0,72	4,42
	Sensitivity	0,72	16,00	0,69	19,27	0,65	17,97	0,81	2,60
	Precision	0,72	11,80	0,69	12,91	0,64	10,70	0,80	2,02
	AUC-ROC	0,78	8,55	0,75	9,70	0,69	9,41	0,93	0,86
HUPA	Accuracy	0,79	10,43	0,69	13,52	0,67	10,62	0,97	0,52
	Specificity	0,80	17,82	0,63	26,85	0,58	21,84	0,96	0,80
	Sensitivity	0,78	16,71	0,74	20,49	0,75	14,54	0,97	0,71
	Precision	0,80	12,62	0,73	17,53	0,71	13,26	0,97	0,70
	AUC-ROC	0,87	8,41	0,76	13,20	0,76	0,08	0,99	0,11

TABLE 7. Features selected for Reinke's edema. Corruption cases are: Clean, Asymmetric, Patient-stratified symmetric, Symmetric.

	MEEI				UEX-Voice				SVD				HUPA			
	C	A	P	S	C	A	P	S	C	A	P	S	C	A	P	S
CPP				■	■		■	■		■	■	■				■
D2		■											■	■		
FMMI																
FZCF	■	■	■	■												
GNE_SNR_SEO																
GNE_SNR_TKEO						■		■								
GNE_mean		■	■	■	■	■	■	■	■	■	■	■		■	■	■
GNE_std																
GQ_prc5_95																
HNR	■								■	■						
HURST	■	■						■			■	■				■
JITTER_abs_dif													■	■		■
LZ2									■							
MALE																
MFCC01	■															
MFCC02										■						
MFCC03		■			■	■	■								■	■
MFCC04																
MFCC05																
MFCC06							■	■								
MFCC07			■	■				■								
MFCC08																
MFCC09																■
MFCC10								■								
MFCC11													■		■	
MFCC12									■	■	■	■				
MFCC13						■	■					■				■
MFSW		■	■	■								■				
PERMUTATION		■			■					■		■			■	■
PPE																
RPDE									■							
SHANNON	■											■	■		■	■
SHIMMER_abs_dif													■	■		
ZCR								■								
TOTAL	5	7	4	5	4	4	5	8	5	6	4	8	5	4	5	9

recordings [46]. The overoptimistic performance of symmetric MCT shows the methodological failure and origin of the great difference between symmetric and the rest of MCT schemata: the same subject can have, and in fact has, recordings both in training and testing sets. The presence of subjects in both sets helps the classifier, which learns

TABLE 8. Mean and coefficient of variation (CV) for accuracy, specificity, and sensitivity obtained for Reinke's disease.

		Clean		Asymmetric		Patient		Symmetric	
		Mean	CV	Mean	CV	Mean	CV	Mean	CV
MEEI	Accuracy	0,96	4,43	0,91	5,33	0,92	4,11	0,98	0,58
	Specificity	0,91	11,37	0,81	15,50	0,78	14,97	0,96	1,87
	Sensitivity	0,98	4,03	0,96	5,35	0,99	1,25	0,99	0,40
	Precision	0,97	5,81	0,92	9,64	0,98	2,54	0,98	0,70
	AUC-ROC	0,99	0,71	0,98	2,25	0,98	1,28	0,99	0,03
UEX-Voice	Accuracy	0,76	10,37	0,72	12,09	0,77	6,31	0,97	0,26
	Specificity	0,75	20,56	0,73	21,54	0,71	12,22	0,97	0,34
	Sensitivity	0,77	16,55	0,72	18,80	0,82	8,56	0,96	0,45
	Precision	0,78	12,76	0,79	14,32	0,72	11,60	0,96	0,79
	AUC-ROC	0,87	7,52	0,85	8,53	0,82	7,91	0,99	0,07
SVD	Accuracy	0,71	8,71	0,68	10,07	0,63	7,52	0,95	0,46
	Specificity	0,67	18,08	0,63	17,90	0,58	20,48	0,95	0,78
	Sensitivity	0,75	15,20	0,73	16,21	0,68	18,19	0,96	0,74
	Precision	0,73	11,48	0,71	12,53	0,65	10,34	0,96	0,70
	AUC-ROC	0,79	7,53	0,77	9,33	0,73	7,22	0,99	0,07
HUPA	Accuracy	0,76	11,54	0,74	10,68	0,70	10,11	0,94	0,83
	Specificity	0,78	18,54	0,73	19,05	0,67	17,61	0,96	0,99
	Sensitivity	0,75	19,72	0,76	19,89	0,74	15,51	0,91	1,51
	Precision	0,77	14,03	0,77	14,34	0,72	11,68	0,92	1,30
	AUC-ROC	0,85	8,83	0,84	8,85	0,79	9,62	0,99	0,06

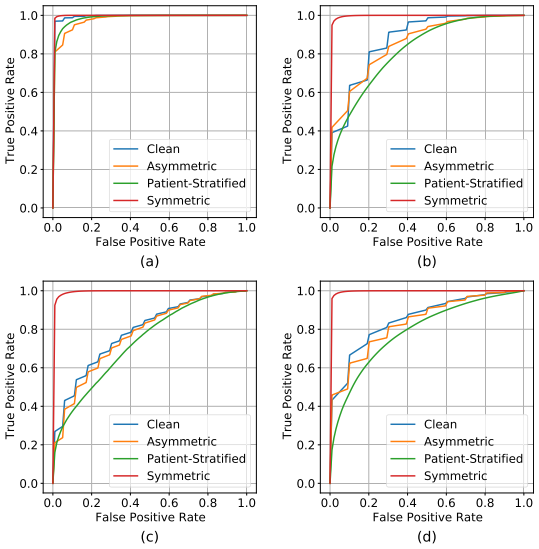


FIGURE 4. Mean ROC curves for Reinke's disease experiments. (a) MEEI database, (b) UEX-Voice database, (c) SVD database, (d) HUPA database.

to distinguish not only disease from normal recordings, but subjects themselves. This fact has a great influence in the outcome but raises strong methodological concerns.

Besides, symmetric MCT shows another weak spot in the number of features selected in each case. We can see that for most database-disease combinations, the number of features required to achieve its results is higher than any other combination. Typical numbers range around 8 selected features with sporadic cases where only 4 or 5 features are needed. On the contrary, the rest of MCT schemata behave the opposite, usually selecting 4 or 5 features with sporadic cases where up to 7 features are needed.

Although less evident than in the symmetric case, patient-stratified symmetry still involves methodological concerns. The lack of presence of subjects in both training and testing sets prevents the results to be overoptimistic, but the sample database size is still artificially increased. Asymmetric corruption and patient-stratified symmetric corruption perform similarly, but a closer look reveals that whereas asymmetry tends to yield better mean metrics, coefficient of variation is usually better in the patient-stratified symmetric case, so there appears to be a trade-off. This can be explained by the multiple repetitions of a subject within training or testing sets, which lowers speaker variability.

The results obtained with asymmetric MCT indicate that this strategy is effective to achieve noise-robustness, since the maximum degradation in mean accuracy across the twelve cases with respect to the clean case is 12.6% for HUPA-polyps combination, followed by HUPA-nodules with 8.45% and most differences below 6%. Furthermore, the results shown when using patient-stratified symmetric MCT approach do not support a performance improvement. Therefore, asymmetric MCT is proposed as the most suitable strategy to follow, being also methodologically rigorous since it does not artificially increase the sample size.

Selected features and their significance play an important role in the outcomes of the experiments. Every feature and *feature family* considered in section IV-A has its own peculiarities, strengths and weaknesses. Some of them depend on non-acoustical characteristics present in the signal, like its length in the case of entropies. This question is solved by maintaining as much homogeneity as possible across recordings in all their "physical" aspects like length, sample rate or bit depth. Moreover, nonlinear analysis requires of a careful selection of hyperparameters in order to obtain

significant results which is addressed making use of some simple strategies found in literature [38], [39], [47].

An analysis of selected features from the experiments based on the two noise conditions that do not increase the sample size, clean and asymmetric MCT, and the three realistic databases, UEX_Voice, SVD, and HUPA, reveals which features are more reliable. Table 9 summarizes those features.

TABLE 9. Most selected features by subgroups.

Subgroup	Features	# experiments	# selections
Clean	CPP	9	6
	LZ2	9	4
	MFCC03	9	4
Asymmetric	CPP	9	4
	GNE_mean	9	4
	PERMUTATION	9	4
UEX_Voice	CPP	6	5
	MFCC03	6	5
SVD	CPP	6	3
	GNE_mean	6	4
HUPA	D2	6	6
Nodules	CPP	6	3
	PERMUTATION	6	3
Polyps	CPP	6	5
Reinke	GNE_mean	6	5

Subgroups identify which parameter is fixed and its value. For noise conditions we fix values clean and asymmetric and for each one of them we iterate over database (UEX_Voice, SVD, HUPA) and disease (nodules, polyps, Reinke). If we look at databases fixed values are UEX_Voice, SVD, HUPA and the counting is carried on noise condition (clean, asymmetric) and disease (nodules, polyps, Reinke). Finally, if we focus on diseases, fixing nodules, polyps, and Reinke's disease, we iterate over noise condition (Clean, asymmetric) and database (UEX_Voice, SVD, HUPA)

Although there is a variety of highlighted features, there are some common features being selected, which are, therefore, the most robust ones as they are valid in a wide range of conditions. Cepstral analysis seems to be very useful as it includes two features: CPP, which seems to be the most reliable, and MFCC03. Glottal-to-Noise excitation also appears in every situation (fixing noise condition, database, and disease). Nonlinear features are also present with PERMUTATION, D2, and LZ2, although the latter one only appears in clean cases.

Obtained from the cepstrum of a sound, CPP has been considered the most successful acoustic feature for vocal quality assessment [48]. High CPP values correspond to a well-defined harmonic structure, whereas periodicity perturbations (commonly present due to the considered pathologies) decrease their values. Being selected in 4 out of 9 experiments under asymmetric MCT, CPP feature seems to be still reliable under noisy conditions. It does not dominate the classification processes as in the clean case, but it is as important as the other two most repeated features (PERMUTATION and GNE_mean).

GNE estimates the excitation due to vocal fold oscillations versus the excitation created by turbulent noise. It uses the correlation of Hilbert envelopes of frequency channels uniformly distributed along the spectrum, and detects turbulent noise as narrow band noise. As our noise is not bandwidth limited, such detection can be performed efficiently. Furthermore, [49] considers GNE calculation robust because it does not require estimations of the fundamental frequency, which is a complicated task, encumbered by the pathological voice, and even more difficult to perform in the presence of environmental noise.

The capability of PERMUTATION to model the characteristics of a biological system even when there is contamination by noise is known from other biomedical applications, such as studies related to brain or heart activity [50]. Its robustness for the detection of benign vocal fold lesions under noisy conditions is demonstrated with this work as it mostly appears under asymmetric MCT.

Despite the fact they are not vocal source-related features, previous scientific work has considered the use of MFCCs for the detection of laryngeal pathologies. In [51] the authors report a lower first formant frequency of vocal polyp patients based on a higher tongue position during phonation, compared to healthy subjects. This means that, for subjects with a laryngeal disorder, also the shape of the vocal tract is changed during phonation.

The system does not select common features for a given disease for different MCT schema, as neither does MCT schemata comparison for different disease as well, if we do not take account of the database. MEEI database seems to prefer nonlinear characteristics, with MFSW or FZCF unlike the other databases, where they have a low number of appearances. UEX-Voice seems to prefer cepstral analysis with a great number of MFCCs, being selected, specifically MFCC3, on the top selected features. SVD concentrates a great number of selected features around Glottal-To-Noise Excitation ratio. For HUPA database entropies seem to be the best predictor. Therefore, apart from the fact that MEEI database metrics are better than those of any other one, database has a greater impact on selected features than disease or corruption.

Table 1 shows sex and age distribution for each combination of database and disease after subject selection process to create a balanced experiment, described before. Age usually does not constitute a problem as it is relatively easy to find pathological voices for each disease in a wide range of ages, as is shown by average and standard deviation values on table 1.

Gender on its side has shown to be a more important issue in voice pathology. Women are more prone to suffer from vocal fold diseases like Reinke's edema because of their vocal fold structure [52], but gender aspects also influence the acoustical feature values obtained in signal analysis [53]. This might explain the differences in feature selection among databases: although sex is never selected as a good predictor, the proportion of male/female subjects in both healthy and pathological recordings varies throughout databases. Table 1 shows an obvious female prevalence in all diseases, and

a female/male proportion that does not match for different database-disease combinations. The effect on acoustical features, feature selection, and therefore in classification task, although interesting, can not be usually addressed due to the imbalance [12].

There is a lack of comparable results due to the novelty of applying MCT to the specific field of voice diagnosis. Moreover, robustness assessment has not been thoroughly discussed beyond some specific pitch related features [54]. However, we can check our results against those obtained in studies that overlap in the use of similar parameters (database, disease, features and/or classifier) as a baseline.

Clean results from our classifiers stand a comparison with previous related work. Accuracy, specificity, sensitivity, precision, and AUC-ROC for MEEI database are shown in Table 4 for nodules (0.95, 0.87, 0.98, 0.93, 0.95), Table 6 for polyps (0.93, 0.82, 0.97, 0.93, 0.96), and Table 8 for Reinke's edema (0.96, 0.91, 0.98, 0.97, 0.99), and establish the baseline to which corruption results will be compared. This baseline is in the vicinity of results obtained in other MEEI research studies: [13] reaches accuracies between 0.91 and 0.97, specificities between 0.73 and 0.90, sensitivities between 0.94 and 0.98, and AUC-ROC between 0.89 and 0.98 diagnosing pathological voices with a feature set consisting of MFCCs, Energy, HNR, NNE, and GNE; [55] achieves 0.95 accuracy, 0.94 specificity, 0.95 sensitivity, and 0.99 ROC using HNR, Normalized Noise Energy (NNE), GNE, and 12 MFCCs with a GMM detector, and accuracies ranging 0.88 - 0.96, specificities ranging 0.87 - 0.98, sensitivities ranging 0.88 - 0.97, and AUC-ROC ranging 0.94 - 0.99, using other feature sets; [15] achieves 0.94 accuracy discriminating nodules and polyps among others. These results, although not directly comparable because of the discrepancies on methodology since they mix diseases, use other features or build a different classifier, consolidate our clean results as a good enough baseline to which compare MCT performance.

SVD and HUPA databases have been available for a shorter period of time, thus they have not been used as thoroughly as MEEI in research, making it more difficult to find comparable studies. However, some results match our accuracy levels. Reference [56] uses different combinations of features, including glottal source features, spectral and cepstral analysis (using MFCCs) to achieve 0.78 accuracy, 0.80 sensitivity, and 0.77 specificity when classifying healthy and pathological voices from HUPA, whereas for SVD yields 0.74 accuracy, 0.75 sensitivity, and 0.71 specificity. Reference [13] uses HMM to detect the pathological voices present in the dataset with accuracy, specificity, sensitivity, and AUC-ROC ranging 0.68 - 0.82, 0.53 - 0.83, 0.78 - 0.86 and 0.72 - 0.83 respectively, whereas we detect one pathology each time against a balanced normomorphic subset and our results coherently range 0.71 - 0.79, 0.68 - 0.80, 0.75 - 0.78, and 0.77 - 0.87.

Metrics analysis confirms the different performance of MEEI in relation to the other three databases. MEEI mean levels for clean baseline are all in the same range for every disease, with great accuracy, as expected, and sensitivity levels, and very good specificity. This is due to the different recording conditions for normal and pathological speakers, and subsequent selection of disease affected utterances.

Observed performance difference when applying the methodology to any other database comes undoubtedly from the recording conditions. An inter-database MCT analysis is interesting, as we can see how the performance for asymmetric training in MEEI is not comparable to clean training with UEX-Voice, SVD, and HUPA databases, what tells us that MEEI database collecting methodology, including strictly controlled environment along with screening and selection of the included recordings, makes pathological voices easily discernible. This is an issue that has already been addressed, and as such, should be only used as starting point, and for research where classification accuracy is not the main goal [57], which is the case.

MCT applied to speaker and speech recognition, fields where this work is inspired from, gets results that support the use of this technique in this scenario. Word accuracy in [58] drops approximately 1% when a MCT with a SNR of 20 dB is applied to the word recognition problem. Those results encourage us to further study the capabilities of this technique.

VII. CONCLUSION

We have studied the effects of MCT approach in voice disease detection from sustained vowel recordings. We made use of MEEI, UEX-Voice, SVD, and HUPA databases healthy samples and nodules, polyps, and Reinke's edema affected recordings. For every database-disease combination a set of random forest classifiers was trained under four conditions: clean, asymmetric corruption, symmetric corruption, and patient-stratified symmetric corruption and their ability to discriminate between healthy and pathological samples using a set of acoustic features extracted from each condition set was tested.

The noise used in the corruption strategies (asymmetric, symmetric, and patient-stratified symmetric) was chosen and added in a way that it accurately replicates the acoustical conditions that could be found in a typical clinical environment, either in its nature, selecting the appropriate sources, and in its relative level with respect to the specific recording.

Symmetric corruption adds all considered noises to every utterance in the database, performing also a data augmentation schema. That augmentation has a great influence in the outcome, leads to overoptimistic results if no further subject-stratification is performed, and raises methodological concerns due to the artificial increase of dataset size. If the classifier is trained using a patient-stratified schema, accuracy, specificity, and sensitivity values align with those obtained using clean and asymmetric strategies, though variance is usually lower.

Asymmetric corruption, which adds noise to randomly chosen samples from the database, causes only a small degradation in accuracy, specificity, and sensitivity in every case.

However, such degradation is small enough to consider the accuracy-robustness trade-off beneficial. Furthermore, preserving the dataset size makes this strategy the only one that does not raise any concerns about its validity. We strongly advise on using it as the methodology to be used in future research.

The effects mentioned in the two previous paragraphs have been observed in all databases, what allows us to consider that the results can be extrapolated to new unknown inputs. Further work would be necessary to check the consistency of results using larger voice datasets (number of samples per disease and multi-class classification) and increasing the number of noise conditions (noise types and amount of samples per type present in the noise database).

ACKNOWLEDGMENT

The authors would like to thank Dr. Moreno for his medical advising, Sandra Paniagua and Esther de la O. for their work recording the UEX-Voice database in HSPdA, and all the voluntary individuals, patients, and healthy subjects.

REFERENCES

[1] A. Hantzakos, M. Remacle, F. Dikkers, J.-C. Degols, M. Delos, G. Friedrich, A. Giovanni, and N. Rasmussen, "Exudative lesions of Reinke's space: A terminology proposal," *Eur. Arch. Oto-Rhino-Laryngol.*, vol. 266, no. 6, p. 869, 2009.

[2] L. Baghai-Ravary and S. W. Beet, *Automatic Speech Signal Analysis for Clinical Diagnosis and Assessment of Speech Disorders*. New York, NY, USA: Springer, 2012.

[3] J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. Part I: Review of concepts and an insight to the state of the art," *Biomed. Signal Process. Control*, vol. 51, pp. 181–199, May 2019.

[4] S. Hegde, S. Shetty, S. Rai, and T. Dodderi, "A survey on machine learning approaches for automatic detection of voice disorders," *J. Voice*, vol. 33, no. 6, pp. 947.e11–947.e33, 2019.

[5] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis—Jitter, shimmer and HNR parameters," *Procedia Technol.*, vol. 9, pp. 1112–1122, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2212017313002788, doi: 10.1016/j.protcy.2013.12.124.

[6] R. Fraile and J. I. Godino-Llorente, "Cepstral peak prominence: A comprehensive analysis," *Biomed. Signal Process. Control*, vol. 14, pp. 42–54, Nov. 2014.

[7] J. I. Godino-Llorente and P. Gomez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 2, pp. 380–384, Feb. 2004.

[8] Y. D. Heman-Ackah, R. T. Sataloff, G. Laureyns, D. Lurie, D. D. Michael, R. Heuer, A. Rubin, R. Eller, S. Chandran, M. Abaza, K. Lyons, V. Divi, J. Lott, J. Johnson, and J. Hillenbrand, "Quantifying the cepstral peak prominence, a measure of dysphonia," *J. Voice*, vol. 28, no. 6, pp. 783–788, Nov. 2014.

[9] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, K. H. Malki, T. A. Mesallam, and M. Farahat Ibrahim, "Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions," *IEEE Access*, vol. 6, pp. 6961–6974, 2018.

[10] J. R. O. Arroyave, J. F. V. Bonilla, and E. D. Trejos, "Acoustic analysis and non linear dynamics applied to voice pathology detection: A review," *Recent Patents Signal Process.*, vol. 2, no. 2, pp. 96–107, Jul. 2012.

[11] E. S. Fonseca, R. C. Guido, P. R. Scalassara, C. D. Maciel, and J. C. Pereira, "Wavelet time-frequency analysis and least squares support vector machines for the identification of voice disorders," *Comput. Biol. Med.*, vol. 37, no. 4, pp. 571–578, Apr. 2007.

[12] P. Harar, Z. Galaz, J. B. Alonso-Hernandez, J. Mekyska, R. Burget, and Z. Smekal, "Towards robust voice pathology detection," *Neural Comput. Appl.*, vol. 32, pp. 15747–15757, Apr. 2018.

[13] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "An improved method for voice pathology detection by means of a HMM-based feature space transformation," *Pattern Recognit.*, vol. 43, no. 9, pp. 3100–3112, Sep. 2010.

[14] J. I. Godino-Llorente, P. Gomez-Vilda, and M. Blanco-Velasco, "Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 10, pp. 1943–1953, Oct. 2006.

[15] M. Markaki and Y. Stylianou, "Voice pathology detection and discrimination based on modulation spectral features," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 1938–1948, Sep. 2011.

[16] D. Hemmerling, A. Skalski, and J. Gajda, "Voice data mining for laryngeal pathology assessment," *Comput. Biol. Med.*, vol. 69, pp. 270–276, Feb. 2016.

[17] M. Alhussein and G. Muhammad, "Voice pathology detection using deep learning on mobile healthcare framework," *IEEE Access*, vol. 6, pp. 41034–41041, 2018.

[18] A. Ben Aicha, "Contribution of data augmentation for the prenventive detection of vocal fold precancerous lesions," *Procedia Comput. Sci.*, vol. 159, pp. 212–220, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050919313559, doi: 10.1016/j.procs.2019.09.176.

[19] E. S. Fonseca, R. C. Guido, S. B. Junior, H. Dezani, R. R. Gati, and D. C. M. Pereira, "Acoustic investigation of speech pathologies based on the discriminative paraconsistent machine (DPM)," *Biomed. Signal Process. Control*, vol. 55, Jan. 2020, Art. no. 101615.

[20] M. Eye and E. Infirmary, *Voice Disorders Database, Version. 1.03 (CD-ROM)*. Lincoln Park, NJ, USA: Kay Elemetrics Corporation, 1994.

[21] J. D. Arias-Londoño, J. I. Godino-Llorente, M. Markaki, and Y. Stylianou, "On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices," *Logopedics Phoniatrics Vocol.*, vol. 36, no. 2, pp. 60–69, Jul. 2011.

[22] *Saarbrücken Voice Database*. Accessed: May 27, 2019. [Online]. Available: http://www.stimmdatenbank.coli.uni-saarland.de

[23] T. A. Mesallam, M. Farahat, K. H. Malki, M. Alsulaiman, Z. Ali, A. Al-Nasheri, and G. Muhammad, "Development of the Arabic voice pathology database and its evaluation by using speech features and machine learning algorithms," *J. Healthcare Eng.*, vol. 2017, pp. 1–13, Oct. 2017.

[24] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, "Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 4257–4260.

[25] Y. Huang, M. Slaney, M. L. Seltzer, and Y. Gong, "Towards better performance with heterogeneous training data in acoustic modeling using deep neural networks," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 1–5.

[26] M. S. Paniagua, C. J. Pérez, F. Calle-Alonso, and C. Salazar, "An Acoustic-Signal-Based preventive program for university Lecturers' vocal health," *J. Voice*, vol. 34, no. 1, pp. 88–99, Jan. 2020.

[27] J. I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, I. Cobeta-Marco, R. González-Herranz, and C. Ramírez-Calvo, "Acoustic analysis of voice using WPCVox: A comparative study with multi dimensional voice program," *Eur. Arch. Oto-Rhino-Laryngol.*, vol. 265, no. 4, pp. 465–476, Apr. 2008.

[28] N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, and P. Gómez-Vilda, "Methodological issues in the development of automatic systems for voice pathology detection," *Biomed. Signal Process. Control*, vol. 1, no. 2, pp. 120–128, Apr. 2006.

[29] M. Pützer and W. J. Barry, "Instrumental dimensioning of normal and pathological phonation using acoustic measurements," *Clin. Linguistics Phonetics*, vol. 22, no. 6, pp. 407–420, Jan. 2008.

[30] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," 2015, *arXiv:1510.08484*. [Online]. Available: http://arxiv.org/abs/1510.08484

[31] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *J. Speech, Lang., Hearing Res.*, vol. 37, no. 4, pp. 769–778, Aug. 1994.

[32] D. Michaelis, M. Fröhlich, and H. W. Strube, "Selection and combination of acoustic features for the description of pathologic voices," *J. Acoust. Soc. Amer.*, vol. 103, no. 3, pp. 1628–1639, Mar. 1998.

[33] A. Tsanas, "Acoustic analysis toolkit for biomedical speech signal processing: Concepts and algorithms," in *Models and Analysis of Vocal Emissions for Biomedical Applications*, vol. 2. Firenze, Italy: Firenze Univ. Press, 2013, pp. 37–40.
[34] A. Tsanas and P. Gómez-Vilda, "Novel robust decision support tool assisting early diagnosis of pathological voices using acoustic analysis of sustained vowels," in *Proc. Multidisciplinary Conf. Users Voice, Speech Sing.*, 2013, pp. 3–12.
[35] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proc. Inst. Phonetic Sci.*, vol. 17, no. 1193. Amsterdam, The Netherlands, 1993, pp. 97–110.
[36] P. Lieberman, "Some acoustic measures of the fundamental periodicity of normal and pathologic larynges," *J. Acoust. Soc. Amer.*, vol. 35, no. 3, pp. 344–353, Mar. 1963.
[37] J. R. Orozco-Arroyave, E. A. Belalcazar-Bolanos, J. D. Arias-Londono, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, K. Daqrouq, F. Honig, and E. Noth, "Characterization methods for the detection of multiple voice disorders: Neurological, functional, and laryngeal diseases," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 6, pp. 1820–1828, Nov. 2015.
[38] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, vol. 7. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[39] P. Henriquez, J. B. Alonso, M. A. Ferrer, C. M. Travieso, J. I. Godino-Llorente, and F. Diaz-de-Maria, "Characterization of healthy and pathological voice through measures based on nonlinear dynamics," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 6, pp. 1186–1195, Aug. 2009.
[40] E. A. F. Ihlen, "Introduction to multifractal detrended fluctuation analysis in MATLAB," *Frontiers Physiol.*, vol. 3, p. 141, Jun. 2012.
[41] M. Riedl, A. Müller, and N. Wessel, "Practical considerations of permutation entropy," *Eur. Phys. J. Special Topics*, vol. 222, no. 2, pp. 249–262, Jun. 2013.
[42] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *Biomed. Eng. OnLine*, vol. 6, no. 1, p. 23, 2007.
[43] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul./Oct. 1948.
[44] J. R. Orozco, J. F. Vargas, J. B. Alonso, M. A. Ferrer, C. M. Travieso, and P. Henriquez, "Voice pathology detection in continuous speech using nonlinear dynamics," in *Proc. 11th Int. Conf. Inf. Sci., Signal Process. Appl. (ISSPA)*, Jul. 2012, pp. 1030–1033.
[45] A. Lempel and J. Ziv, "On the complexity of finite sequences," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 1, pp. 75–81, Jan. 1976.
[46] L. Naranjo, C. J. Pérez, Y. Campos-Roca, and J. Martín, "Addressing voice recording replications for Parkinson's disease detection," *Expert Syst. Appl.*, vol. 46, pp. 286–292, Mar. 2016.
[47] C. M. Travieso, J. B. Alonso, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, E. Nöth, and A. G. Ravelo-García, "Detection of different voice diseases based on the nonlinear characterization of speech signals," *Expert Syst. Appl.*, vol. 82, pp. 184–195, Oct. 2017.
[48] C. A. Ferrer Riesgo and E. Nöth, "What makes the cepstral peak prominence different to other acoustic correlates of vocal quality?" *J. Voice*, vol. 34, no. 5, pp. 806.e1–806.e6, Sep. 2020.
[49] J. I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, P. Gómez-Vilda, M. Blanco-Velasco, and F. Cruz-Roldán, "The effectiveness of the glottal to noise excitation ratio for the screening of voice disorders," *J. Voice*, vol. 24, no. 1, pp. 47–56, Jan. 2010.
[50] M. Zanin, L. Zunino, O. A. Rosso, and D. Papo, "Permutation entropy and its main biomedical and econophysics applications: A review," *Entropy*, vol. 14, no. 8, pp. 1553–1577, Aug. 2012.
[51] J.-W. Lee, H.-G. Kang, J.-Y. Choi, and Y.-I. Son, "An investigation of vocal tract characteristics for acoustic discrimination of pathological voices," *BioMed Res. Int.*, vol. 2013, pp. 1–11, Jul. 2013.
[52] N. Çomunoğlu, C. S. Batur, and A. M. Önenerk, "Pathology of nonneoplastic lesions of the vocal folds," in *Voice and Swallowing Disorders*. Rijeka, Croatia: IntechOpen, 2019.
[53] M. Brockmann, M. J. Drinnan, C. Storck, and P. N. Carding, "Reliable jitter and shimmer measurements in voice clinics: The relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task," *J. Voice*, vol. 25, no. 1, pp. 44–53, Jan. 2011.
[54] D. D. Deliyski, H. S. Shaw, and M. K. Evans, "Adverse effects of environmental noise on acoustic voice quality measurements," *J. Voice*, vol. 19, no. 1, pp. 15–28, Mar. 2005.
[55] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 2, pp. 370–379, Feb. 2011.
[56] S. R. Kadiri and P. Alku, "Analysis and detection of pathological voice using glottal source features," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 367–379, Feb. 2020.
[57] K. Daoudi and B. Bertrac, "On classification between normal and pathological voices using the MEEI-KayPENTAX database: Issues and consequences," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 1–6.
[58] H.-G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ASR-Autom. Speech Recognit., challenges New Millenium ISCA Tutorial Res. Workshop*, 2000, pp. 1–8.

MARIO MADRUGA received the degree in computing engineering with a capstone project involving machine learning and automatic classification, in 2010, and the degree in electrical engineering with great interest in acoustics, in 2017. He is currently pursuing the Ph.D. degree with the Universidad de Extremadura. He joined the Department of Mathematics as a Research Assistant, where he started his Ph.D. degree in researching speech processing for biomedical applications. In his final and following years, he worked in several companies as a Developer and a Systems Administrator.

YOLANDA CAMPOS-ROCA received the M.S. and Ph.D. degrees in telecommunication engineering from the Universidade de Vigo, Spain, in 1994 and 2000, respectively. From 1996 to 2000, she performed research stays that accumulate almost three years at the Fraunhofer Institute for Applied Solid State Physics, Freiburg, Germany, where she has been a Guest Researcher from the University of Vigo or a Staff Member. In 2000, she joined the School of Technology, Universidad de Extremadura, Cáceres, Spain, as an Assistant Professor and becoming an Associate Professor in 2002. Her current research interests include circuit design in the microwave and millimeter-wave range and speech processing for biomedical applications.

CARLOS J. PÉREZ received the master's degree in mathematical science in 1996 and the Ph.D. degree in mathematics in 2003. He is currently a Full Professor of statistics with the Department of Mathematics, University of Extremadura, Spain. His main research interest includes the area of Bayesian statistical inference and classification. He has authored or coauthored more than 60 JCR-indexed journal articles about statistical methodology and applications in diverse knowledge fields, including computer aided diagnosis systems based on acoustic features extracted from voice recordings. He has participated in many research projects from competitive calls and contracts. He also has been a Reviewer for some journals as *Expert Systems with Applications*, *Reliability Engineering and Safety Systems*, *Journal of Applied Statistics*, or *Annals of Applied Statistics*.

Chapter 6

Addressing smartphone mismatch in Parkinson's disease detection aid systems based on speech

Title:

Addressing smartphone mismatch in Parkinson's disease detection aid systems based on speech

Authors and affiliation:

Mario Madruga[a], Yolanda Campos-Roca[b], Carlos J. Pérez[a]

[a]Universidad de Extremadura, Departamento de Matemáticas, Spain
[a]Universidad de Extremadura, Departamento de Tecnología de los Computadores y las Comunicaciones, Spain

Journal:

Biomedical Signal Processing and Control

DOI:

10.1016/j.bspc.2022.104281

Abstract:

Objective

Voice analysis based systems offer low-cost, highly available automatic diagnostic aid for Parkinson's disease (PD) detection anywhere a smartphone with a broadband connection is available. However, reliability depends on factors affecting the communication channel. In this paper the effects of recording device mismatch are analyzed. Multicondition training (MCT) is proposed to improve robustness against that mismatch.

Methods

An experiment on 30 PD patients and 30 healthy subjects was designed. 3 vocalizations of sustained \a\ were recorded using a smartphone. These recordings, along with a simulation of 8 additional smartphones, were analyzed. Acoustical features were extracted and averaged per patient and recording device. Machine learning was used to distinguish healthy from PD patients by using different combinations of train-test smartphones.

Results

By using the same device for training and testing, a 10% best-worse mean accuracy drop is observed. The gap among different devices reaches 37%. MCT retains 90% of the maximum accuracy and exceeds a 20% mean accuracy while lowers dispersion of the aggregated results obtained with single condition. Smartphone position shows a direct impact on performance.

Conclusion

Recording device has a major effect on results. It is also found that positioning of the recording device might also be influential. Using MCT appears to improve robustness.

Significance

Results support the use of mobile devices to create an automated PD detection test. It is also encouraged to consider the use of MCT to obtain more robust and reliable results across different devices.

Keywords: Parkinson's disease, Microphone simulation, Machine learning, Diagnosis aid, Channel mismatch robustness.

Addressing smartphone mismatch in Parkinson's disease detection aid systems based on speech

Mario Madruga mariome@unex.es[1], Yolanda Campos-Roca ycampos@unex.es[2], and Carlos J. Pérez carper@unex.es[1]

[1] *Universidad de Extremadura, Departamento de Matemáticas, Spain*
[2] *Universidad de Extremadura, Departamento de Tecnología de los Computadores y las Comunicaciones, Spain*

Abstract

Objective: Voice analysis based systems offer low-cost, highly available automatic diagnostic aid for Parkinson's disease (PD) detection anywhere a smartphone with a broadband connection is available. However, reliability depends on factors affecting the communication channel. In this paper the effects of recording device mismatch are analyzed. Multicondition training (MCT) is proposed to improve robustness against that mismatch. **Methods:** An experiment on 30 PD patients and 30 healthy subjects was designed. 3 vocalizations of sustained \a\ were recorded using a smartphone. These recordings, along with a simulation of 8 additional smartphones, were analyzed. Acoustical features were extracted and averaged per patient and recording device. Machine learning was used to distinguish healthy from PD patients by using different combinations of train-test smartphones. **Results:** By using the same device for training and testing, a 10% best-worse mean accuracy drop is observed. The gap among different devices reaches 37%. MCT retains 90% of the maximum accuracy and exceeds a 20% mean accuracy while lowers dispersion of the aggregated results obtained with single condition. Smartphone position shows a direct impact on performance. **Conclusion:** Recording device has a major effect on results. It is also found that red positioning of the recording device might also be influential. Using MCT appears to improve robustness. **Significance:** Results support the use of mobile devices to create an automated PD detection test. It is also encouraged to consider the use of MCT to obtain more robust and reliable results across different devices.

Keywords: Parkinson's disease, Microphone simulation, Machine learning, Diagnosis aid, Channel mismatch, robustness.

1 Introduction

Parkinson's disease (PD) is a neurodegenerative disorder usually classified as a motor function disease. It is characterized by the presence of bradykinesia, rigidity and tremor [1]. It is estimated that more than 8 million people worldwide suffer from PD. The population group aged over 65 accumulates most of the patients, and the percentage rapidly grows as the population reaches 80 years old. The prevalence shows an age-standardized rate of 106.28 per 100,000 inhabitants, with an increasing percentage change of 155.5% in the 1990-2019 period [2].

A reliable diagnostic test for PD has yet to be available. A range of novel techniques have been developed in order to obtain an early PD diagnosis. Examples are found in [3], using electroencephalograms; [4] finds markers in magnetic resonance images; or [5], analyzing motion of upper and lower extremities. However, their availability as a general diagnostic method is low.

Voice analysis has been proposed as a non-invasive low-cost method for PD detection and assessment. 75%-95% of people with PD suffers from some sort of speech impairment [6], so voice analysis is a potential candidate to become an additional biomarker that can be used for PD diagnosis. This has led to the research of early detection of PD by analyzing different aspects of voice impairment, for which sustained vowels [7], running speech [8], and diadochokinesis tests [9] have been considered. Also, a variety of machine learning techniques have been proposed, from classical approaches [10] to state-of-the-art deep learning methods [11]. [12, 13] provide a thorough review of voice assessment approaches in the context of PD and other diseases.

One of the advantages that these non-invasive diagnostic techniques offer is ubiquity. The omnipresence of mobile technology allows to carry a recording device with broadband connectivity in the form of a smartphone. This technology gives both practitioners and patients access to advanced diagnostic aid tools almost everywhere. In fact, PD related telemedicine systems have long been developed [14], with a recent focus on mobile devices [15, 16].

The concept of channel robustness is commonly applied in relation to speech classification systems, meaning that perturbations affecting the channel do not critically decrease the system performance. The use of the term robustness with this meaning is often present in the scientific literature related to speech classification systems [17].

Channel robustness covers several factors that may produce variations in the outcomes of the experiments (noise, differences in the recording device...). In this work we focus on recording device variability. Most studies refer to a single recording device for all of their voice samples, while those showing a variety of devices, it is due to the use of a variety of databases. As a consequence, they do not offer isolation of a single element on the channel, since recording environments are markedly different. This leads to lack of generalization, a common problem in machine learning known as domain adaptation. Training datasets are often small compared to the target population, and testing data sources do not often match training data [18].

These differences can even cause an unnoticed bias, leading to unwanted discrimination [19]. However, little effort has been made in studying the variability induced by differences in the communication channel. To the authors' best knowledge, [20] is the only published study about the robustness of telemonitoring systems against the impact of such differences, in this case mobile telephony network. In fact, it points out the need of a detailed microphone comparison.

In the present study we isolate the recording device and focus the attention on the effects of this element of the communication channel. The research hypothesis is that the recording setup has significant impact on the outcome of the classifier, especially if the training process is made using a dataset obtained with a different setup than that used to record new unseen samples.

First, we study robustness against recording device variability of an automatic detection aid system. Then, we propose multicondition training (MCT) [21] as a useful generalization technique: we test its ability to improve robustness against differences between training and application devices. We show that this technique increases the ability of the machine learning model to distinguish healthy from PD diagnosed subjects when tested against previously unseen recording devices.

2 Materials and methods

The influence of smartphone as recording device has been studied by means of simulation. We used an in-house collected voice database, and designed a methodology to add the recording behavior of an assortment of smartphones to the recordings. Later, we trained a machine learning classifier to test the differences on classification accuracy due to the smartphone change. The following subsections provide details on each element of the experiment.

2.1 Participants

60 subjects volunteered for the experiment: 30 of them affected with PD, and 30 of them healthy. This number is in line with other research on PD assessment using vocal recordings [22, 23]. All subjects affected by PD were recruited in collaboration with the *Asociación Regional de Parkinson de Extremadura (ARPE)*. The inclusion criteria were that all of them should have been formally diagnosed with PD, and that their medical reports were available.

Healthy subjects were later recruited to approximately match the age and sex distribution of PD patients with the only requirement of neither having been diagnosed with PD nor having PD related symptomatology at recording time. All participants signed an informed consent. The research protocol was supervised and approved by the Bioethics Committee of the *Universidad de Extremadura*.

The group of people with PD is composed of 24 men and 6 women, with mean (standard deviation) age of 70.27 (9.54). The time since diagnosis was 9.93 (6.16) years. The Hoehn and Yahr stages ranged from 2 to 3, with a median stage of 2.5, i.e., patients in a mild-to-moderate condition. All the subjects were medicated with levodopa and the mean time since the last intake was 2.21 (1.32) hours.

2.2 Vocal task and recording equipment

For each subject included in the study, three different recordings were performed in a single session. Subjects were asked to vocalize an open \a\, for at least 5 seconds, as steadily as possible in both pitch and volume. Open \a\ is commonly used in research on automatic detection and assessment systems for voice impairment related diseases. Its ubiquity throughout different languages, and the simplicity of the experimental settings involved are the main reasons [10, 24].

The voice recordings were made using the same smartphone, model BQ Aquaris V, at a same sampling frequency of 44.1 kHz, and resolution of 16 bits. The setup for each recording session was the same: the distance from the speaker's mouth was about 30 cm, with the smartphone horizontally held, touchscreen up, and oriented so that the microphone points directly towards the source.

All the recordings were performed in the same room at the ARPE facilities under similar acoustical conditions. The room was not acoustically treated, although at recording time it was quiet. A trained person was present at all times in each recording session to ensure that all the participants properly followed the protocol, and to register the required complementary information.

Before any computation was made, all of the recordings were trimmed down in order to eliminate any leading or trailing silence. Also, one second segments were used to extract the considered speech features (see section 2.4), a duration deemed long enough [25]. This process was performed using Audacity software (release 2.0.5).

2.3 Recording device simulation

Different devices record the same sound in a disparate way, given the dissimilarities in design, component selection, and construction. The divergence can go from subtle, when comparing two specimens of the same model, to wide, when comparing models from different manufacturers, age, price range, or other features.

For our purposes, the ideal situation would be being able to record the same vocalization simultaneously using as many smartphones as possible. However, this situation is far from feasible: two devices can not be in the same position, and location influences the voice acquisition process since the human voice is not omnidirectional [26]. Furthermore, to the authors' best knowledge, database collection for any PD study has not considered the recording device variability problem.

Smartphone influence goes far beyond pure microphone frequency response. The recording system of a modern smartphone might include signal processing such as noise cancellation, compression or equalization. However, vendors do not offer information on their recording stacks. For that reason, recording device simulation seems to be an adequate alternative.

Having access to the original recording device, and to an assortment of smartphones, we can experimentally determine their individual frequency responses. We can process the recordings so that we subtract the effects of the original device, and estimate what the recording would have been if some another device had been used instead.

For the smartphone simulation, eight different devices were considered: Apple iPhone (model A1533); Apple iPhone S and Apple iPhone S(2) (model A1688), without and with an external battery attached respectively, which alters the microphone opening; iPhone SE (model A2296); OnePlus Nord (model AC2003); Realme 8 (model RMX3241); Redmi Note 9 Pro (model M2003J6B2G); and Samsung A51 (model SM-A515F/DSN). The selection criteria was having a variety of manufacturers, with high market penetration and relatively affordable. Microphone placement for each of them is shown in Fig. 1.

We followed the IEC 60268-4:2018 standard for microphone testing to the extent possible. It describes the way a microphone should be tested in order to obtain its characteristics, including frequency response and directional pattern. The standard is intended for stand alone microphones, thus not all requirements could be fulfilled since smartphone recording is a black box where processing is unknown.

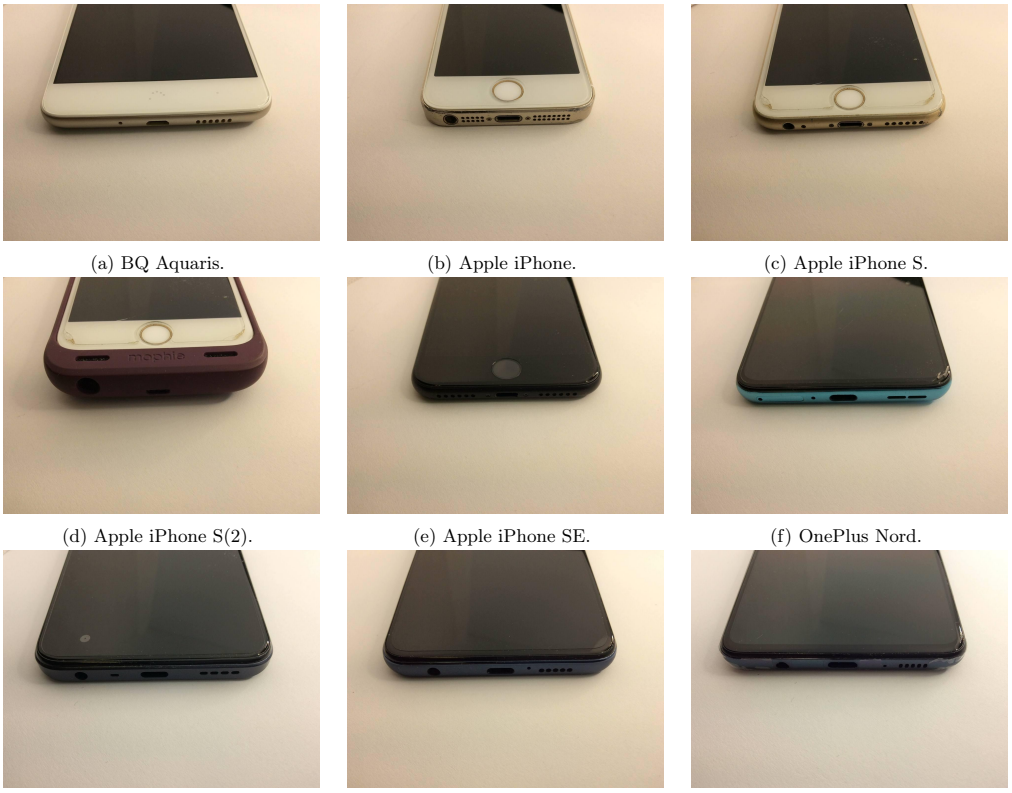
Testing was made in an anechoic chamber located at Array Processing Lab (*Universidad de Valladolid*). The loudspeaker model was Hedd Audio Type 07, which has a frequency response of -3 dB in the range 38-40,000 Hz. As for the reference microphone, we used Behringer ECM8000, with a frequency response of -3 dB in the range 20-20,000 Hz. The frequency swipes and recordings were made using Audacity software (release 2.0.5) and the sound interface was TASCAM US-322.

A total of 4 different orientations were tested: 3 different pitch angles (rotation around X-axis): 0°, position 1, Fig. 2a; 45°, position 2, Fig. 2b; and 90°, position 3, Fig. 2c. The device microphone as the center point for rotation was always considered, thus microphone placement relative to the sound source remained the same. Also, an extra position was tested by placing the smartphone on a horizontal surface as could be a table, which is not considered in the standard, Fig. 2d.

For positions 1-3 the sound source was located at a distance of 30 cm from the microphone and, due to the technical difficulties of placing the reference microphone and the device under test in the exact same spot in space, substitution method was discarded and simultaneous comparison method was used. For position 4, distance to the sound source was located 20 cm over the horizontal plane where the smartphone is lying, and the distance to the source was still 30 cm. In this case, substitution method was used, placing the reference microphone without the horizontal surface present in the same spot as the smartphone microphone would later be placed.

A continuous frequency sweep was performed in the 0-22,050 Hz range which, along with distance to the source, follows IEC 60268-4:2018 standard. Fig. 3 shows the magnitude of the Fourier Transform for a sample recording using each device simulation in each position.

Given that 1 second length recordings were used, at a sampling rate of 44,100 Hz, frequency responses were obtained with a resolution of 1 Hz. The frequency gains for BQ Aquaris-V were subtracted from each recording spectral analysis to obtain a "clean" recording without device influence. Then, frequency gains for each device were added so we could simulate each device influence. The gains were applied by transforming the signal to frequency domain using Fourier Transform, operating with the gains obtained for each device, and reconstructing the signal by means of Inverse Fourier Transform.



(a) BQ Aquaris. (b) Apple iPhone. (c) Apple iPhone S.

(d) Apple iPhone S(2). (e) Apple iPhone SE. (f) OnePlus Nord.

(g) Realme 8. (h) Redmi Note 9. (i) Samsung A51.

Figure 1: Microphone placement for all the devices.

2.4 Feature extraction

35 features were initially considered, including Cepstral Peak Prominence, Correlation Dimension, First Minimum in Mutual Information, Glottal to Noise Excitation, Harmonic to Noise Ratio, Hurst's Exponent, Jitter, Lempel Ziv Complexity, Mel Frequency Cepstral Coefficients, Multifractal Spectrum Width, Permutation Entropy, Pitch Period Entropy, Recurrence Period Density Entropy, Shannon's Entropy, Shimmer, and Zero Crossing Rate.

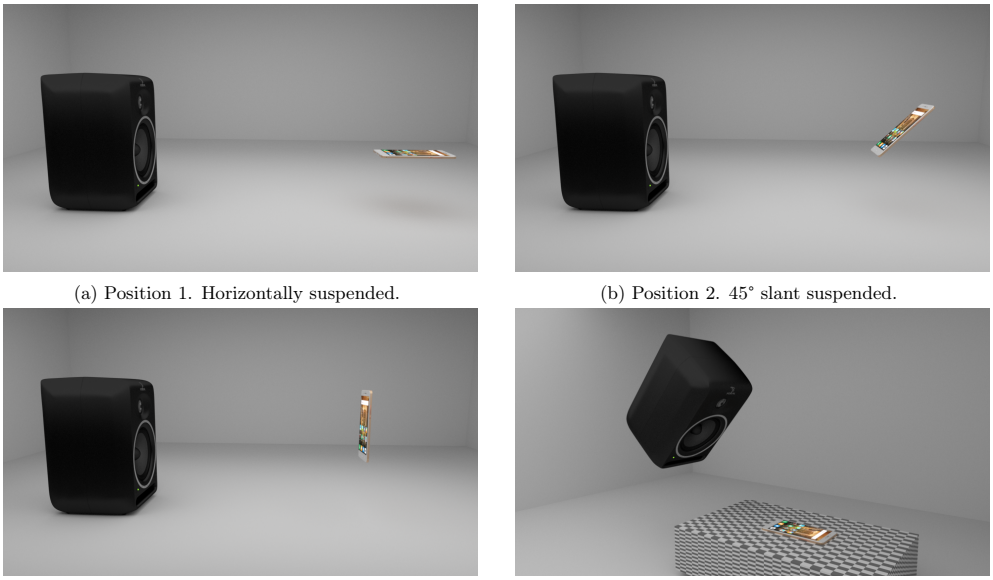
More detailed information on the considered features can be found in [12, 27]. They have been widely used in studies on pathological speech since they measure different speech impairment aspects. Also, a feature selection process is performed (see Subsection 2.5) to select and employ only the most useful ones.

2.5 Variable selection and classification

As stated in section 2.2, three vocal samples were collected from each participant. Those samples were individually processed, simulating 8 additional devices in 4 positions, and totaling 36 device/position combinations. Features were extracted for all the subjects, obtaining a data matrix of 180 voice samples $\times$ 35 acoustical features for each combination. Later, the values for each feature were averaged per patient, reducing the matrix size to 60 $\times$ 35 and the result was used as input data.

Different experiments were conducted by changing train and test datasets. For each device/position combination we built a machine learning model consisting in three steps: feature selection, grid search, and classification. The goal was to maximize accuracy. This process follows the methodology shown in [27]. In this case, we used a passive aggressive classifier because it yielded high accuracy levels and low computation times in early research stages.

The initial number of features is large compared to the number of voice samples, with a ratio close to 1/2. This could lead to overfitting problems in the train phase, and limited statistics for valuable results. Therefore, reducing the number of features is a critical step in the pipeline.



(a) Position 1. Horizontally suspended.

(b) Position 2. 45° slant suspended.

(c) Position 3. Vertically suspended.

(d) Position 4. On a table.

Figure 2: Test setup for each source-smartphone positioning.

Feature selection was performed using Recursive Feature Elimination with Cross Validation (RFECV). This technique obtains an optimal size smaller feature subset by discarding the least important features in an iterative process. This was repeated 500 times using a 2-fold cross validation scheme. The number of repetitions was chosen looking for stability, and is based in the law of large numbers (LNN), which states that the larger the number of trials, the closer to the expected value the average will be. The resulting subsets were stored and used in a sorting mechanism so that the features could be ranked by their prevalence in the selections. This ranking would be later used to obtain a small working subset.

Grid search looks for the best hyperparameters given a dataset and a classifier. A passive aggressive classifier was used, matching the one used in the feature selection step, and training data was set to match feature selection as well. At this point, the system must be oblivious to the testing samples to be used.

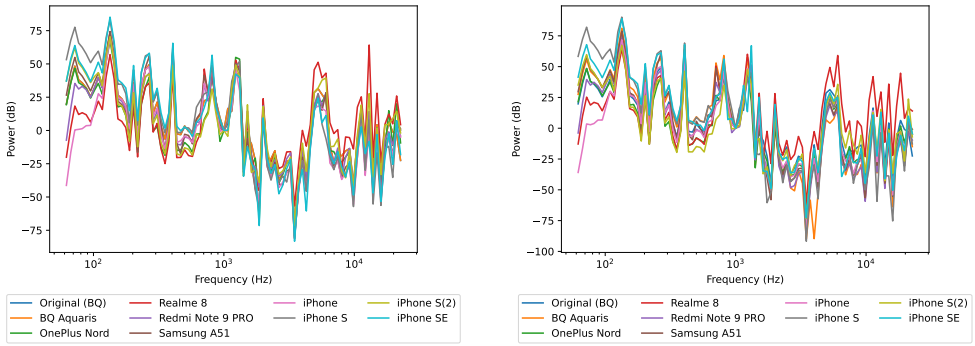
Finally, for each train/test device and position combination, we found the feature subset that yielded the best accuracy for that specific combination, which brought to an end the feature selection process. Those subsets are small compared to the initial one, therefore small compared to the dataset size, alleviating the feature number to sample size ratio problem. To obtain these small sets we trained the classifier 35 times using the n most important features according the rank obtained in the feature selection phase, being $n = 1 \dots 35$. Finally, we found which classifier yielded the highest accuracy, thus finding a small feature subset for each cross validation split. This was repeated for each train/test device and position combination.

We used a stratified shuffle and split strategy as model validation and generalization technique. Cross validation enables us to generalize the performance of the machine learning model and obtain an estimate of the classification accuracy if tested against new samples, as long as those samples come from a dataset with similar statistical characteristics. We are looking specifically towards the influence of recording devices on the outcomes. Therefore, it is of interest to maintain every other constraint constant. Stratification ensures that the proportion of healthy/pathological subjects is constant across training and testing sets. The number of splits was 1000, with a 2/3 (40 subjects) training size and 1/3 (20 subjects) testing size. The number of splits was chosen based on LNN.

Also, random sampling was conditioned so the *n-th* split for each phone-position combination selects the same individuals for testing and training, thus the differences in accuracy can only be attributed to splitting.

2.6 Multicondition training

Whenever a classifier is built with samples from a single data acquisition setup, the system may specialize in the environmental characteristics of the experimental setup, and may lack accuracy if tested with samples from other sources. Therefore, it is necessary to develop strategies to avoid this problem. MCT, which takes into consideration



(a) Position 1. Horizontally suspended.

(b) Position 2. 45° slant suspended.

(c) Position 3. Vertically suspended.

(d) Position 4. On a table.

Figure 3: Test setup for each source-smartphone positioning.

the variability of acquisition conditions in the training dataset [21], has proven to be useful to improve robustness of classifiers for Reinke's space diseases in an environment affected by noise [28].

Differences in the recording equipment and its relative position to the subject may also degrade the performance when this source of variability has not been taken into account in the training process. The robustness of a multicondition trained classifier based on the different smartphone frequency responses has been tested. The performance of this system is compared with the aggregated results obtained when training with a smartphone recordings and testing with a different smartphone; the aimed improvement should be assessed not only in the mean accuracy, but also in the dispersion of the results obtained.

Among the different multicondition strategies available, [28] has shown that asymmetry (using only one recording per subject independently of the number of recording conditions present in the dataset) is the right strategy.

Training phase for MCT follows the same schema than single condition training: A feature selection phase, followed by grid search and classification. Two different approaches were studied: First, each recording in the train set was affected by a randomly selected device, and all the recordings in the test set were affected by the same device, named all to one; secondly, both train and test sets were affected by a randomly selected device, named all to all. The devices were chosen so that the proportion of recordings affected by each device was constant, with the limitation of split size and number of devices being coprime integers.

Finally, the *umpteenth* split for each cross validation step selects the same individuals for each set, which are also the same individuals for the *n-th* split in single condition training (see Subsection 2.5). Thus, the differences in results between steps in cross validation are to be attributed only to splitting, and the differences between single, all to one MCT, and all to all MCT, can be attributed merely to the training strategy.

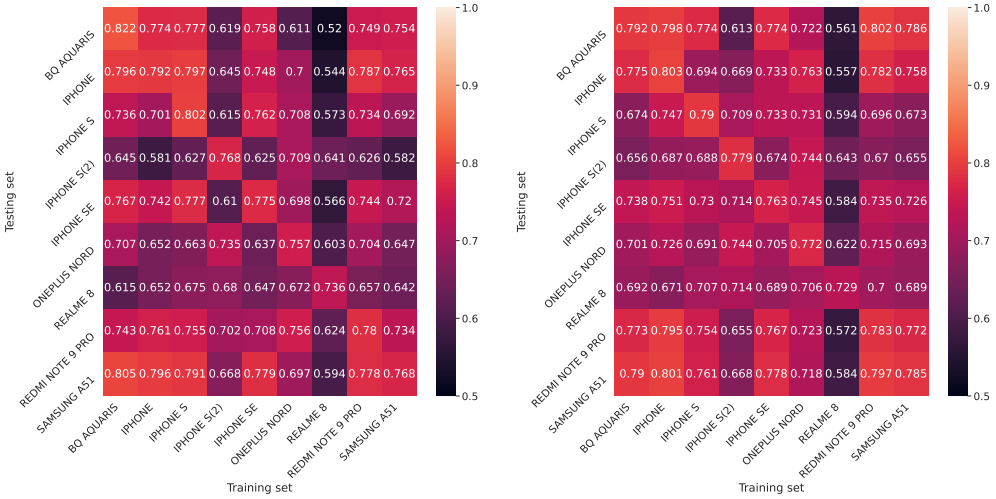
2.7 Statistical analysis

Descriptive statistics such as mean, standard deviation, and coefficient of variation have been considered. Coefficient of variation is a dimensionless relative dispersion measure that is defined as $CV = s/\bar{x}$, where s stands for standard

deviation and $\bar{x}$ for mean. Statistical hypothesis tests have been used to report statistically significant differences between groups. When normality condition could be assumed, unpaired t-tests for the homoskedastic and heteroskedastic cases were applied because of their statistical power [29]. Otherwise, the non-parametric counterpart (Mann-Whitney U test) was applied [30]. Both tests provide a p-value that can be thought of as the probability of finding the data under the assumption of the null hypothesis, i.e. under the hypothesis of no difference between groups. P-values lower than 0.05 reported statistically significant differences.

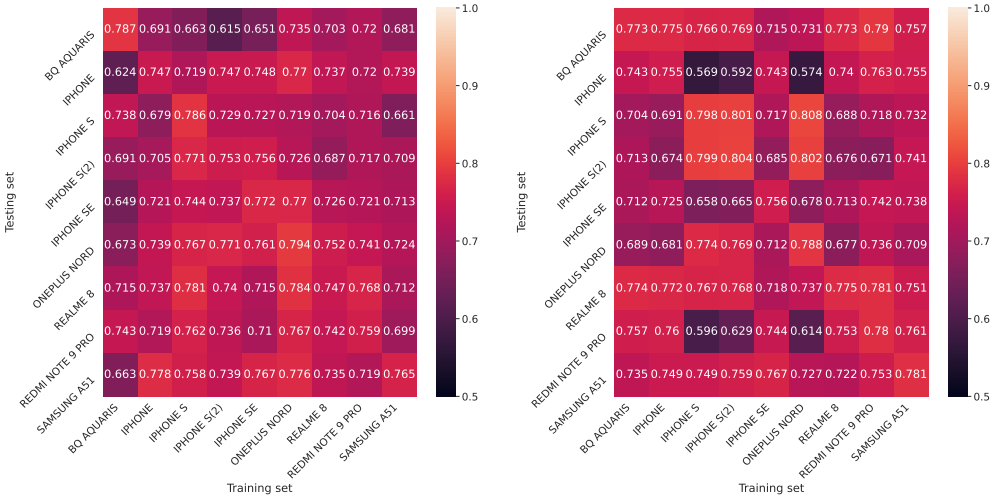
3 Results and discussion

We have studied the influence that changing the recording device and its relative position to the subject might have on the performance of the system.



(a) Position 1. Horizontally suspended.

(b) Position 2. 45°slant suspended.



(c) Position 3. Vertically suspended.

(d) Position 4. On a table.

Figure 4: Classification accuracy obtained with every training-testing device combination for each of the device positions considered.

Fig. 4 shows the accuracy obtained for all the combinations for training device, testing device, and position, as

discussed in section 2.3. It is noticeable how the BQ Aquaris-BQ Aquaris-Position 1 combination yields the best accuracy overall (0.822). The original recording device and experimental setup should be expected to get the best results since the recordings have not been processed and nothing had to be simulated.

Not every position shows an even behavior. Fig. 4c shows a higher homogeneity in results for position 3, with a more equal "heat" across all training-testing combinations than Figs. 4a, 4b, and 4d. For a smartphone common use case, the microphone points towards speaker's face when the user is making a phone call; that direction is normal to the smartphone screen so it is pointing towards sound source in position 3 experiments, which explains the good results.

Position 4 yields surprising results, overperforming positions 1-3 in 49, 43, and 47 out of 81 combinations, respectively. It is commonly advised for any measurement procedure where microphones are involved that the microphone should be far enough from reflective surfaces [31]. However, in our experiments, placing the phone on a horizontal surface shows better behavior than positions 1 and 2.

Table 1 is consistent with this analysis. It shows the mean value and coefficient of variation for all the accuracies shown in each Fig. 4 subfigure. Mean accuracy increases from position 1 to position 2, and from position 2 to position 3, while the coefficient of variation decreases. This explains the homogeneity perceived for position 3 where the "heat" seems more evenly distributed in the subfigure. Also, accuracy values are higher in general.

	Position			
	1	2	3	4
Average	0.701	0.718	0.729	0.731
CV	0.103	0.085	0.052	0.073

Table 1: Average accuracy and coefficient of variation for each subfigure in Fig. 4, training and testing with a single device.

Position 4 shows higher mean accuracy than position 3, and positions 1 and 2 consequently. However, the coefficient of variation is higher than that of position 3, showing a slight advantage for the latter, while it is still lower than the coefficients of positions 1 and 2. This places position 4 as the second best setup, very close to position 3. Given the sound source position relative to the microphone, the smartphone angle is $\alpha = \arcsin(20/30) \approx 42°$, close to that of position 2. The reason for this improvement is not clear: reflections on the surface and resonances should be accounted for, and, instinctively, one might expect a degradation in performance, but the data shows the opposite.

Every combination other than BQ Aquaris-BQ Aquaris should be affected by the simulation, with some undetermined side effects than might induce error into the system. However, looking at the main diagonal in Figs.4a-4d, where the training and testing recording/simulated device is the same, it is shown that the performance of all the systems is similar regardless its position, showing accuracies in the 0.73-0.82 range. This combination is always at least a 89% of the BQ Aquaris-BQ Aquaris-Position 1 combination, thus retaining most of the classifying ability.

Furthermore, the matched diagonal elements seem to yield better results than mismatched experiments. This is supported by the statistical analysis shown in Fig. 5 and Table 2. Mean accuracy for matched devices is almost even across all positions, in the vicinity of 0.77, whereas mismatched experiments lose between 11% and 7% depending on the position. Error bars shown in Fig. 5 underline the improving effect of matched over mismatched experiments. A hypothesis test has also been applied. The results reveal that in all positions there exist statistically significant differences in accuracies between matched and mismatched conditions, being the values for mismatched lower. All p-values were lower than 0.001.

Based on the matched-mismatched differences in accuracy, we proceeded to train the system under an MCT schema, testing its capacity to improve the system robustness. Fig. 6b shows results for the experiments carried out: For each position we train the classifier with a mixture of multiple recording devices and test their abilities against an individual device recordings. It is worth noting that for MCT the train/test split is stratified in both PD/healthy

Position	Type	N	Mean	Stand. Dev.	P-value
Position 1	Matched	9	0.778	0.025	<0.001
	Mismatched	72	0.692	0.071	
Position 2	Matched	9	0.777	0.022	<0.001
	Mismatched	72	0.710	0.061	
Position 3	Matched	9	0.768	0.018	<0.001
	Mismatched	72	0.724	0.037	
Position 4	Matched	9	0.779	0.017	<0.001
	Mismatched	72	0.725	0.053	

Table 2: Count, mean, and standard deviation for one-one comparison.

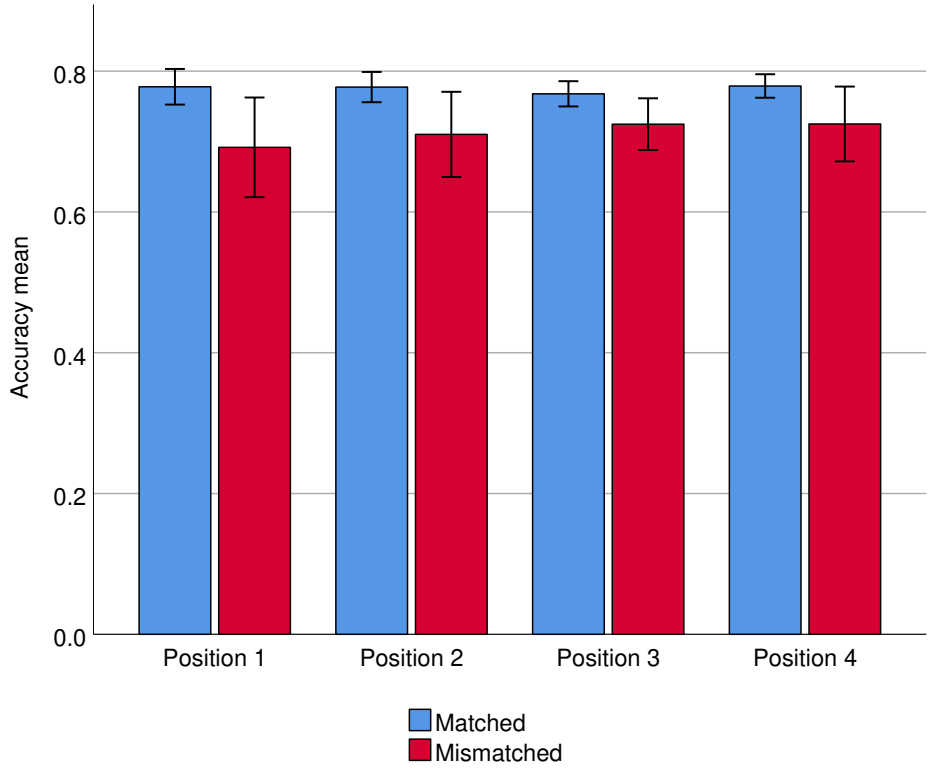


Figure 5: Classification accuracy obtained with every training-testing device combination for each of the device positions considered.

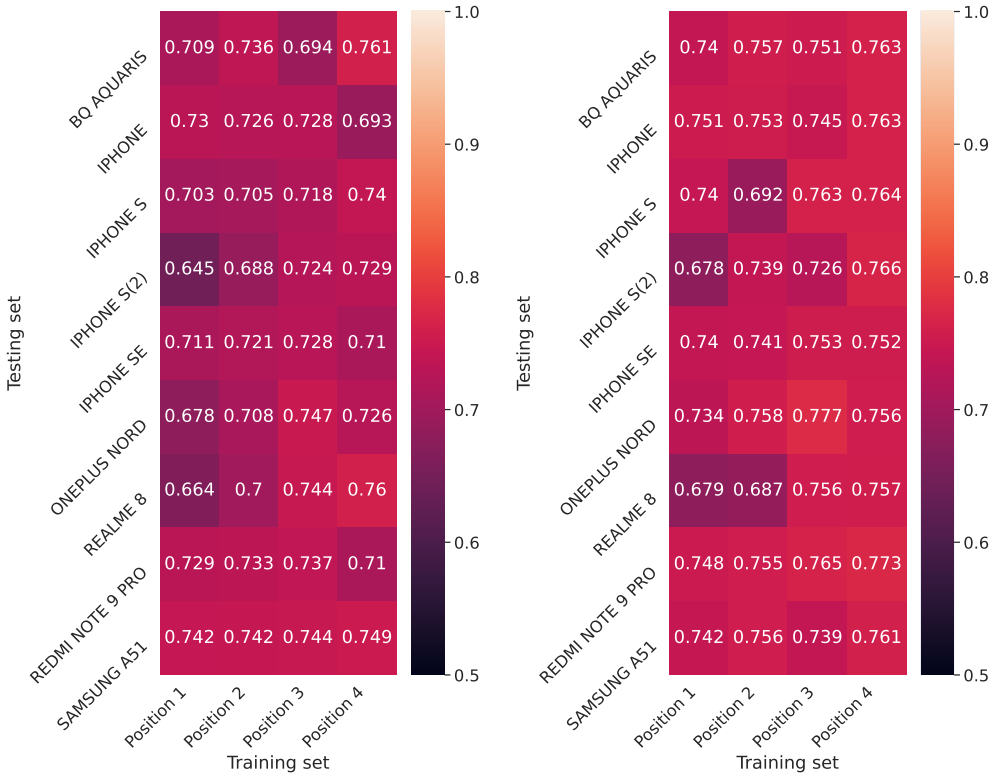
ratio and recording device prevalence, so differences in results can be attributed, like in single condition experiments, to the patients selected for each cross validation split.

For comparison purposes, we show in Fig. 6a the row-wise mean accuracy obtained in Figs.4a-4d. Each column summarizes results for single phones as test devices in a position. We can see that in this situation MCT improves the mean classifier performance for each position and for each recording device under test, since the results exceed the equivalent mean values for all the combinations. Exceptions are found in position 2-iPhone S(2), position 2-Realme 8, position 3-Samsung A51, and position 4-Realme 8 simulations. However, these exceptions in MCT barely underperform a 2% from the mean, whilst the mean improvement in accuracy due to MCT is about 4%, with peaks of 8%.

The fact that MCT gets better average results and lower error (Fig. 7) points out that MCT might contribute to build more robust systems. This is also backed by the statistical analysis results shown in Table 3: the difference between MCT and SC is more than one standard deviation apart and shows statistically significant differences (p-value < 0.05). This underscores the MCT usefulness to improve robustness. Position 2 is an exception to this, although its p-value is 0.058, very close to statistical significance.

Comparing position performance, both position 1 and position 2 seem more homogeneous with MCT (Fig. 6b) than they do without using it (Fig. 4a). Also, the growing trend of accuracies for positions 1, 2, and 3 appears to remain with MCT, and position 4 still rivals with position 3 results. This qualitative analysis is supported by Table 4: Compared with Table 1, averages obtained per position are higher with MCT in all cases, and coefficients of variation are lower as well. It is remarkable how, in this case, position 4 yields the best results, beating those obtained for position 3.

Finally, Table 5 shows the results obtained for an all to all MCT experiment (using a train set and a test set built with a mixture of all recording devices). The average accuracy values for positions 1-4 are consistent to the mean accuracy shown in Table 4 as should be expected: the low all to one CV values suggest that an all to all experiment should yield an average close to the mean average of the all to one experiments, which is the case. In fact, the difference



(a) Row-wise mean (mean per testing device) of the accuracies of each subfigure in Fig. 4.

(b) Mean accuracy for MCT in each position.

Figure 6: Comparison between mean accuracy obtained attending to position. Mean testing device accuracy versus MCT mean accuracy.

Position	Type	N	Mean	Stand. Dev.	P-value
Position 1	SC	9	0.701	0.033	0.040
	MCT	9	0.728	0.028	
Position 2	SC	9	0.718	0.018	0.058
	MCT	9	0.738	0.028	
Position 3	SC	9	0.729	0.017	0.006
	MCT	9	0.753	0.015	
Position 4	SC	9	0.731	0.024	0.004
	MCT	9	0.762	0.006	

Table 3: Count, mean, standard deviation, and p-value for MCT and SC.

	Position			
	1	2	3	4
Average	0.728	0.738	0.753	0.762
CV	0.039	0.038	0.020	0.008

Table 4: Average accuracy and coefficient of variation for each column in Fig. 6b, MCT with all the devices and testing with a single device.

of Table 5 values from those shown in Table 4 is lower than 0.01% in all cases, and specifically lower than 0.001% for positions 3 and 4. Those results lie within error in positions 3 and 4, which are consistently yielding the most stable

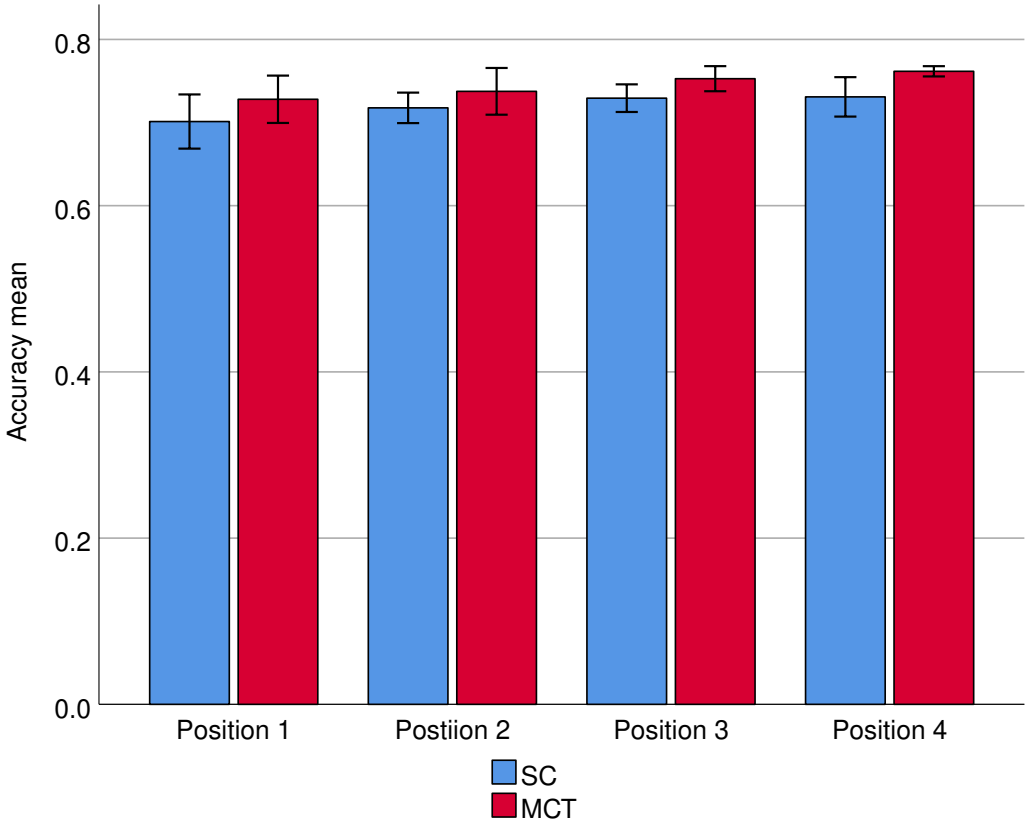


Figure 7: Classification accuracy obtained with every training-testing device combination for each of the device positions considered.

	Position			
	1	2	3	4
Average	0.724	0.734	0.754	0.762

Table 5: Mean accuracies for MCT, all devices in training and testing sets.

results.

For the sake of completeness, since in diagnostic tools sensitivity and specificity are important metrics, we also include their values as supplementary data. In the case of MCT, the mean sensitivity (specificity) values obtained after testing will all the devices, are 0.737 (0.711), 0.744 (0.724), 0.772 (0.737), 0.782 (0.743), for positions 1, 2 ,3 and 4, respectively. Therefore, the proposed MCT system is more sensitive than specific. In the case of a screening test for a disease, the role of sensitivity is more critical than that of specificity

Although the goal of this study is to analyze the effects of changing the recording device, accuracy obtained for the realistic scenario, where we use BQ Aquaris smartphone for both train and test phases, is on par with related literature. For example, [32] uses a variety of phonemes from PC-Gita and Viswanthan's databases for PD detection. Using \a\ phoneme it reaches an accuracy of 0.693 and 0.858, respectively. On its side, [33] uses Neurovoz, ItalianPVS and mPower databases with accuracies of 0.854, 0.990, and 0.754, respectively, always using \a\ phoneme.

There have been some efforts in studying the reliability of smartphones as a source of data for voice health assessment [34, 35]. They show that certain features are more affected than others, and consider smartphones as a valid recording device. They compare the error of an assorted set of smartphones performance against a studio microphone, but do not get into the automatic assessment phases.

Regarding selected features, there exists a high variability on experimental conditions: 9 single condition feature

selection experiments per position; 1 multicondition feature selection experiment per position in all to one configuration; and 1 in all to all configuration. However, an examination of the most selected features shows that, for each experiment, the 10 most used features are a subset of the following: Glottal to Noise Excitation, Lempel Ziv Complexity, Mel Frequency Cepstral Coefficients 3, 5, 6, 9, 10, 11, 13, Cepstral Peak Prominence, and First Zero in Correlation Function.

Those features are usually considered in scientific literature [12]. Some of them also stand out as reliable: in [33], experiments with different databases show that MFCCs are usually ranked among the most important features they considered. The prevalence of the aforementioned features across every experiment shows that, for PD, these might be the most robust ones among the features considered, which should be further investigated.

All of the experiments were designed having in mind that future health telemonitoring, specifically voice assessment, and particularly PD diagnosis and monitoring, will probably be linked to the development of smartphones and their capabilities. Many efforts have been already made, like mPower initiative [36], recruiting volunteers and recording their voices among other motor and cognitive tests for PD monitoring, or Parkinson Voice Initiative [37], collecting telephone-quality recordings from subjects from seven different countries. In the case of PD, [20] suggests the necessity of a detailed comparison of different microphones from different smartphones to complete the analysis of the full communication path in a hypothetical telemonitoring system.

This paper fills that gap. Results obtained in matched conditions show that most modern smartphones provide adequate recording systems for this particular application. Furthermore, the quality of the recording device is not nearly as important as a right setup for experimentation. It is worth stressing that in this paper we do not intend to recommend a specific recording device, but to underline the importance of training with a variety of sound sources. Environmental influence has been tested in previous work [27, 28]. The present paper complements them in channel description even though those studies revolve around voice conditions other than PD.

However, the results can be transposed to any other condition. The experiments test the influence of recording device and their positioning in the outcome of a statistical learning algorithm. The fact that we can compare positioning of the same device allows us to discard any other influential factor, since the experimental setup fully isolates the considered variables. The differences in simulation between two different positions given a smartphone, or between two smartphones having selected a position, can only be attributed to that change, as the anechoic chamber eliminates any noise source other than those inherent to the recording system, and the ones already present in the original recording.

Moreover, the present results can be further extended to any other telemonitoring setup. In this paper, we have considered .wav lossless voice recordings. Other efforts in telemedicine development use real time connections. To the authors' best knowledge, influence of other channels than that of cell phone networks have not been tested. However, there is a wide range of commercial voice over IP solutions, and it is a hot topic in communications development mostly due to current teleworking needs. All of these solutions will necessarily be placed after voice sampling, and therefore the recording setup would have an influence on them all, whether it is live or recorded.

4 Conclusions

We have studied the effects of smartphone selection and placement in the accuracy of an automatic detection aid system for PD based on voice recordings. Experimental results indicate that it is a good practice to test the system using recordings obtained with the same device used for testing. If we acknowledge the variability in recording devices used for a widespread technology, results may vary. Differences up to 37% were found when using other smartphone than that used for training.

We have also proposed a methodology to overcome the limitation in recording device selection by using MCT. This technique offers lower results dispersion with an increase in accuracy compared to the averaged results of single condition. However, further studies would be required to increase the statistical power of the results, involving a higher number of voice samples.

Also, we have found that recording device position relative to the speaker has a high impact on results. Holding the phone vertically right in front of the speaker yields the best results, and placing the phone atop a table is the second best option.

Acknowledgments

This research is part of R&D&I Projects PID2021-122209OB-C32 and MTM2017-86875-C3-2-R, funded by MCIN/AEI/10.13039/501100011033/; Grants GR21057 and GR21072, funded by Junta de Extremadura and the European Regional Development Fund (ERDF/FEDER); and Grant FPU18/03274 (Ministerio de Universidades).

References

[1] O.-B. Tysnes, A. Storstein, Epidemiology of Parkinson's disease, Journal of neural transmission 124 (8) (2017) 901–905. doi:10.1007/s00702-017-1686-y.

[2] Z. Ou, J. Pan, S. Tang, D. Duan, D. Yu, H. Nong, Z. Wang, Global trends in the incidence, prevalence, and years lived with disability of Parkinson's disease in 204 countries/territories from 1990 to 2019, Frontiers in public health (2021) 1994.

[3] S. L. Oh, Y. Hagiwara, U. Raghavendra, R. Yuvaraj, N. Arunkumar, M. Murugappan, U. R. Acharya, A deep learning approach for Parkinson's disease diagnosis from EEG signals, Neural Computing and Applications 32 (15) (2020) 10927–10933. doi:10.1007/s00521-018-3689-5.

[4] N. Amoroso, M. La Rocca, A. Monaco, R. Bellotti, S. Tangaro, Complex networks reveal early MRI markers of Parkinson's disease, Medical image analysis 48 (2018) 12–24. doi:10.1016/j.media.2018.05.004.

[5] M. Belić, V. Bobić, M. Badža, N. Šolaja, M. Đurić-Jovičić, V. S. Kostić, Artificial intelligence for assisting diagnostics and assessment of Parkinson's disease—a review, Clinical neurology and neurosurgery 184 (2019) 105442. doi:10.1016/j.clineuro.2019.105442.

[6] W. Pawlukowska, A. Szylińska, D. Kotlęga, I. Rotter, P. Nowacki, Differences between subjective and objective assessment of speech deficiency in Parkinson disease, Journal of Voice 32 (6) (2018) 715–722. doi:https://doi.org/10.1016/j.jvoice.2017.08.018.

[7] T. Zhang, Y. Zhang, H. Sun, H. Shan, Parkinson disease detection using energy direction features based on EMD from voice signal, Biocybernetics and Biomedical Engineering 41 (1) (2021) 127–141. doi:10.1016/j.bbe.2020.12.009.

[8] W. Rahman, S. Lee, M. S. Islam, V. N. Antony, H. Ratnu, M. R. Ali, A. A. Mamun, E. Wagner, S. Jensen-Roberts, E. Waddell, T. Myers, M. Pawlik, J. Soto, M. Coffey, A. Sarkar, R. Schneider, C. Tarolli, K. Lizarraga, J. Adams, M. A. Little, E. R. Dorsey, E. Hoque, Detecting Parkinson disease using a web-based speech task: Observational study, Journal of Medical Internet Research 23 (10) (2021) e26305. doi:10.2196/26305.

[9] D. Montaña, Y. Campos-Roca, C. 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 (2018) 89–97. doi:https://doi.org/10.1016/j.cmpb.2017.11.010.

[10] G. Solana-Lavalle, J.-C. Galán-Hernández, R. Rosas-Romero, Automatic Parkinson disease detection at early stages as a pre-diagnosis tool by using classifiers and a small set of vocal features, Biocybernetics and Biomedical Engineering 40 (1) (2020) 505–516. doi:10.1016/j.bbe.2020.01.003.

[11] M. Hireš, M. Gazda, P. Drotár, N. D. Pah, M. A. Motin, D. K. Kumar, Convolutional neural network ensemble for Parkinson's disease detection from voice recordings, Computers in Biology and Medicine (2021) 105021doi:10.1016/j.compbiomed.2021.105021.

[12] J. A. Gómez-García, L. Moro-Velázquez, J. I. Godino-Llorente, On the design of automatic voice condition analysis systems. Part I: Review of concepts and an insight to the state of the art, Biomedical Signal Processing and Control 51 (2019) 181–199. doi:10.1016/j.bspc.2018.12.024.

[13] J. A. Gómez-García, L. Moro-Velázquez, J. I. 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 (2019) 128–143. doi:10.1016/j.bspc.2018.09.003.

[14] A. Tsanas, Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning, Ph.D. thesis, University of Oxford UK, D. Phil. Thesis (2012).

[15] J. R. Orozco-Arroyave, J. C. Vásquez-Correa, P. Klumpp, P. A. Pérez-Toro, D. Escobar-Grisales, N. Roth, C. D. Ríos-Urrego, M. Strauss, H. A. Carvajal-Castaño, S. Bayerl, L. R. Castrillón-Osorio, T. Arias-Vergara, A. Künderle, F. O. López-Pabón, L. F. Parra-Gallego, B. Eskofier, L. F. Gómez-Gómez, M. Schuster, E. Nöth, Apkinson: the smartphone application for telemonitoring Parkinson's patients through speech, gait and hands movement, Neurodegenerative Disease Management 10 (3) (2020) 137–157. doi:10.2217/nmt-2019-0037.

[16] H. Yoon, N. Gaw, A novel multi-task linear mixed model for smartphone-based telemonitoring, Expert Systems with Applications 164 (2021) 113809. doi:10.1016/j.eswa.2020.113809.

[17] J. Li, L. Deng, R. Haeb-Umbach, Y. Gong, Robust automatic speech recognition: a bridge to practical applications, Academic Press (2015).

[18] W. M. Kouw, M. Loog, A review of domain adaptation without target labels, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (3) (2021) 766–785. doi:10.1109/TPAMI.2019.2945942.

[19] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, M. H. Chin, Ensuring fairness in machine learning to advance health equity, Annals of internal medicine 169 (12) (2018) 866–872. doi:10.7326/M18-1990.

[20] A. Tsanas, M. A. Little, L. O. Ramig, Remote assessment of Parkinson's disease symptom severity using the simulated cellular mobile telephone network, IEEE Access 9 (2021) 11024–11036. doi:10.1109/ACCESS.2021.3050524.

[21] D. Garcia-Romero, X. Zhou, C. Y. Espy-Wilson, Multicondition training of gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition, in: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2012, pp. 4257–4260. doi:10.1109/ICASSP.2012.6288859.

[22] L. Moro-Velazquez, J. A. Gomez-Garcia, J. I. Godino-Llorente, J. Villalba, J. Rusz, S. Shattuck-Hufnagel, N. 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 (2019) 205–220.

[23] M. Novotný, P. Dušek, I. Daly, E. Růžička, J. Rusz, Glottal source analysis of voice deficits in newly diagnosed drug-naïve patients with Parkinson's disease: correlation between acoustic speech characteristics and non-speech motor performance, Biomedical Signal Processing and Control 57 (2020) 101818.

[24] L. Naranjo, C. J. Perez, Y. Campos-Roca, J. Martin, Addressing voice recording replications for Parkinson's disease detection, Expert Systems with Applications 46 (2016) 286–292.

[25] A. J. Romann, B. C. Beber, C. A. Cielo, C. R. d. M. Rieder, Acoustic voice modifications in individuals with Parkinson disease submitted to deep brain stimulation, International archives of otorhinolaryngology 23 (2019) 203–208. doi:10.1055/s-0038-1675392.

[26] C. Pörschmann, J. M. Arend, Investigating phoneme-dependencies of spherical voice directivity patterns, The Journal of the Acoustical Society of America 149 (6) (2021) 4553–4564. doi:10.1121/10.0005401.

[27] M. Madruga, Y. Campos-Roca, C. J. Pérez, Impact of noise on the performance of automatic systems for vocal fold lesions detection, Biocybernetics and Biomedical Engineering 41 (3) (2021) 1039–1056. doi:10.1016/j.bbe.2021.07.001.

[28] M. Madruga, Y. Campos-Roca, C. J. Perez, Multicondition training for noise-robust detection of benign vocal fold lesions from recorded speech, IEEE Access 9 (2020) 1707–1722. doi:10.1109/ACCESS.2020.3046873.

[29] G. D. Ruxton, The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test, Behavioral Ecology 17 (4) (2006) 688–690. doi:10.1093/beheco/ark016.

[30] P. E. McKnight, J. Najab, Mann-Whitney U Test, John Wiley & Sons, Ltd, 2010, pp. 1–1. doi:10.1002/9780470479216.corpsy0524.

[31] Acoustics - Measurement of room acoustic parameters - Part 1: Performance spaces, Standard, International Organization for Standardization, Geneva, CH (Jun. 2009).

[32] N. D. Pah, M. A. Motin, D. K. Kumar, Phonemes based detection of parkinson's disease for telehealth applications, Scientific Reports 12 (1) (2022) 1–9.

[33] A. S. Ozbolt, L. Moro-Velazquez, I. Lina, A. A. Butala, N. Dehak, Things to consider when automatically detecting Parkinson's disease using the phonation of sustained vowels: Analysis of methodological issues, Applied Sciences 12 (3) (2022) 991.

[34] F. Schaeffler, S. Jannetts, J. M. Beck, Reliability of clinical voice parameters captured with smartphones–measurements of added noise and spectral tilt, in: Proceedings of the 20th Annual Conference of the International Speech Communication Association INTERSPEECH, Graz, Austria, 15-19 September 2019, ISCA, 2019.

[35] S. Jannetts, F. Schaeffler, J. Beck, S. Cowen, Assessing voice health using smartphones: bias and random error of acoustic voice parameters captured by different smartphone types, International journal of language & communication disorders 54 (2) (2019) 292–305.

[36] B. M. Bot, C. Suver, E. C. Neto, M. Kellen, A. Klein, C. Bare, M. Doerr, A. Pratap, J. Wilbanks, E. R. Dorsey, S. H. Friend, A. D. Trister, The mPower study, Parkinson disease mobile data collected using researchkit, Scientific data 3 (1) (2016) 1–9. doi:10.1038/sdata.2016.11.

[37] S. Arora, L. Baghai-Ravary, A. 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 145 (5) (2019) 2871–2884. doi:10.1121/1.5100272.

Chapter 7

Results and conclusion

This chapter summarizes the most significant contributions arising from the exposed research work. Also, a conclusion and a discussion on future research lines is included.

7.1 Summary of the results

This thesis represents a scientific advance in terms of robustness of automatic detection systems for diseases that affect voice production. Regardless of its final purpose, the work has covered the whole development cycle for machine learning models. As such, conclusions and results can be extracted for every step in the process. Concrete solutions for specific problems associated to the nature of the task in hand have been proposed. Also, different health conditions have been studied, covering organic vocal fold pathologies, such as nodules, polyps and Reinke's edema, and a neurological disease (PD), and similar challenges were raised for all of them. Domain adaptation and high bias emerges as a methodological problem that has been addressed by introducing multicondition training. The results presented have been obtained by following a rigorous methodology based on balanced datasets regarding age and sex in all the experiments in which the same approaches are applied on different databases in a comparative way; comparable disease stages between datasets have been considered where possible; also, cross validation is based on multiple iterations to reduce results variability. The following subsections address the most significant results.

7.1.1 Performance assessment of automatic voice evaluation systems for disease detection

Differences in recording conditions have a direct impact on the performance of the resulting machine learning model. These differences include, but are not limited to the following aspects: room, background noise, and recording device. Model generalization becomes a difficult task due to specialization in the specific recording conditions of the sample population. Alteration of a single factor can have a great impact. This behavior has been observed for every disease under consideration, and for each isolated noise source studied. Lack of qualified supervision in the data collection process is also an aspect that might critically impact performance.

Mismatched noise and recording device conditions

Mismatched noise conditions are challenging scenarios for the applications considered. This work has shown and quantified the degrading impact of additive noise on automatic systems based on acoustic features for detection of nodules and Reinke's edema. An aspect to highlight is that the study used real-world nonstationary noise specifically recorded for this purpose in a clinical environment, showing a lower impact than that observed when white noise was added, but still detrimental (Madruga et al., 2021a).

Healthcare is becoming mobile and ubiquitous based on a wide variety of available devices with the ability to collect data. Future exploitation of the full potential of automatic voice condition analysis systems requires considering mismatched recording device conditions. A novel contribution of this thesis is the assessment of the impact that smartphone model and position may have in PD detection aid systems based on speech. The investigation showed differences in detection accuracy up to 37% when using other smartphone model than that used for training. It also showed that recording device position relative to the speaker has a high impact on the results, being the best position holding the phone vertically right in front of the speaker (Madruga et al., 2023).

Tolerable noise levels

Not all the noises that have been induced in the communication channel are measurable. For example, smartphone simulation effects cannot be measured in terms of Signal-to-Noise ratio. However, the additive noises studied allow for an estimation of the maximum noise level allowed for a machine learning CAD model to be moderately good. In absence of mitigation strategies, the

minimum Signal-to-Noise ratio can be established at 30 dB, by allowing an accuracy degradation not higher than 10% if we consider the in-house database, real noise and both nodules and Reinke's edema pathologies (Madruga et al., 2021a).

Impact of non-controlled conditions

mPower PD database provides data sent by volunteers under a variety of realistic acoustic environments and without control. This means that each volunteer sent their own voice recordings and complementary information without any professional supervision. A decisive step in the field of automatic voice condition analysis for PD detection would be to obtain systems that are reliable in such real-life scenarios. On the other hand, an in-house database, collected using the same smartphone and room under trained supervision, was recorded. By application of the same methodology on the in-house and mPower databases and the realization of cross-database experiments it has been shown that the performance of a PD detection system decreases in the case of an unsupervised database and strongly drops in cross-database tests (Carrón et al., 2021).

7.1.2 Strategies to performance robustness improvement

Once the system performance is assessed and the problem is identified, strategies to improve robustness were defined, implemented and tested. The main results are summarized next.

Semi-controlled scenarios

The performance degradation observed using voice recordings from non-controlled scenarios allows to encourage the use of semi-controlled conditions, i.e., voices recorded at home by the patients themselves following a strict recording protocol and control of the information about patients by professional supervision.

The relevant clinical information is provided by the physicians. However, general practitioners (in the context of triage for diagnosis) or patient and caregivers (in a PD monitoring application...) should receive some initial training after which a test should be mandatory to ensure that the speech protocol is fully understood and that the user has some control on the acoustic environment regarding noise level. Within this framework, recordings would be submitted via smartphone from different environments. The semi-controlled scenarios have shown great potential to be useful in real clinical applications.

Multicondition training

Multicondition training has proven to be an effective technique to tackle the problems associated with restricted development populations. These problems lead to high bias since the variability shown is small compared to the target population. Moreover, research in this field has so far restricted the recording environment, further reducing the sampling variability.

The multicondition training strategies proposed in this work are based on asymmetric approaches, in which the development datasets equal the size of the original datasets. Symmetric approaches are affected by methodological concerns and the only purpose to test them within this work in comparison to asymmetric ones is to emphasize these issues (Madruga et al., 2021b).

The use of multicondition training strategies, bounded by the statistical limitations due to database size, has shown to be beneficial. Inclusion of multiple noise sources (Madruga et al., 2021b) or a variety of recording devices (Madruga et al., 2023) in the training phase leads to an improvement of robustness. Not only the individual performance for noise mismatched train-test sets is improved, but also dispersion is lower. Hence, global performance is improved, and reliability enhanced. These effects are shown by the fact that, by using multicondition training, the tolerable noise level for classifier training can be lowered to 20dB instead of the 30dB level estimated in Madruga et al. (2021a). It is also noteworthy that multicondition training improves model performance irrespective of the classifier used (Madruga et al., 2021b).

Replication strategy

Voice databases have been shown to be relatively small in this particular research field. This is likely to increase bias in the models. Intra-subject variability associated with the participants has been addressed. Imperfections in technology and the very biological variability produce replicated recordings for which results from signal analysis are not identical, even if they correspond to the same subject at a concrete moment, in the same acoustic environment, and using the same recording device.

Taking within-subject variability into account has also made it possible to propose strategies to obtain more stable results. Specifically, regularization techniques were proposed, and have proven to improve model performance by taking into consideration multiple experiments coming from the same individual, aggregating the results into a unique feature set.

Experimental results consistently show that the classifier should only see each participant once, either in train or test phase. This result, in addition to the regularization technique used per individual, suggests that the best way to lower model bias is taking multiple samples from each participant, apply regularization, and use these aggregations as unique inputs per patient. In an application context that considered Reinke's edema detection, the replication-based regularization methods proposed have proven to provide a more stable performance than independence-based methods (Naranjo et al., 2021b).

Feature robustness and gender dimension

Classifiers and variable selection methods play a very relevant role as a strategy to make the process more robust. A wide range of methods have been tested. For each disease and noise source studied, a full model was developed. This includes the feature selection phase for each scenario. Therefore, any feature that happens to be commonly selected, especially under noisy conditions, may be deemed as noise robust. Passive-aggressive method as a classifier and a variable selection approach based on relevance of the features have provided the best results in the multicondition training frameworks. The following features have been the most relevant in noisy scenarios, regardless of the noise origin: Permutation entropy (PERMUTATION) stands as a good predictor for nodules, Cepstral Peak Prominence (CPP) for polyps and Glottal Noise Excitation (GNE) for Reinke's edema (Madruga et al., 2021b).

Gender perspective in this research is relevant. There is a gender imbalance in organic voice diseases and in PD prevalence. Vocal apparatus disorders affect women more than men, whereas PD affects men more than women. This has been considered in the experiments by using different sample sizes and including a gender variable in the machine learning algorithms, although part of this information is underlying in the extracted features.

Novel publicly available electronic health record dataset

In the scientific literature, there is a lack of datasets related to this context. This work also contributed to the publication of open access electronic health record datasets of acoustic features extracted from healthy, nodules, polyps and Reinke's edema effected voices. For noise multicondition training, the data can be downloaded from Madruga et al. (2021c,d), for analyzing data coming from controlled voice recordings, see Carrón et al. (2023), and for voice recordings from non-controlled situations see Bot et al. (2016); finally, for a replication-based strategy, data can be found in Naranjo et al. (2021b).

7.2 Conclusion and further research

The proposed CAD systems have a great potential to assist diagnosis and improve patient monitoring of many detectable-by-voice diseases. The procedure is non-invasive, low cost, and potentially applicable remotely. It can help general practitioners to conduct triage and help in diagnosis and tracking the disease. The detectable-by-voice diseases can be highly benefited by smartphone-

based systems, due to different aspects such as increasing incidence, diagnosis prone to errors, continuous monitoring, and remote control.

To ensure that a model is achieving its intended purpose, it's necessary to consider, optimize and check its performance in terms of robustness. The results obtained show the viability of CAD techniques for clinical diagnostic aid, enhancing robustness by overcoming some limitations present in the previous state-of-the-art. These limitations are related to at least the following aspects: mismatched conditions with respect to the communication channel (noise and device); methodological issues, such as the artificial increase of the sample size by using replicated recordings as if they were independent; unbalanced datasets; and other methodological weaknesses present in previous approaches. We can conclude that robustness of the proposed techniques is superior to that of the state-of-the-art approaches available at the publication time.

Nevertheless, despite the methodological and empirical advances provided in this PhD thesis, there is further research that can be conducted to bring these systems closer to real world clinical practice. Further studies with other vocal tests are interesting to carry out. This PhD thesis revolves around a single test: sustained /a/ sound. This is not the only option. It has the advantage of universality, since it is a common sound on most languages. However, it focuses on the vocal folds control. Other tests focus on other speech features. One example is diadochokinetic test for PD. In this regard, one research paper showing the viability of the proposed techniques for these tests is already under peer review (M. Madruga, Y. Campos-Roca, Carlos J. Pérez. *Enhancing robustness of automatic PD detection from diadochokinesis tests through the use of noise multicondition training*).

In that paper, speech features extracted from diadochokinetic tests have been used to measure articulatory aspects of speech impairment in PD and the migration from lab conditions to real world settings is faced by taking into account the corruption of speech by environmental noise. A multicondition training approach is considered for PD detection. Firstly, the experiments considered single-condition (clean and specific noises) training. In a second step, noise-based multicondition training is applied and the test is performed under different noisy conditions. The mean accuracy rates show relevant improvement percentages. To the authors' best knowledge, this is the first strategy addressed in the scientific literature to deal with the potential corruption of speech by environmental noise in the development of automatic PD detection systems from diadochokinetic tests.

Database size plays an important role in training a new model. The methods proposed help to overcome the limitations associated with low variability of recording conditions. This is the case for the databases used for this research work, which are collected in a controlled and repeatable environment. In this case, the noise was simulated in various stages of the communication channel. A better solution for real world medical applications would be to develop models using heterogeneous and large databases. This would increase the recording conditions variability, and would also yield more robust models, since multicondition training would be embedded in the training dataset.

The feature extraction algorithms and machine learning methodology developed in this PhD thesis are directly applicable for moderate size databases. However, if the database size is large enough, deep learning algorithms could be used. Deep learning faces the same problems addressed by this research work. Researchers already use similar techniques to avoid bias due to dataset size, namely data augmentation. However, results obtained for replication strategies suggest that in this case, intra-subject variability should also be taken into account. The integration of multicondition training in a well developed deep learning strategy for voice recordings could lead to even better results. Therefore, further research is encouraged.

In this work the models are trained to discriminate healthy from pathological voices by taking into account only one disease. It is a common issue for this research field. Multi-class classification, where a machine learning model is able not only to differentiate between healthy and pathological voices, but also detect the specific disease that might affect each individual, has not been properly addressed. This would allow CAD tools to move up from screening to diagnosis. In the case of PD, the different Hoern and Yahr stages are of interest, whereas for organic voice diseases, differentiating from vocal fold nodules, Reinke's edema or other related voice disorders is of great

help for triage or for diagnosis aid. The development of Bayesian hierarchical additive generalized models properly addresses nonlinearity and may provide good results.

Another related problem is monitoring the progression of voice-detectable diseases. This is highly relevant so that patients can receive continuous monitoring of the disease progression, since the voice tests can be remotely applied. This would allow customizing the dose of medication and its administration according to the evolution of patient symptoms. To address this problem, longitudinal models are needed. Specifically, longitudinal additive generalized models are flexible enough to handle this kind of data. The real problem is to recruit patients for the long term. This requires a long and stable collaboration with medical institutions or patients' associations. In this way, an agreement has been signed with the Spanish Federation of PD, which gives access to a great number of PD subjects that may compose a sample of sufficient size for future experiments. The collaboration with the Otorhinolaryngology Unit of the *Hospital San Pedro de Alcántara* can be reactivated any time.

Finally, in a longer term, the considered features can be combined with others such as eye-related measurements, gait or tapping to produce a multimodal approach for PD. The features should be able to be automatically collected by an easy-to-use mobile app, so the data generation process is straightforward and no other specific devices are needed. The rest of the proposed and future methodology would keep the same.

References

Amoroso, N., La Rocca, M., Monaco, A., Bellotti, R., and Tangaro, S. (2018). Complex networks reveal early MRI markers of Parkinson's disease. *Medical image analysis*, 48:12–24.

Barry, W. and Pützer, M. (2016). Saarbrücken voice database.

Belić, M., Bobić, V., Badža, M., Šolaja, N., Đurić Jovičić, M., and Kostić, V. S. (2019). Artificial intelligence for assisting diagnostics and assessment of Parkinson's disease. A review. *Clinical neurology and neurosurgery*, 184:105442.

Bot, B. M., Suver, C., Neto, E. C., Kellen, M., Klein, A., Bare, C., Doerr, M., Pratap, A., Wilbanks, J., Dorsey, E. R., Friend, S. H., and Trister, A. D. (2016). The mPower study, Parkinson disease mobile data collected using researchkit. *Scientific data*, 3(1):1–9.

Carrón, J., Campos-Roca, Y., Madruga, M., and Perez, C. J. (2023). Dataset for acoustic features extracted from voice recordings of people suffering Parkinson's disease and healthy subjects. *Mendeley Data*.

Carrón, J., Campos-Roca, Y., Madruga, M., and Pérez, C. J. (2021). A mobile-assisted voice condition analysis system for Parkinson's disease: Assessment of usability conditions. *BioMedical Engineering OnLine*, 20(1):1–24.

Echternach, M., Döllinger, M., Köberlein, M., Kuranova, L., Gellrich, D., and Kainz, M. (2020). Vocal fold oscillation pattern changes related to loudness in patients with vocal fold mass lesions. *Journal of Otolaryngology-Head & Neck Surgery*, 49(1):1–9.

Garcia-Romero, D., Zhou, X., and Espy-Wilson, C. Y. (2012). Multicondition training of gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing*, page 4257–4260. IEEE.

Godino-Llorente, J. I., Osma-Ruiz, V., Sáenz-Lechón, N., Cobeta-Marco, I., González-Herranz, R., and Ramírez-Calvo, C. (2008). Acoustic analysis of voice using WPCVox: A comparative study with Multi Dimensional Voice Program. *European Archives of Oto-Rhino-Laryngology*, 265(4):465–476.

Gómez-García, J., Moro-Velázquez, L., Arias-Londoño, J., and Godino-Llorente, J. (2021). On the design of automatic voice condition analysis systems. Part III: Review of acoustic modelling strategies. *Biomedical Signal Processing and Control*, 66:102049.

Gómez-García, J. A., Moro-Velázquez, L., and Godino-Llorente, J. I. (2019a). On the design of automatic voice condition analysis systems. Part I: Review of concepts and an insight to the state of the art. *Biomedical Signal Processing and Control*, 51:181–199.

Gómez-García, J. A., Moro-Velázquez, L., and Godino-Llorente, J. I. (2019b). 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.

Hantzakos, A., Remacle, M., Dikkers, F., Degols, J.-C., Delos, M., Friedrich, G., Giovanni, A., and Rasmussen, N. (2009). Exudative lesions of Reinke's space: A terminology proposal. *European Archives of Oto-rhino-laryngology*, 266(6):869–878.

Hegde, S., Shetty, S., Rai, S., and Dodderi, T. (2019). A survey on machine learning approaches for automatic detection of voice disorders. *Journal of Voice*, 33(6):947.e11–947.e33.

Kouw, W. M. and Loog, M. (2021). A review of domain adaptation without target labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):766–785.

Madruga, M., Campos-Roca, Y., and Pérez, C. J. (2021a). Impact of noise on the performance of automatic systems for vocal fold lesions detection. *Biocybernetics and Biomedical Engineering*, 41(3):1039–1056.

Madruga, M., Campos-Roca, Y., and Pérez, C. J. (2021b). Multicondition training for noise-robust detection of benign vocal fold lesions from recorded speech. *IEEE Access*, 9:1707–1722.

Madruga, M., Campos-Roca, Y., and Pérez, C. J. (2021c). Dataset for acoustic features extracted from healthy, nodules and Reinke's edema affected voices. *Mendeley Data*.

Madruga, M., Campos-Roca, Y., and Pérez, C. J. (2021d). Dataset for acoustic features extracted from healthy, nodules, polyps and Reinke's edema effected voices for noise multicondition training. *Mendeley Data*.

Madruga, M., Campos-Roca, Y., and Pérez, C. J. (2023). Addressing smartphone mismatch in Parkinson's disease detection aid systems based on speech. *Biomedical Signal Processing and Control*, 80:104281.

Massachusetts Eye & Ear Infirmary (1994). Voice disorders database, version. 1.03 (cd-rom). *Lincoln Park, NJ: Kay Elemetrics Corporation*.

Montaña, D., Campos-Roca, Y., and Pérez, C. J. (2018). 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.

Naranjo, L., Perez, C. J., Campos-Roca, Y., and Madruga, M. (2021a). Replication-based regularization approaches to diagnose Reinke's edema by using voice recordings. *Artificial Intelligence in Medicine*, 120:102162.

Naranjo, L., Perez, C. J., Campos-Roca, Y., and Madruga, M. (2021b). Reinke's edema diagnosis with replicated acoustic features. *Mendeley Data*.

Oh, S. L., Hagiwara, Y., Raghavendra, U., Yuvaraj, R., Arunkumar, N., Murugappan, M., and Acharya, U. R. (2020). A deep learning approach for Parkinson's disease diagnosis from EEG signals. *Neural Computing and Applications*, 32(15):10927–10933.

Pawlukowska, W., Szylińska, A., Kotlęga, D., Rotter, I., and Nowacki, P. (2018). Differences between subjective and objective assessment of speech deficiency in Parkinson disease. *Journal of Voice*, 32(6):715–722.

Rahman, W., Lee, S., Islam, M. S., Antony, V. N., Ratnu, H., Ali, M. R., Mamun, A. A., Wagner, E., Jensen-Roberts, S., Waddell, E., Myers, T., Pawlik, M., Soto, J., Coffey, M., Sarkar, A., Schneider, R., Tarolli, C., Lizarraga, K., Adams, J., Little, M. A., Dorsey, E. R., and Hoque, E. (2021). Detecting Parkinson disease using a web-based speech task: Observational study. *Journal of Medical Internet Research*, 23(10):e26305.

Shehab, M., Abualigah, L., Shambour, Q., Abu-Hashem, M. A., Shambour, M. K. Y., Alsalibi, A. I., and Gandomi, A. H. (2022). Machine learning in medical applications: A review of state-of-the-art methods. *Computers in Biology and Medicine*, 145:105458.

Tysnes, O. and Storstein, A. (2017). Epidemiology of Parkinson's disease. *Journal of neural transmission*, 124(8):901–905.

Zhang, T., Zhang, Y., Sun, H., and Shan, H. (2021). Parkinson disease detection using energy direction features based on EMD from voice signal. *Biocybernetics and Biomedical Engineering*, 41(1):127–141.