On the use of electroglottography and speech signals for automatic classification of patients with voice pathologies

Néstor Rafael Calvo Ariza

Tesis de maestría presentada para optar al título de Magíster en Ingeniería de Telecomunicaciones

Director
Juan Rafael Orozco Arroyave, Doctor (PhD) en Procesamiento de Señales
Asesor
Tomas Arias Vergara, Doctor (PhD) en Ciencias de la computación

Universidad de Antioquia
Facultad de Ingeniería
Maestría en Ingeniería de Telecomunicaciones
Medellín, Antioquia, Colombia
2024

Cita	Calvo Ariza [1]
Referencia Estilo IEEE (2020)	[1] N. R Calvo Ariza, “On the use of electroglottography and speech signals for automatic classification of patients with voice pathologies”, Tesis de maestría, Maestría en Ingeniería de Telecomunicaciones, Universidad de Antioquia, Medellín, Antioquia, Colombia, 2024.

Maestría en Ingeniería de Telecomunicaciones, Cohorte XVIII.

Grupo de Investigación en Telecomunicaciones Aplicadas (GITA).

Centro de Investigación Ambientales y de Ingeniería (CIA).

On the use of electroglottography and speech signals for automatic classification of patients with voice pathologies

UNIVERSIDAD
DE ANTIOQUIA
1 8 0 3

Master's Thesis in Telecommunication Engineering

Nestor Rafael Calvo Ariza

Director: Prof. Dr.-Ing. Juan Rafael Orozco Arroyave
Advisor: Dr. Tomas Arias Vergara

Faculty of Engineering

Department of Electronics and Telecommunications

University of Antioquia.

Acknowledgments

I want to thank my parents, Nestor Calvo and Mireya Ariza, for all the support and love I have received from them while completing this work. They are my motivation and my role models. I also want to thank my younger brother, Andres Felipe, who has supported me in many difficult moments and motivated me to keep growing and be an example for him, as well as my older siblings, Nestor Antonio and Lauren, who have always been watching my progress and have given me valuable life advice. I also want to thank my close family, who have always been there for me; I hope to continue being a source of pride for them. Additionally, I want to thank my girlfriend, Yina, who has given me unconditional support and love in difficult times and encouraged me to keep going. I love you.

Now, I would like to thank my colleagues Cristian Rios, Daniel Escobar, Santiago Moreno, Jeferson Gallo, Diego Lopez, Jaime Vergara, Fredy Mercado, and Christian Garzón, as well as the other colleagues of the GITA lab with whom I have shared many years. They have been people with whom I have been able to have very enriching conversations and who have helped me in one way or another during this process.

Last but not least, I would like to thank my director, Prof. Dr.-Ing. Juan Rafael Orozco Arroyave and my advisor, Dr. Tomas Arias Vergara, for all the support and knowledge you provided during this process. Thank you for your experiences, patience, and willingness to discuss and review my work.

I also want to thank the University of Antioquia for the financial support I received while developing this master's thesis through the CODI project number 2023-58010.

Abstract

Voice production is a crucial aspect of human life; problems with the voice can affect the quality of life by influencing how we communicate. Speech production involves various muscles and neural connections so that voice pathologies can arise from multiple sources. Early detection of these disorders is critical to maintaining or improving the patient's condition. However, diagnosing these pathologies is often time-consuming and subject to physicians' assessment. The increased popularity of Artificial Intelligence (AI) has led to the creation of machine learning and deep learning models that perform an automatic analysis based on patterns found in the data. These AI techniques offer the potential to simplify the diagnostic process, providing more consistent and objective assessments. However, they require a previous analysis of the data, the features that will be extracted, and the classifiers. This work analyzes and compares multiple techniques to classify voice pathologies using the Saarbrücken Voice Database, a German database containing multiple voice pathologies and healthy controls performing different tasks. This work aims to apply and compare different machine learning and deep learning techniques to find the best classifier, considering an unbiased analysis for age and gender. Also, this work aims to showcase the capabilities of a novel feature set called phase plots, which represented glottal cycles as elliptical trajectories superimposed in a 2D plane. Additionally, the study explores the impact of incorporating complementary information through early and late fusion methods on the classification process. By integrating these techniques, the study aims to enhance the accuracy and robustness of voice pathology classification. The findings of this work highlight the potential of automated techniques in voice pathology detection.

Contents

Chapter 1

Introduction

1.1 Motivation

Nowadays, one-third of the labor force uses their voices in daily routine [1]. Most voice pathologies result from wrong vocal use, such as vocal hygiene, laryngeal infection, and vocal fatigue. However, due to the complexity and the number of parts of our body involved in speech production, identifying the causes and detecting voice pathologies can be challenging. Additionally, some voice pathologies can be caused by neurological problems, which increases the difficulty in the detection process. These pathologies often require specialized physicians and a multidisciplinary approach [2].

Machine learning advances have helped researchers create technology capable of extracting patterns from different bio-signals. These patterns are usually analyzed by physicians or automatic models that classify if the subject has "features" of a person with a voice pathology, generating early reports. Early detection of the pathology would allow early treatment. In the long run, these early classifications will be a tool that physicians can use in conjunction with their medical analysis to assess a patient. Afterward, these tools can be integrated into a cellphone or a PC so the patient can be continuously monitored.

Research has found multiple combinations of features, classifiers, bio-signals, and techniques to achieve this classification. The main focus of this work is to test different strategies for the classification and compare the methods tested in the German Saarbrücken Voice Database (SVD), which contains voice and electroglottography (EGG) recordings of subjects with different voice disorders.

This work employs multiple feature extraction techniques to extract relevant information from the electroglottography and speech signals of patients with voice disorders and healthy subjects. It compares the performance of this task alongside different classifiers to identify the combinations that yield the best results in the classification task. Additionally, a new set of features called phase plots is tested and compared with the more traditional feature sets. Lastly, a multimodal approach is explored to discover whether combining the information from both signals can be advantageous for the classification problem.

1.2 State of the art

Many works have been dedicated to exploring techniques and architectures that allow the use of speech to perform pathology analysis. The majority of these works use 3 datasets: Massachusetts Eye and Ear Infirmary Database (MEEI) [3], Arabic Voice Pathology Database (AVPD) [4], and Saarbruecken Voice Database [5]. Because this work only uses SVD, only the works that used SVD at least once are considered in this brief review.

Works will be categorized by their samples from the database, the pathologies used, the extracted features, the classification methods used, and the results obtained. Finally, the SVD contains speech and electroglotography signals; this opens a window for analyzing "what happens when we combine both signals." So fusion, if there is any, will be taken into account.

Lastly, the analysis of voice pathologies is usually carried out from three research branches. The most common one is to analyze the problem in a more general approach, considering all the pathologies and aiming to differentiate between a healthy subject (HC) and a subject with a voice pathology (VP) [6], [7]. Another approach is to assess the voice quality of subjects with a voice pathology using a scale like the Dysphonia Severity Index (DSI) [8] or the Grade of dysphony, Roughness, Breathiness, Asthenicity, and Strainess scale (GRBAS) [9]. The last focus is pathologies, either narrowing down the first approach by selecting one pathology [8] or focusing on differentiating between two or more pathologies [10]. This work focuses exclusively on the first classification approach.

There are many studies about classifying between HC and VP subjects. In [11], the authors used both the SVD and the MEEI datasets. They extracted Mel-Frequency Cepstral Coefficients (MFCC) and noise-related fea-

tures such as Harmonic-to-Noise Ratio (HNR), Normalized Noise Energy (NNE), and Glottal-to-Noise Excitation Ratio (GNE). Gaussian Mixture Models (GMM) were used for the classification process, and metrics such as Accuracy (ACC), Receiver Operating Characteristic curve (ROC), Sensitivity (SEN), and Specificity (SPE) were reported. All three vowels and four intonations were used in the SVD, creating 12 subsets. The authors mentioned that they could not guarantee that a subject is in the same subset for all 12 combinations, and they comprehend that this can generate "optimistic results". 30 folds were created; 29 were used for training the classifier, and the remaining was used for testing. Results showed accuracies of more than 70% for the classification task using the vowels, and the best result (87.9%) is obtained when the three (3) vowels are combined; this is also one of the first works that analyze task combination techniques with the SVD.

In [12], the authors used the vowel /a/ as their selected task for the analysis. 50 HC and 70 VP subjects were selected from the database; 24 had Chronic laryngitis, 6 had Cysts, 19 had Reinke edemata, and 21 had Spasmodic dysphonia. Articulation features were extracted, specifically 13 MFCC, the first and second derivatives, using a Hamming window of 30 ms with an overlap of 15 ms. The authors also used Linear Discriminant Analysis (LDA) to reduce the dimensionality of the features and an artificial neural network for the classification. An accuracy of 75.13% was reported for the MFCC and 87.8% when combining the MFCC with both the first and second derivative and the features transformed by LDA.

Another work that used classical features such as MFCC is [13]; it includes features like Fundamental Frequency (F0), Jitter, and Shimmer, which can be grouped as phonation features together with HNR. The authors followed a 10-fold cross-validation strategy to train different classifiers with features extracted from sustained vowel /a/ recordings. Support Vector Machine (SVM), Decision Tree (DT), Bayesian Classifier (BC), and Logistic Model Tree (LMT) were used. The authors also performed a feature selection using a correlation method to find the features with a high correlation to the classes and an information gain method that assesses which features give more information. SVM and DT yielded the best results with accuracies of 85.7% and 83.6%, respectively.

With the increasing popularity of neural networks and deep learning methods, multiple works adopted different architectures to avoid the feature extraction method. For instance, [14] proposes an automatic voice pathol-

ogy detector using deep neural networks to analyze the sustained vowel /a/. In this work, the authors used a Long-Short-Term-Memory (LSTM) with a Convolutional Neural Network (CNN) responsible for receiving the audio signal, previously segmented into 64 ms windows with 30 ms overlapping. The convolutional layers perform automatic feature extraction, and these extracted features are the input of the LSTM, which is in charge of the final classification. The authors reported an accuracy of 68.1% when testing using 960 samples for training, 206 for validation, and 874 for testing.

In [15], the authors presented a framework for continuously evaluating a patient's condition based on speech signals. They used two databases for the model's training: the MEEI and the SVD. The MEEI database contains samples of the sustained vowel /a/. The signal is divided into 40 ms frames, with a 20 ms overlapping. The Fourier transform is applied to extract the spectrograms that are subsequently analyzed by two convolutional network architectures: the VGG16 network and the CaffeNet. Three experiments are reported. The first is training a model with the MEEI database, which is tested using the SVD database; the second experiment trains the model with the SVD and tests it with the MEEI.

The authors in [16] presents a deep learning model that seeks to detect and reconstruct dysarthria in speech; this work used neural networks but also attacks the problem that neural networks are often considered black boxes. This problem was solved by analyzing the characteristics encoded by the neural network and reconstructing the signal with a decoder, showing that the network can encode features that may be interpretable. This work proposed an architecture composed of a Recurrent Convolutional Neural Network model (RCNN); the output of this network passes through a dense bottleneck layer, considerably reducing the features. These bottleneck layers are usually very common among autoencoders. To evaluate the model, they use a Leave-One-Subject-Out (LOSO) cross-validation scheme and report an accuracy of 92.9%. This work provides insight into a new architecture strategy, which, in turn, allows for internal knowledge of the characteristics and what is being observed by the network.

In [17], classical features such as MFCCs and Linear Prediction Cepstrum Coefficients (LPCCs) were extracted from spectrograms with a 40-ms window and a 20-ms frame shift. Additionally, the authors used Higher-Order Statistics (HOS) to identify speech impairment. Specifically, the 3rd and 4th statistics were selected, named normalized skewness and normalized kurto-

sis. The authors utilized a Feed-Forward Network (FFN) and a combination of a CNN and a fully connected network for the classification process. For both classifiers, the authors used the features mentioned previously as their input layer. The vowels /a/, /i/, and /u/ were considered, and the authors performed experiments focusing on gender to address the imbalances in the SVD database. An accuracy of 82.7% was reported, using the LPCC features but only with male subjects. The highest accuracy considering both genders was 76.6% and was obtained by extracting the MFCCs from the vowel /a/ and using the CNN with the fully connected classifier.

A work using EGG and audio signals to detect voice pathology is [18]. This study employs a cloud framework similar to [15] to group and analyze different signals to make pathology detection. In this work, the authors used a GMM, which receives both the audio and the EGG signal. Traditional features like jitter and shimmer, among others, are extracted from the audio signal. For the case of the EGG signal, commonly used features such as the quotients, peak-related features, and some cepstral features are used. The GMM was evaluated using the SVD database. The results show a model that achieves an accuracy of 92.8% using just the speech, 77.7% using only the EGG signal, and 94.2% using both signals. These findings showcase that combining both signals can yield good results for the voice pathology detection task.

In [19], the authors extracted traditional features like MFCCs and LPCCs, similar to [17]. The authors used the whole SVD database. To solve the imbalances between classes, the authors proposed an oversampling method called the Synthetic Minority Oversampling Technique (SMOTE); this oversampling method allowed the authors to create artificial samples to balance the database between HC and VP. The authors performed a similar classification process to the one used in [17]. The highest accuracy reported was 98.8% using a combination of MFCC and LPCC features with the oversampled data.

Finally, both [20] and [21] are very recent works that propose an automatic classification of voice pathologies using a combination of the signals. In [20], the authors use a CNN combined with a fully connected layer; for this work, the authors stacked the three vowels using just the normal pitch and the ones with different pitches. The vowels are introduced as raw data to the model. The authors reported different metrics like precision, recall, and F1-score. The model that yielded the best result was the stacked vowels with a normal

pitch with an 80% F1-score.

In the case of [21], the authors use the features extracted with the CNN and perform an early fusion to combine the handcrafted features. This new set of features is then used in an SVM classifier to classify healthy and pathological. The highest accuracy was 90.1% and was obtained when the handcrafted features were combined with those obtained from the CNN.

The analysis of voice pathologies has gained increasing attention over the past decade. Figure 1.1 illustrates the trend in the number of research works focused on voice pathology detection over the last ten years. The figure also highlights the subset of studies that utilized the SVD as one of their primary datasets. Additionally, Figure 1.2 and Figure 1.3 present a comparison of the most commonly used features and classifiers for voice pathology detection over the years.

In terms of feature extraction, MFCCs are among the most frequently used features. Moreover, using CNNs has become increasingly popular, often in combination with MFCCs, to compare results obtained from automatic feature extraction against more traditional approaches. It is important to note that these CNNs are typically trained on spectrogram images derived from the audio signals.

On the classification side, SVMs are frequently the classifier of choice for voice pathology detection. This is due to their robustness and proven effectiveness in such classification tasks. Another commonly used classifier is the Artificial Neural Network (ANN), which, when combined with CNNs, can directly process raw audio signals and perform classification without the need for explicit feature extraction. However, the adoption of these deep learning methods introduces challenges related to the interpretability of the results [22].

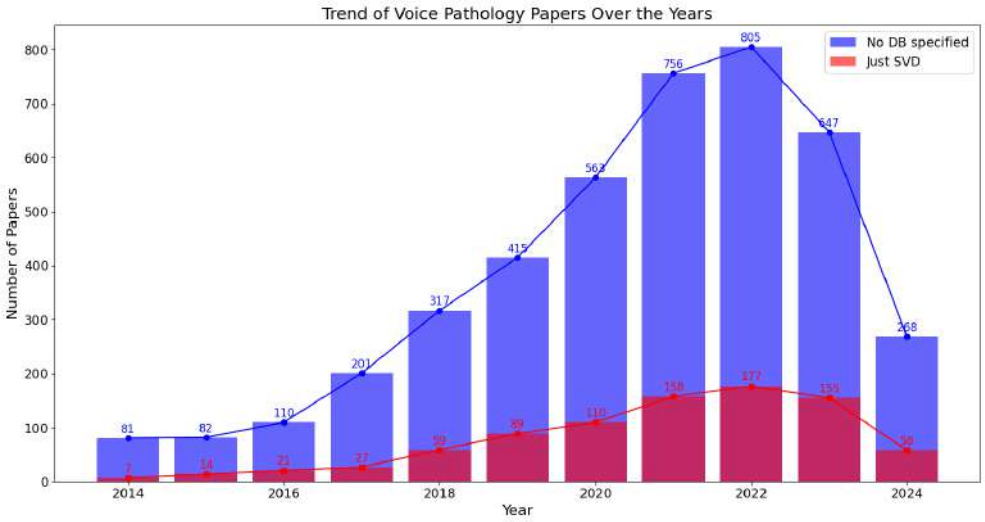


Figure 1.1. Voice pathology papers from 2014 to 2024

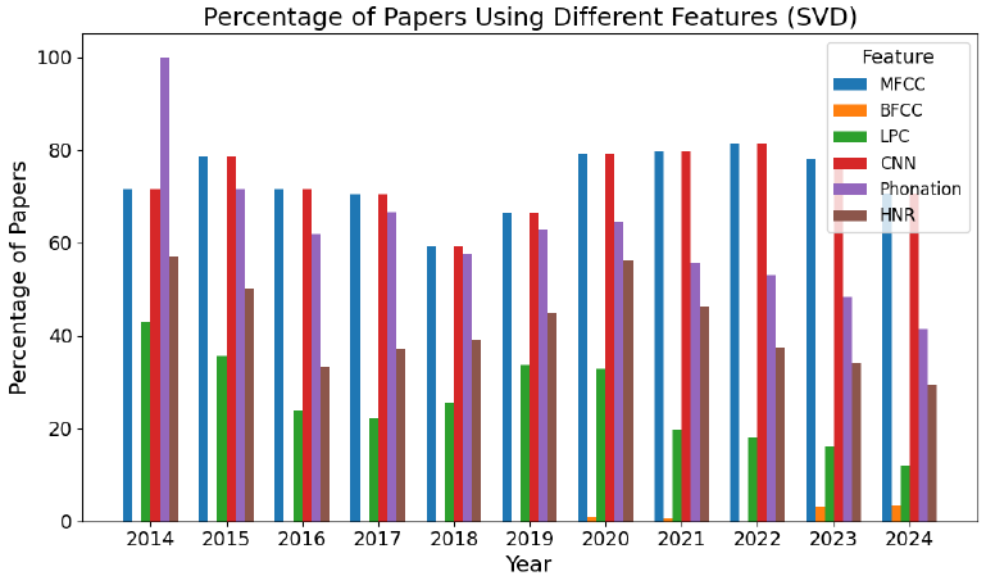


Figure 1.2. Most common features used in voice pathology detection

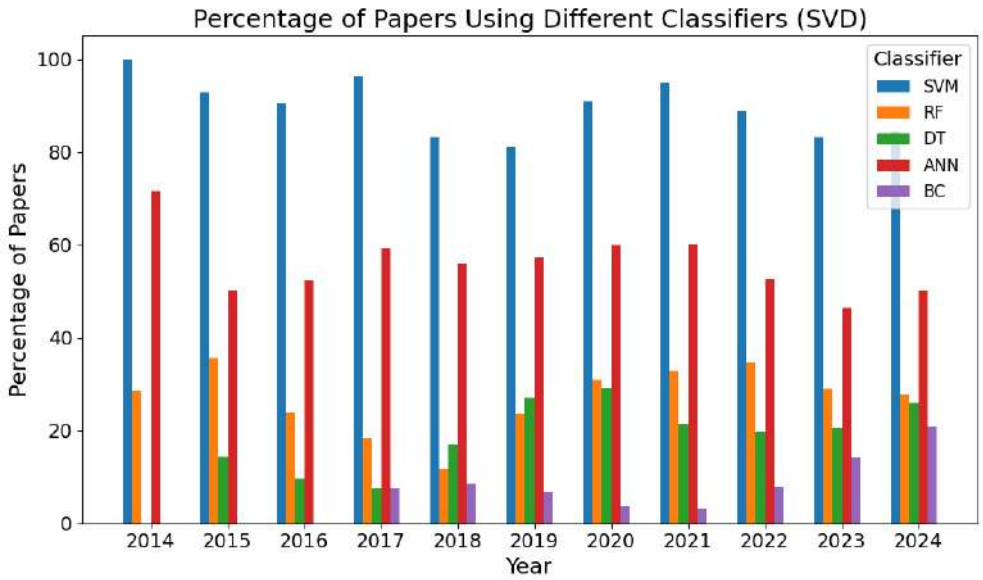


Figure 1.3. Most common classifiers used in voice pathology detection

Several studies in the state-of-the-art review have demonstrated promising results in detecting voice pathologies. However, a deeper analysis reveals several critical considerations. One major issue is the imbalances in age and gender within the SVD database. Some studies address these imbalances by applying oversampling techniques [19] or performing classifications within specific gender groups [17]. Others either focus on particular pathology groups or overlook these imbalances altogether.

These challenges open up multiple research directions. One promising avenue involves exploring more advanced oversampling and data augmentation techniques that balance the dataset and preserve the essential pathological information in the generated data. Another approach is to reduce the number of subjects used during the training process, mitigating the effects of imbalances on model performance.

Furthermore, testing different sets of features and classifiers remains a popular research strategy. This includes methods like automatic feature extraction, with a focus on interpretability, where convolutional neural networks (CNNs) are used to generate embeddings from spectrogram images, which are then utilized by more traditional models. Other promising strategies involve exploring alternative audio signal representations to extract novel features, applying transfer learning, or using fusion techniques to integrate external information into the classification process.

The reviewed works are summarized in Table 1.1. Some of the main limitations of these studies include a lack of analysis regarding the influence of other vowel tasks in the classification process [12]–[15], [18], [19], [21], a focus on specific voice pathologies [20], and limiting the analysis solely to speech signals [17].

The present study aims to identify which biomarkers are most effective for the different tasks and how the information from the two modalities can be combined.

Table 1.1. Summary of the State of the Art. **HC**: Healthy controls. **HD**: Hyperfunctional dysphonia. **L**: Laryngitis. **VP**: Voice pathologies. **MFCC**: Mel Frequency Cepstral Coefficients. **LPCCs**: Linear Prediction Cepstrum Coefficients. **HOS**: Higher-Order Statistics. **CNN**: Convolutional Neural Network. **ANN**: Artificial Neural Network. *****: Data was generated via oversampling

	Year	Participants (M/F)	Class	Task	Features	Classifier	Accuracy
[20]	2023	687 (259/428) - 207 (42/165) - 127 (75/52)	HC HD L	Vowel /a/, /i/, /u/	CNN generated features	Fully connected	81.0%
[19]*	2023	1354 (-/-) - 1354 (-/-)	VP, HC	Not reported	MPCC LPC	Fully connected	98.8%
[21]	2022	687 (259/428) - 1354 (657/727)	VP, HC	Vowel /a/	MFCC LPC Age F0	SVM	90.1%
[17]	2021	518 (259/259) - 518 (259/259)	VP, HC	Vowel /a/, /i/, /u/	MFCC LPCC HOS	Fully connected	82.7%
[13]	2018	685 (257/428) - 685 (257/428)	VP, HC	Vowel /a/	CNN Jitter Shimmer F0 MFCC	SVM DT Bayesian Classification Logistic Model Tree	84.1%
[15]	2018	244 (145/99) - 262 (137/125)	VP, HC	Vowel /a/	CNN	Fully connected	93.9%
[18]	2017	Not reported	VP, HC	Vowel /a/	Closed and open quotient MFCCs	GMM	93%
[14]	2017	1353 (-/-) - 687 (-/-)	VP, HC	Vowel /a/	CNN	Fully connected	68.1%
[12]	2016	50 (-/-) - 70 (-/-)	VP, HC	Vowel /a/	MFCCs First and second derivate	ANN	87.8%
[11]	2012	1320 (609/711) - 650 (400/250)	VP, HC	Vowel /a/, /i/, /u/	Acoustic features Noise related features	GMM	87.9%

1.3 Problem statement

Voice pathologies can arise from various muscular issues combined with contributing factors. These disorders may be due to organic, neurological, or overuse-induced conditions of the voice [23]–[25]. Organic causes include the growth of nodules, cysts, or other abnormal masses that hinder the normal function of the muscles responsible for voice production. Neurological conditions, such as paralysis, spasms, or involuntary movements of the vocal cords, may stem from brain or nervous system problems. Additionally, abuse of the voice through excessive shouting or unnatural pitch levels can lead to voice disorders.

The variability in the presentation of voice pathologies complicates their diagnosis. Age and gender play essential roles in manifesting these pathologies, and similar conditions may present with different characteristics. Most diagnostic methods rely on clinicians' subjective judgment, which can result in inconsistent diagnoses and treatments. Therefore, objective tools such as Electroglottography (EGG) analysis and speech signals are necessary, as they offer more reliable and uniform diagnostic conclusions.

Electroglottography is a noninvasive technique that records electrical impedance between electrodes placed on the neck, providing information on the vibration patterns of the vocal folds. Speech signal, on the other hand, captures the acoustic properties of the voice, offering insight into the sound output from the vocal tract. Both types of signals have their unique advantages and limitations.

EGG signals directly measure vocal fold activity, making them useful for analyzing pathologies related to irregular vocal fold activity, which is common in many voice disorders [26]. However, EGG has limitations in measuring aspects related to speech production, such as airflow characteristics. Factors such as patient anatomy, electrode placement, skin conductivity, and neck tension can also introduce noise and variability in EGG recordings [27].

Speech signals, on the other hand, are highly effective for investigating various voice pathologies. These signals enable the measurement of parameters like loudness, pitch, jitter, and formants, which often vary in the presence of a voice disorder. Unlike EGG, speech signals capture both the source (vocal fold vibration) and the filter (shaping of the signal by the vocal tract) involved in speech production. However, speech signals are prone to noise, often from environmental factors, and can be influenced by the speaker's

accent, emotions, and pronunciation [28].

The combination of EGG and speech signals could prove helpful, as each complements the limitations of the other. While EGG focuses on vocal fold movement, speech signals offer a broader perspective on voice production. Combining these two signals could improve the detection of voice pathologies and lead to the development of more robust diagnostic models.

1.4 Research question

To what extent can features extracted from speech signals be effectively transferred to EGG data to improve the detection of voice pathologies?

1.5 Objectives

1.5.1 General Objective

To design and evaluate classical and modern machine learning methods for discriminating between pathological and healthy control subjects, considering information extracted from EGG and Speech signals.

1.5.2 Specific Objectives

- Assess the efficacy of machine learning models in classifying patients with voice pathologies using features derived from speech signals.
- Assess the efficacy of machine learning models in classifying patients with voice pathologies using speech features extracted from EGG signals.
- Determine the impact of integrating EGG and speech signals on the accuracy and informativeness of models for voice pathology classification.

1.6 Contribution of this study

Multiple works have used the SVD database to classify voice pathologies, but only some of those works ensure that the results they obtain come from an unbiased dataset. To contribute to this way of analysis, the following are the main outcomes of this work.

- Achieved a detection accuracy of 64.6% for voice pathology using models trained on a balanced dataset with EGG signals.
- Enhanced classification performance in the uni-modal approach by incorporating phase plots in an early fusion strategy, leading to an accuracy of 88.5%.
- Verified the impact of dataset imbalances on model performance in voice pathology detection, providing insights into the effects of training data distribution.

1.7 Outline

This work is divided into four chapters. Chapter 2 explains the theoretical background used in this work. Chapter 3 discusses the data. Chapter 4 shows the methodology followed and the results obtained. Lastly, chapter 5 shows the conclusions of the work.

Chapter 2

Theoretical background

This chapter briefly explains some basic concepts in speech production and how the human body coordinates to perform the task. This chapter aims to inform the reader about the most common speech analysis and the features that will be extracted for the automatic analysis.

Lastly, the final part of the chapter shows the classifier and the mathematical formulation behind the classification process.

2.1 Physiological process of speech production

Speech production is a complex physiological process that involves multiple anatomical structures and neural pathways, as shown in Figure 2.1. The process begins with activating the speech motor cortex, a region in the brain's frontal lobe that controls the movement of the articulators, including the lips, tongue, jaw, and vocal cords [29].

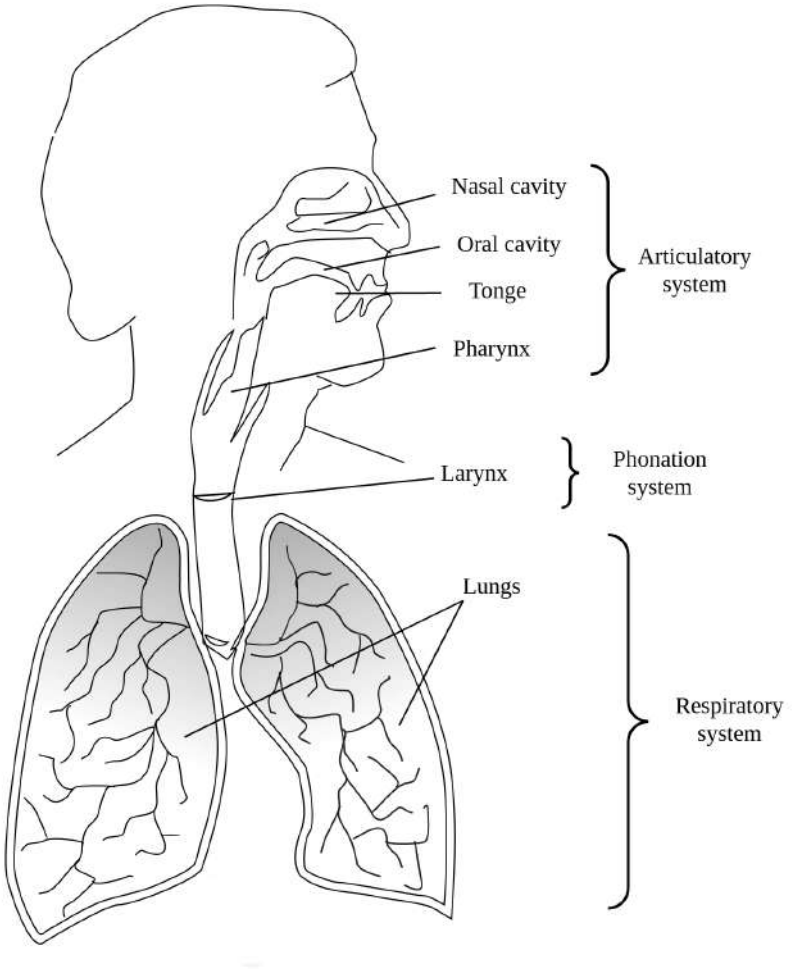


Figure 2.1. Important areas on speech production

This process involves a delicate balance between muscle control, airflow, and vibration, which must work together ideally to produce intelligible speech. Disruptions of these mechanisms can result in speech disorders that can impair communication and quality of life. For example, the vocal cords transform the energy that comes from the lungs into acoustic energy radiated by the lips; when a disruption occurs, this can be seen in changes in the vibration of these vocal cords. The vibration on the vocal cords has a waveform that goes up and down in the mucous membrane surface in regular cycles [30].

2.2 Electroglottography

Human sound is produced by combining different muscles, including the vocal folds. Vibration in the vocal folds generates a quasi-periodic sound. These vibrations are caused by the air that comes from the lungs through the trachea. This generates the fundamental frequency of those vibrations depending on how soft or rigid the vocal folds are [31].

Due to their importance in speech production, research has focused on analyzing vocal folds. The best way is to observe the movement of these folds directly when the speech is being produced. This can be achieved by a laryngeal endoscopy, usually done when a patient presents problems with vocal quality or hoarseness [32]. This endoscopy can be performed directly or indirectly. The easiest and least invasive of the two is the indirect way, where the physician introduces a mirror and points it with a light; this mirror goes to the back of the mouth and allows the physician to see parts of the laryngeal area [32].

The direct way can be divided into two types: rigid and flexible. The flexible examination requires the patient to be upright, and the physician introduces a fiberscope or flexible endoscope through the nose that consists of two optical wires inside a flexible unit. The rigid examination uses the same fiberscope, but this time, it is inserted through the mouth [33]. In both cases, the patient is requested to do sustained vowels. At the same time, the physician observes the closing of the glottis during the phonation process through the fiberscope camera. Afterward, the clinicians compared the degree of closure of the vocal fold based on a scale from 1 to 6, with 6 being fully open.

It is worth mentioning that the patient is awake during all procedures, and there is no sedation process. Sometimes, a small portion of a numbing substance is sprayed [33] or applied with cotton pledgets [34]. Some works have also reported that if the procedure needs to be done on infants, three people are required to hold the patient in place, and the infant is wrapped in a sheet to restrain the movement that the discomfort of the fiberscope can cause [34].

All this creates a procedure that is too invasive and generates discomfort. Currently, this procedure is the most straightforward and gives the best information. However, alternatives such as electroglottography are also used.

Electroglottography, sometimes also known as electrolaryngography (ELG), is a method that captures the opening of the glotias via an electric current sent through a couple of electrodes. These electrodes are placed on opposite sides of the neck close to the glottis, and small currents with high frequency are sent from one electrode to another. The basic concept behind this is that when the glottis is closed, the folds around the glottis are touching, allowing the current to flow through the muscles without too much resistance. This generates the maximum value of amplitude in the EGG signal.

On the other hand, when the folds are open, air is in between. This increases the resistance, reducing the amplitude of the signal received. Research has shown that other substances, such as mucus or cysts, can also generate changes in the signal.

As we can notice, this method allows physicians and researchers to know how the vocal folds are behaving non-intrusively. A comparison of speech and EGG can be seen in Figure 2.2.

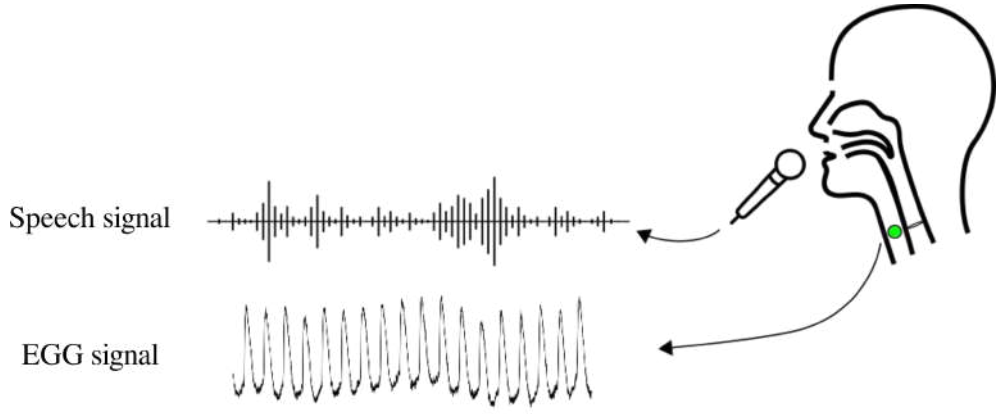


Figure 2.2. Comparison of speech and EGG signal

2.3 Feature extraction

2.3.1 Nonlinear features

Studies have shown evidence that the vocal tract is not linear [35], that there is the existence of nonlinear structures in the speech signals [36], and that voice pathology, depending on the impairment, can generate nonlinear pressure-flow in the glottis or nonlinear collision in the vocal fold area [37].

Defining the speech process as a linear model means that in the case of an increase in the velocity of the sound, the passive systems, like the glottis, will have an increase in the resonance, and that is not usually the case [36].

The existence of non-linearity in speech dynamics has led research to treat it as a turbulence or dynamical system. In the area of the fluid, turbulence problems are analyzed either stochastically with concepts like autocorrelation functions or deterministically with the chaos theory [38].

The dynamic systems have the characteristics of using phase spaces to define the changes over time, which can be defined either by differential equations or dimensional maps [39], as shown in Equation 2.1.

$$\begin{aligned} x_{n+1} &= \mathbf{F}(x(n)) \\ \frac{d}{dt}x_t &= \mathbf{f}(x(t)) \end{aligned} \tag{2.1}$$

The phase spaces, known as state spaces, can determine future states based on a fixed present state in a fully deterministic system. Analyzing the phase space dynamics allows us to study the dynamics of the problem, i.e., the vocal tract [40].

The trajectories described by the phase space are called *attractor*. This attractor can show graphically how "chaotic" a system can be, generating different types of attractors like a fixed-point type when the system is non-chaotic, a limit-cycle type when the system becomes periodic after some time, and the strange type that represents a chaotic system [41].

In real-world problems, we don't have phase spaces; we only have a series of data. Our job is to transform this data in a state space; this is done using the delay reconstruction method [41]. Supposedly, we want to find the *nth* element in our phase space; this can be reconstructed using Taken's theorem [42]. This theorem states that a dynamical system representation can be formed using a time-delay version of the original series. We can define the vector $\mathbf{X}_n(t)$ as shown in Equation 2.2, where τ is the delay and m is the dimension. Figure 2.3 shows the comparison of the phase space or attractors with $m = 3$ from an HC and a VP using the speech signal and the vowel /a/

$$\mathbf{X}_n(t) = [x(t), x(t-\tau), x(t-2\tau), ..., x(t-(m-1)\tau)] \tag{2.2}$$

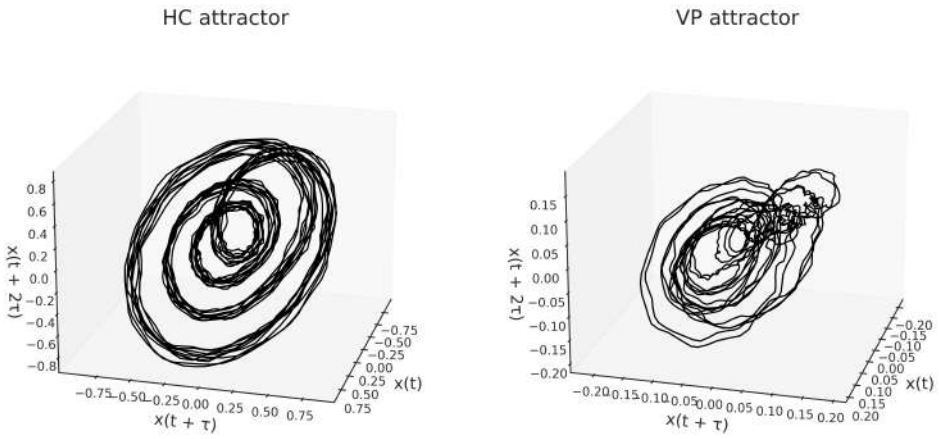


Figure 2.3. Attractor comparison between HC subject and VP subject. **HC**: Healthy Control. **VP**: Voice Pathology

Multiple nonlinear features can be extracted. In this work, we computed Largest Lyapunov Exponent (LLE), Sample Entropy (SampEn), and Hurst Exponent (HE).

Largest Lyapunov Exponent

This feature quantifies the predictability or stability of a system in the presence of changes [43]. It is computed based on the concept of attractors, with the exponents defined as the average differences between neighboring points in the attractor.

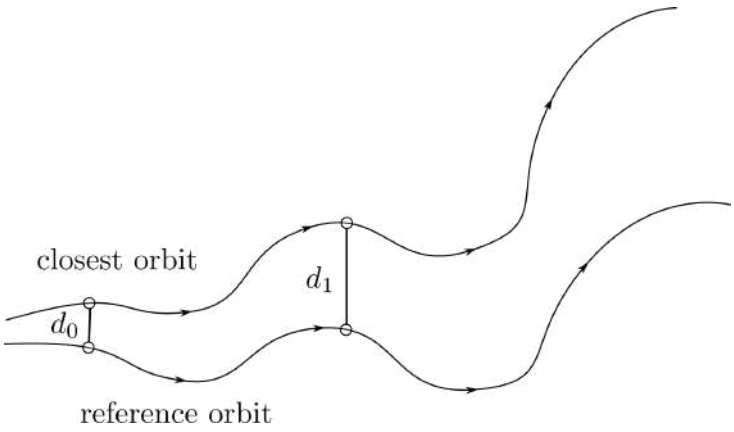


Figure 2.4. Distance between orbits for calculation of the Lyapunov Exponent

Figure 2.4 shows an example of a zoomed attractor where two orbits are close. A reference orbit is selected, as well as a point d_0 from an initial t_0. Afterward, we can define a point in d_1 in the attractor for a time t_1, where $t_1 > t_0$. The ratio between the distance can be expressed exponentially [44] as shown in Equation 2.3.

$$\frac{d_1}{d_0} = e^{\lambda(t_1 - t_0)} \tag{2.3}$$

Where λ is the Lyapunov exponent. It is worth mentioning that when $\lambda > 0$, both orbits will be drifting from each other exponentially, $\lambda < 0$, the orbits will end in 0, and if $\lambda = 0$, the change over time is not exponential.

For the calculation of the exponents, distances below a predefined threshold ϵ are discarded, and the highest of these exponents is selected as the feature. The exponent provides a quantitative measure of the level of chaos or irregularity present in the signal. It can potentially differentiate between standard speech and speech affected by a disorder.

Sample Entropy

This feature measures the level of regularity or unpredictability in time series data. It is used to quantify a signal's complexity by calculating the likelihood that similar patterns will repeat themselves inside the signal. SampEn was defined as an improved estimate for the randomness of data compared to the approximate entropy (ApEn). Studies have shown that ApEn introduces statistical biases that SampEn solves [45].

For the calculation of SampEn, a template of consecutive data points with size m is generated, as shown in Figure 2.5. This template is then compared with another set of m points within the dataset, and the number of times the template appears in the dataset is counted. It is worth noting that when the template finds itself, this is not counted as this is one of the biases in the approximate entropy calculation [46].

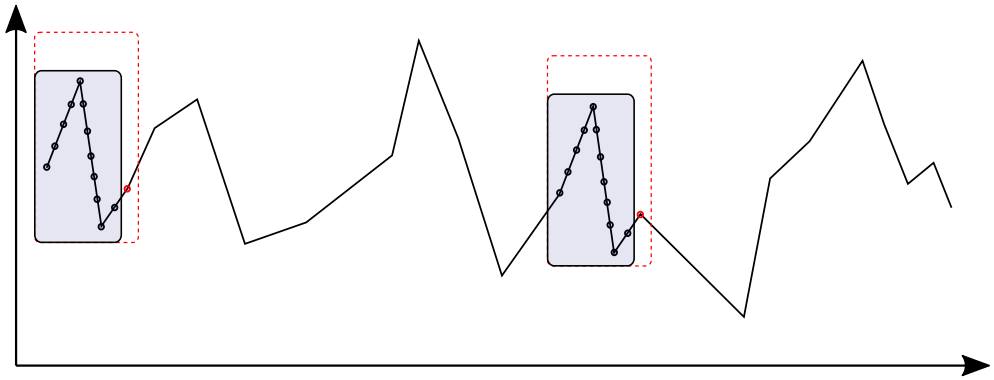


Figure 2.5. Points comparison for sample entropy

This number is commonly named B and can be represented as $B = \sum_{i=1}^{N-m} \sum_{j=1,j\neq i}^{N-m} (d[x_m(i) - x_m(j)] < r)$ as shown in Equation 2.4, where d is the euclidean distance. The value r is a tolerance set to verify if two points are similar, normally set to $0.2 * std(x_m)$, being std the standard deviation [47].

Lastly, the algorithm counts the number of times the template $m+1$ is similar to a set of points of size $m+1$. This value is commonly known as A and can be represented as $A = \sum_{i=1}^{N-m} \sum_{j=1,j\neq i}^{N-m} (d[x_{m+1}(i) - x_{m+1}(j)] < r)$, as shown in Equation 2.4. Finally, the negative logarithm of the ratio between A and B is calculated as the sample entropy.

$$SampEn = -log \frac{\sum_{i=1}^{N-m} \sum_{j=1,j\neq i}^{N-m} (d[x_{m+1}(i) - x_{m+1}(j)] < r)}{\sum_{i=1}^{N-m} \sum_{j=1,j\neq i}^{N-m} (d[x_m(i) - x_m(j)] < r)} \tag{2.4}$$

Hurst Exponent

This feature quantifies the possible “long-term memory" or the presence of long statistical dependencies in a time series that are not attributed to cycles. It was defined by [48] when analyzing the problem of river water storage in

reservoirs for irrigation. Hurst found that if we divide the variation rank R by the standard deviation of the signal S this can be represented as shown in Equation 2.5

$$\frac{R}{\sigma} = \left(\frac{N}{2}\right)^{H} \tag{2.5}$$

Where N is the length of the segment, and H is the Hurst exponent. Hurst also showed that the value of H changes depending on the phenomena [49].

First, the series is divided into segments to calculate the Hurst exponent from a time series. Each segment is divided into segments, and the means, cumulative deviations, and range of cumulative deviations are calculated. Lastly, the segment's standard deviation is computed. The Hurst exponent is the logarithm of the ratio between the range and the standard deviation.

The range of values for this exponent typically spans from 0 to 1. A value of 0.5 corresponds to a random process or no correlation between values. Values higher than 0.5 mean that the time series is persistent, so increments usually follow increments. On the other hand, values lower than 0.5 represent an anti-persistent series where values usually decrease after an increment in the time series.

2.3.2 Phonation features

Phonation is a critical part of speech production; it includes analyzing the vocal folds, the diaphragm, and the glottal cavities, among others. Depending on their position, these muscles produce sounds with different tones, pitch, and volume.

People with voice pathologies present problems such as lower volume, breathiness, or raspiness. If the stability of the audio is analyzed, problems in the vocal folds can be detected.

Fundamental Frequency

Phonated speech is created by quasi-periodic vibrations in the vocal folds [50]. The frequency in this periodicity is known as fundamental frequency. Figure 2.6 shows a 40 ms segment of a healthy subject during the sustained vowel /a/; the figure shows the quasi-periodicity in the voice production, where T_0 is the fundamental period.

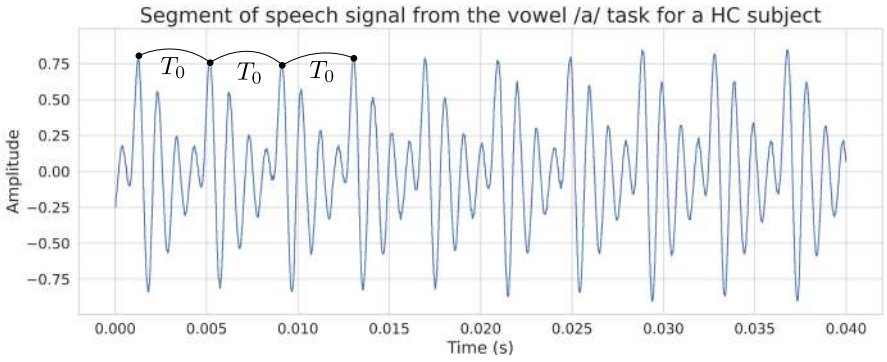


Figure 2.6. Example of quasi-periodicity in speech. **HC**: Healthy Control.

In the case of the speech signals, F_0 is sometimes considered the pitch and is defined as the rate of the vibration of the vocal folds [51]. F_0 is usually extracted from short-time frames (e.g., 40ms) using the autocorrelation, which compares segments of the signal with other segments offsets to find a match.

Temporal perturbation of the fundamental frequency - jitter

The word "quasi" in quasi-periodic means "more or less". Voice production not being fully periodic always means that the F_0 value is not the same during the whole audio length. This behavior is called a frequency perturbation and can be represented as the frequency variability between cycles [52]. The measurement of this perturbation is called jitter. Figure 2.7 and Figure 2.8 show the contrast of the frequency perturbation between a healthy subject versus a patient with a voice pathology, the figures show that $|T_{i-1} - T_i| < |T_{j-1} - T_j|$.

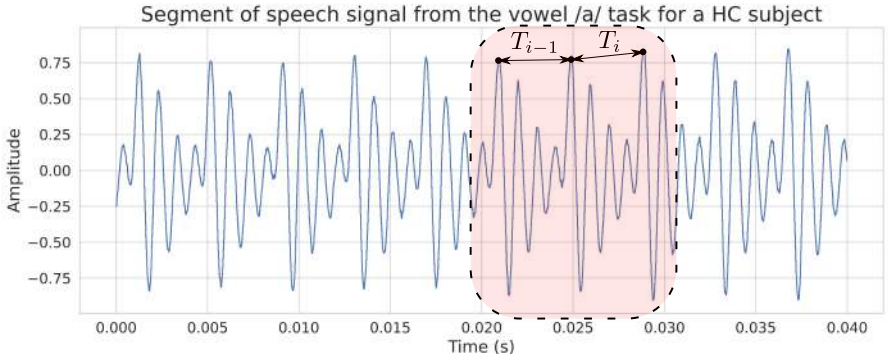


Figure 2.7. Temporal perturbation in an HC subject. **HC**: Healthy Control.

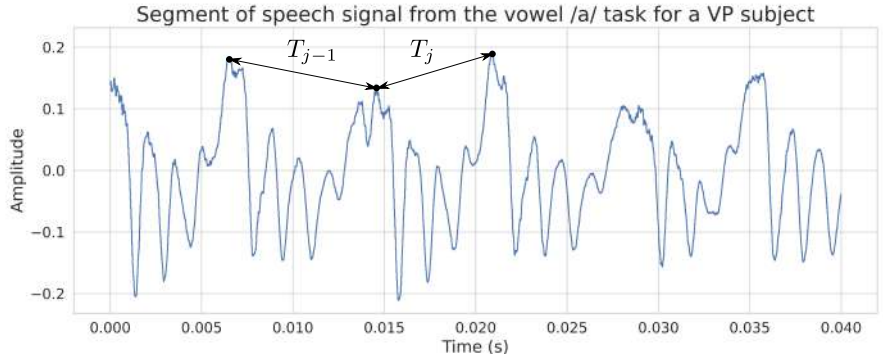


Figure 2.8. Temporal perturbation in a VP subject. **VP**: Voice Pathology.

Different types of Jitter can be calculated:

- Jitter (absolute) is the cycle-to-cycle variation of the fundamental frequency calculated between two periods.

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

- Jitter (relative) is the average difference between two consecutive periods divided by the average period.

$$Jitter(relative) = \frac{Jitter(absolute)}{\frac{1}{N}\sum_{i=1}^{N} T_i} * 100$$

Studies have shown that relative Jitter values are usually less than 0.5% for healthy subjects [53] and subjects with impairments like vocal tremor, aphonia, or roughness in the voice present higher jitter variations [52].

Amplitude perturbation of the fundamental frequency - shimmer

Shimmer, also known as amplitude perturbation, is similar to jitter but with a distinct focus. While jitter measures variations in the duration of successive periods, shimmer examines fluctuations in the amplitude or the signal's peak value within each fundamental period.

It is usually represented in dB, and values for healthy subjects are less than 0.35 dB [54]. Figure 2.9 and Figure 2.10 show a similar contrast as the ones shown in jitter, but in this case, we focus on the perturbations in the amplitude. These plots are generated from the EGG signal and the phrase task, showing that the metric can be extracted for both signals and gives information for different tasks.

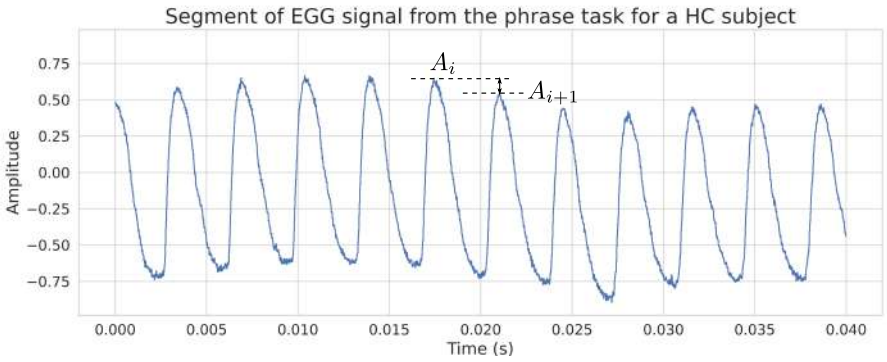


Figure 2.9. Amplitude perturbation in a HC subject. **HC**: Healthy Control.

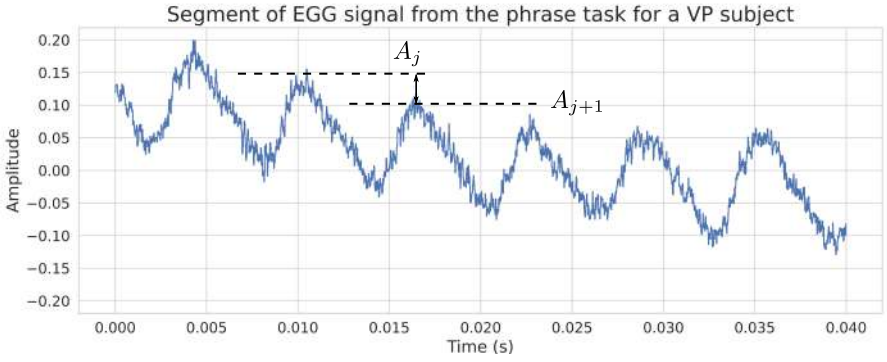


Figure 2.10. Amplitude perturbation in a VP subject. **VP**: Voice Pathology.

It is worth noticing that not only $|A_{i-1} - A_i| < |A_{j-1} - A_j|$ but also the VP subject exhibits bigger fluctuations in the amplitude for the following

fundamental periods.

Similar to Jitter, different types of shimmers can be calculated.

- Shimmer (dB) is the variability of the peak-to-peak amplitude in decibels

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

- Shimmer (relative) is the average difference between two consecutive periods divided by the average period.

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

Relative average perturbation (RAP)

It represents the average deviation of a period from its mean value and two adjacent periods, normalized by the period's average [55]. The equation for the RAP metric can be found in Equation 2.6

$$RAP = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|T_i - \left(\frac{1}{3}\sum_{n=i-1}^{i+1}T_n\right)\right|}{\frac{1}{N}\sum_{i=1}^{N}T_i} * 100 \tag{2.6}$$

Pitch perturbation quotient (PPQ)

This type of jitter is defined as the ratio of perturbations over certain periods to the average period. It is commonly represented by the symbol PPQ followed by a number, e.g., PPQ5, which represents the ratio of disturbances over five periods divided by the average period [55], as shown in Equation 2.7.

$$PPQ5 = \frac{\frac{1}{N-1}\sum_{i=2}^{N-2}\left|T_i - \left(\frac{1}{5}\sum_{n=i-2}^{i+2}T_n\right)\right|}{\frac{1}{N}\sum_{i=1}^{N}T_i} * 100 \tag{2.7}$$

Amplitude perturbation quotient (APQ)

Similarly to PPQ, APQ is another metric that asses perturbations related to shimmer. APQ quantifies the changes in amplitude across multiple fundamental periods [54]. Like PPQ, APQ is denoted at the end with a number,

e.g., APQ3, APQ5. The number denotes the number of fundamental periods considered for the averaging. This average at the end is divided by the mean of the amplitude as shown in Equation 2.8.

$$APQ3 = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|A_i - \left(\frac{1}{3}\sum_{n=i-1}^{i+1} A_n\right)\right|}{\frac{1}{N}\sum_{i=1}^{N} A_i} * 100 \tag{2.8}$$

2.3.3 Articulation features

Figure 2.1 shows that the last part of the speech production process involves the articulatory system. This system is based on different sections, as shown in Figure 2.11, that mold the vibrations from the vocal folds [56]. It includes organs such as the tongue, lips, teeth, or mouth areas like the hard palate or the alveolar ridge, among others [57].

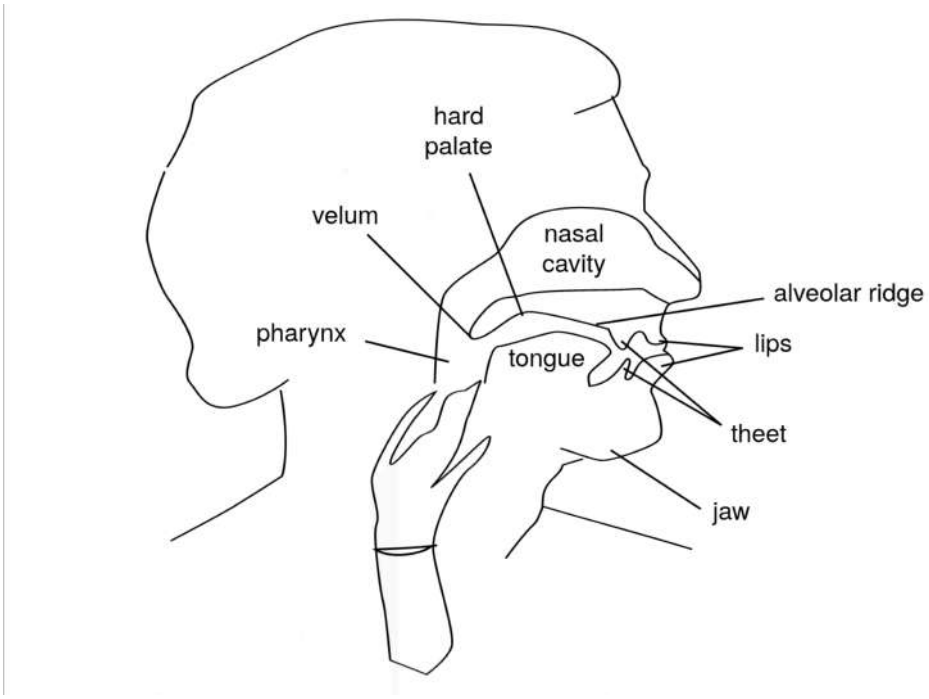


Figure 2.11. Articulators in vocal tract

Depending on the phonemes, multiple articulators move. In the vowel's case, for example, the change is mainly done in the thong and the lip's positions because they are produced by allowing the free pass of the airflow and without friction [57]. For instance, in the production of the vowel /a/,

the thong is usually low while the lips are open, and for the vowel /u/, the thong is closer to the roof of the mouth, and the lips are closer and round [58].

Articulation features describe and analyze how those different limbs and muscles work together in speech production. Voice pathologies can affect different muscles and nerves used to move these muscles efficiently [59]–[61], so modeling how these muscles interact is crucial in speech analysis.

Articulation can be analyzed in both voiced and unvoiced segments. Studies have shown that patients with voice pathologies, e.g., Parkinson's disease, have difficulties with some articulatory movements [62] or speeding when doing repetitive muscle movement [63].

Articulation features are extracted using Disvoice [64]; a total of 122 descriptors are generated, but all the descriptors, including voiced and unvoiced segments, are removed, leaving only the first and second formant as well as their firsts and second derivatives. The Mel Frequency Cepstral Coefficients of the audio replace the removed descriptors.

Mel Frequency Cepstral Coefficients

Our hearing system is based on the external part of our ear that captures the audio signals, a middle part that converts the sound waves to pressure, and an internal part that takes these pressure signals and sends information to the brain [65]. This inner part is mainly formed by the cochlea, which has a membrane that vibrates in a set of frequencies. This non-uniform vibration focuses primarily on lower frequencies [65].

This knowledge has led researchers to estimate how our ear captures sound by defining a set of bandpass filters separated linearly in lower frequencies and logarithmically in higher frequencies [66].

MFCCs are one of them; they focus on extracting audio information based on how the human auditory system works. This system doesn't work linearly; for a tone with frequency f, a pitch can be measured on a 'Mel' scale. The Mel (Melody) scale focuses on that pitch, the relative tone the ear perceives, rather than the actual frequency. It was established based on human perception experiments.

Equation 2.9 shows the equation to find the value of a frequency in a 'Mel' scale.

$$\text{MEL}(f) = 2595 * log_{10}\left(1 + \frac{f}{700}\right) \tag{2.9}$$

An analysis in the frequency domain is required to apply the filter bank to the signal. This filter bank consists of triangular filters with the spacing and bandwidth defined by the Mel-frequency constant, as shown in Figure 2.12. The number of filters will depend on the constants we want to analyze. This shows that the number of filters and the type of window used to divide the audio are hyper-parameters selected when the MFCC is extracted.

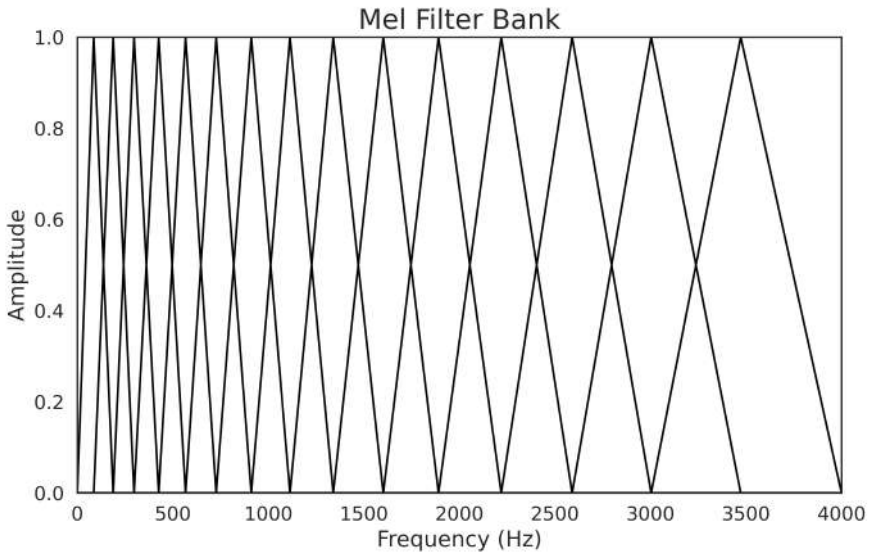


Figure 2.12. MFCC triangular filters. **MFCC**: Mel Frequency Cepstral Coefficients

To extract the MFCCs, the audio signal must pass through multiple steps shown in Figure 2.13. First, a pre-emphasis filter is used to increase the relevance of high-frequencies. This increment in the energy helps when we work with fricative phonemes such as 's' or 'f' [67]. The filter is calculated using Equation 2.10, where $s(n)$ is the input signal and α is the filter's cutoff frequency, and it takes values around 0.94.

$$y(n) = s(n) - \alpha s(n-1) \tag{2.10}$$

The next step is to frame the audio in small chunks, usually from 20 to 40 ms, with an overlapping of 50%. Now that the frames have a fixed length, they are passed through a Hamming window to avoid discontinuities at the

endpoints. The function of the Hamming window is multiplied by the frame and is represented in Equation 2.11, where N is the length of the frame and $0 \leq n \leq N-1$.

$$W(n) = 0.54 - 0.46cos\left(\frac{2\pi n}{N-1}\right) \tag{2.11}$$

Each frame is converted into the frequency domain using the Discrete Fourier Transform (DFT), and once in the frequency domain, the frames are filtered using the Mel filter banks.

Finally, to extract each coefficient, the representation in the Mel-scale is brought back to the time domain using a Discrete Cosine Transform (DCT).

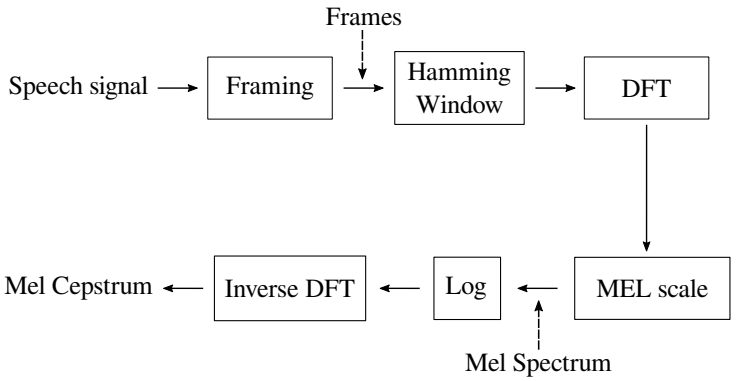


Figure 2.13. MFCC extraction process diagram. **MFCC**: Mel Frequency Cepstral Coefficients

Twenty MFCC features are extracted using a 200 ms Hamming window with 50 % overlap. Because it is a static analysis, four statistics (mean, std, skewness, and kurtosis) are extracted for each descriptor (104 features, 80 descriptors for the MFCC, and 24 descriptors with the formants).

2.3.4 Bark Frequency Cepstral Coefficients

Bark-frequency cepstral Coefficients (BFCCs) work similarly to the MFCC. These coefficients are based on the Bark scale, an alternative to represent how humans perceive sounds. While the Mel scale focuses on the perceived pitch, the Bark scale is based on how the basilar membrane performs a spectral analysis, which can be modeled with band-passing filters with a bandwidth of one critical band or one Bark [68].

Both are perceptual scales that aim to reflect how humans perceive sound from different perspectives. The extraction process is the same for the

MFCCs in Figure 2.13, but changing the scale to the Bark scale, Figure 2.14 shows a comparison of the triangles bandpass filter bank for both MFCC and BFCC. It is worth noting that before 500 Hz, both scales are equal; this is based on the premise that the bands are almost linear up to 500 Hz [69]; after that point, we can notice the small differences between the two scales. Equation 2.12 shows how to convert frequency in Hertz to Bark [70].

$$\mathrm{BARK}(f) = 6\ln\left(\left(\frac{f}{600}\right) + \sqrt{\left(\frac{f}{600}\right)^2 + 1}\right) \tag{2.12}$$

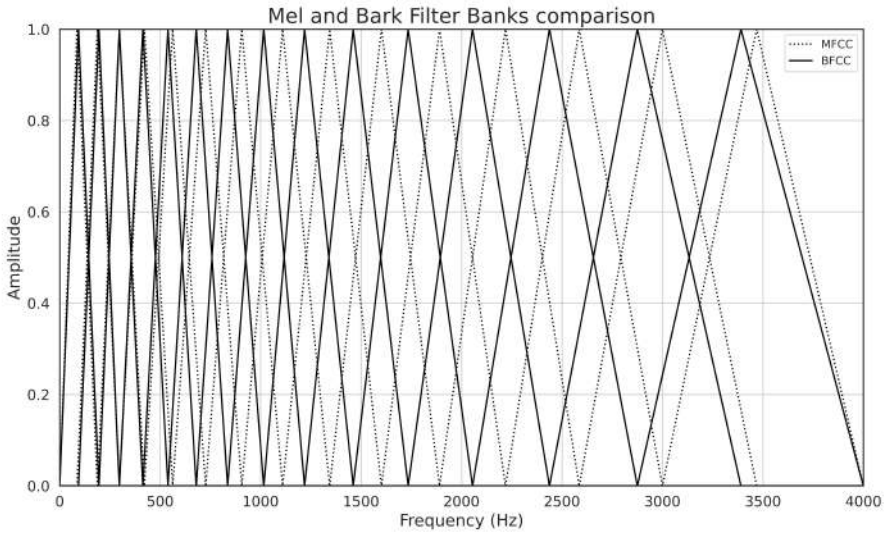


Figure 2.14. Comparison of BFCC and MFCC triangular filters. **MFCC**: Mel Frequency Cepstral Coefficients. **BFCC**: Bark Frequency Cepstral Coefficients

2.3.5 Phase plot analysis

In general, phase plots can be obtained by plotting the real and imaginary parts of an analytic signal of the form

$$z(t) = x(t) + jy(t) \tag{2.13}$$

where $x(t)$ is the acoustic/EGG signal and $y(t)$ is the imaginary part obtained with the Hilbert transform. As shown in [71], when the phase plots

are extracted from glottal signals, glottal cycles are represented as elliptical trajectories superimposed in a 2D plane. In the case of acoustic signals, the phase plots result in more complicated shapes due to the non-linearities present in the signal.

Figure 2.15 shows an example of the phase plot extracted from a segment of a sustained phonation.

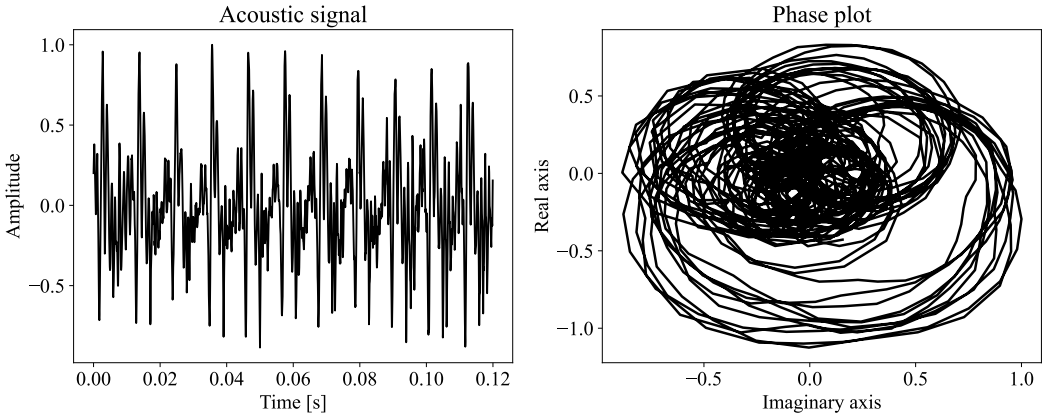


Figure 2.15. Example of the phase plot extracted from a segment of the sustained phonation of a vowel.

From the Figure, it can be observed that analyzing abnormal vibration in the signal can be complex due to the complexity of the plot, which is the result of the components (e.g., harmonics) that compose the acoustic signal. Thus, we propose the following procedure to perform analysis with phase plots:

- Extract the temporal fine structure (TFS) [72] of the acoustic/EGG signal to reduce the complexity of the phase plot. The TFS is obtained by dividing the acoustic/EGG signal by the amplitude envelope of the signal, i.e., the magnitude of the analytic signal. The top part of Figure 2.16 shows an example.

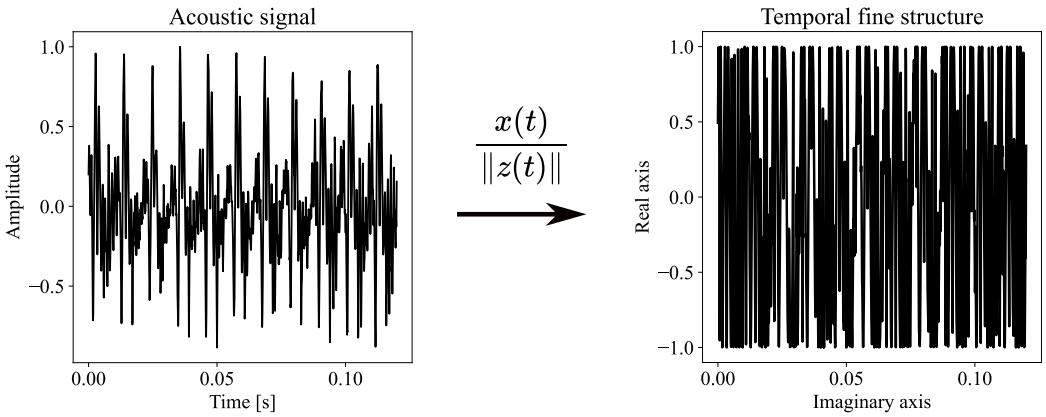


Figure 2.16. Extraction of the TFS.

- Compute the Hilbert transform of the resulting TFS to obtain $z'(t) = x'(t) + jy'(t)$ and plot the real and imaginary parts to get the phase plot, i.e., $x'(t)$ vs. $y'(t)$. The middle part of Figure 2.17 shows an example.

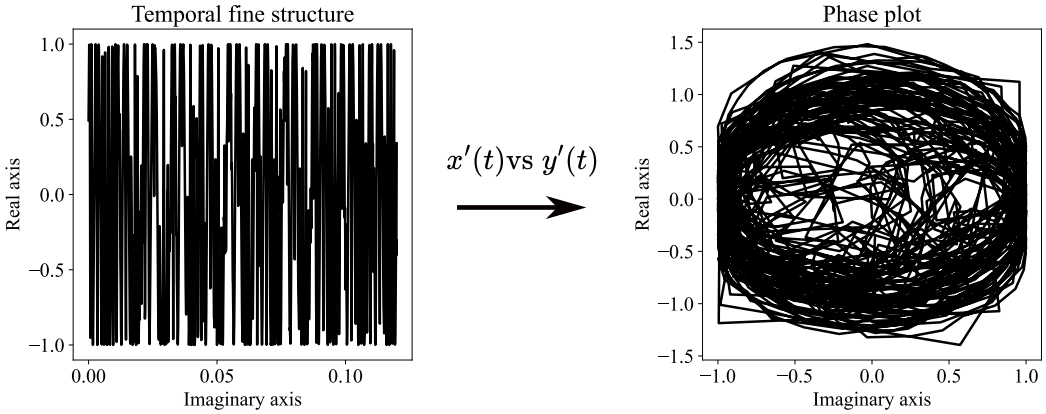


Figure 2.17. Extraction of the phase plot from the analytic signal of the TFS ($z'(t)$).

- Transform the resulting phase plot into an image with dimensions 256×256. The phase plot is converted into a heatmap by computing a bi-dimensional histogram with 248 bins and applying a 2D Gaussian filter with a standard deviation of 12 to smooth the data points. The bottom part of Figure 2.18 shows an example.

An example of the phase plots (as heatmaps) obtained from the recordings of a healthy control and a patient can be observed in Figure 2.19. It can be

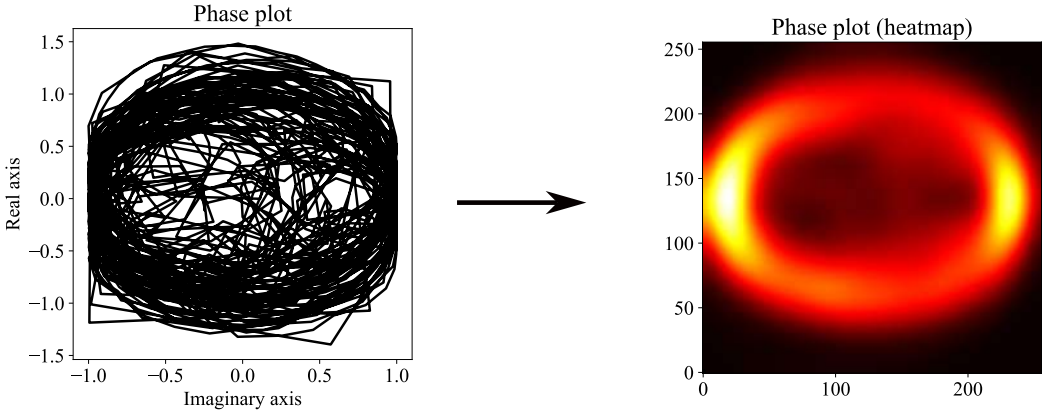


Figure 2.18. Phase plot converted into a heatmap.

observed that the phase plot of the patient is more “noisy" than the one of the healthy subjects.

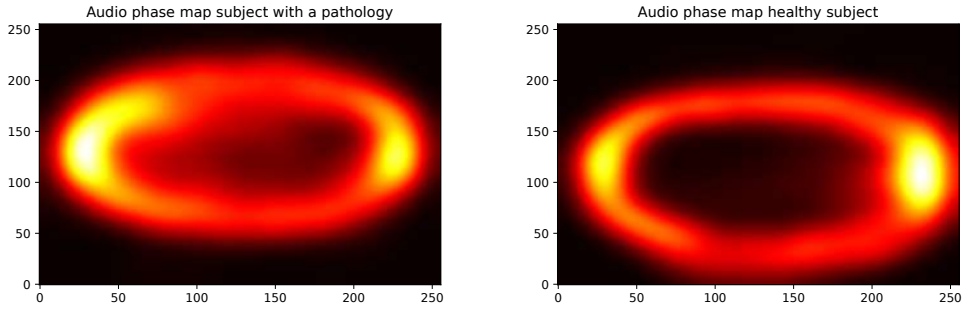


Figure 2.19. Phase plot of a patient (left) and a healthy subject (right)

Because the generation of the phase plots requires the use of the Hilbert transform to calculate the real and complex part of the signal, which at the same time uses the Fourier transform, then the computational complexity in terms of time can be represented in O notation as $O(n \log n)$, being n the number of samples in the audio that depends on the length of the audio and the sample frequency.

Phase plots are significant because they provide a condensed graphical representation of the glottal cycle during vocal fold vibration. Traditionally, these graphical representations have been subjectively analyzed. Research has shown that certain patterns in these plots correspond to specific vibratory behaviors, which in turn can indicate the presence of certain types of voice pathologies.

For example, phase plots are built on rims that represent a cycle. Since the audio contains multiple cycles, the rims overlap in a circular pattern. Clinicians often observe features such as the shape of the rims, the degree of vibratory irregularities, the luminosity or intensity in certain areas, and the curvature of the rims to diagnose voice problems [73].

Some examples of these patterns are illustrated in Figure 2.20. In one case, a male subject over 50 years old with a pathology is shown. The heatmap reveals high luminosity on the right side of the rims and a flattened right edge, which may be associated with a long open phase that abruptly transitions to a closed phase. The intensity in that area suggests that this pattern occurs multiple times during the glottal cycles, generating many points in that part of the plot.

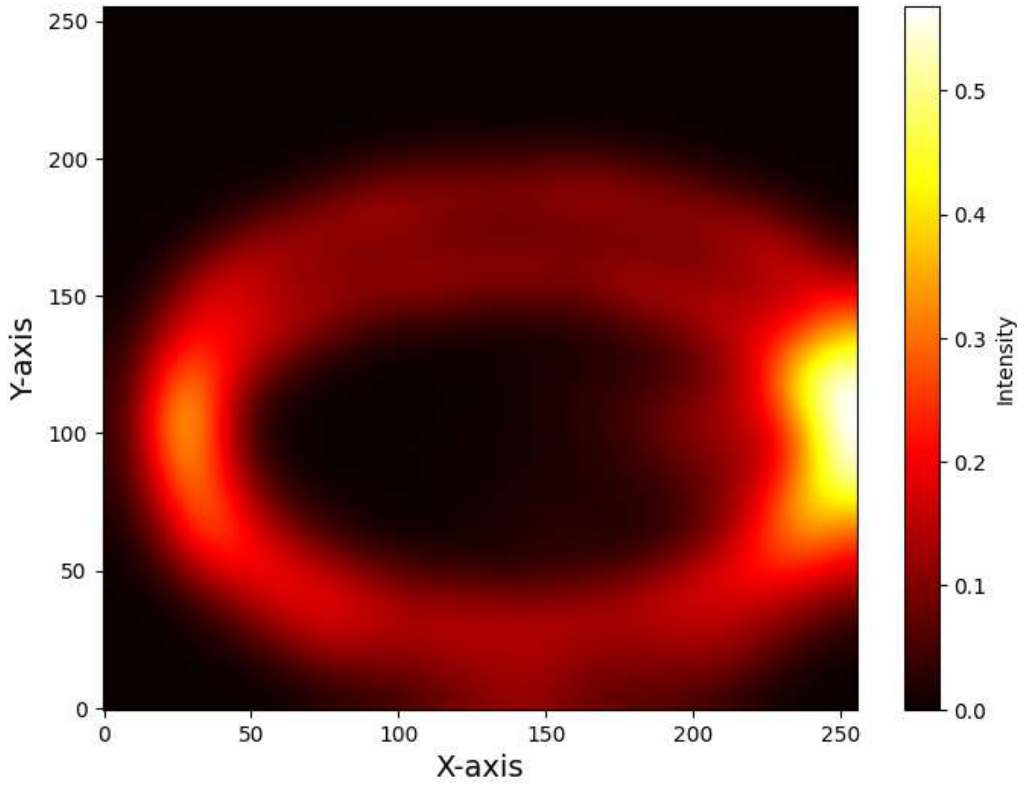


Figure 2.20. Heat map of phase plot for a male with over 50 years old and a voice pathology

A similar pattern is observed in the heatmap shown in Figure 2.21, this time for a female subject between 20 and 30 years old with a voice pathology.

This subject exhibits a similar pattern on the right side of the heatmap, but also displays two rims with higher luminosity compared to the previous heatmap, indicating the presence of two distinct patterns during the glottal cycle.

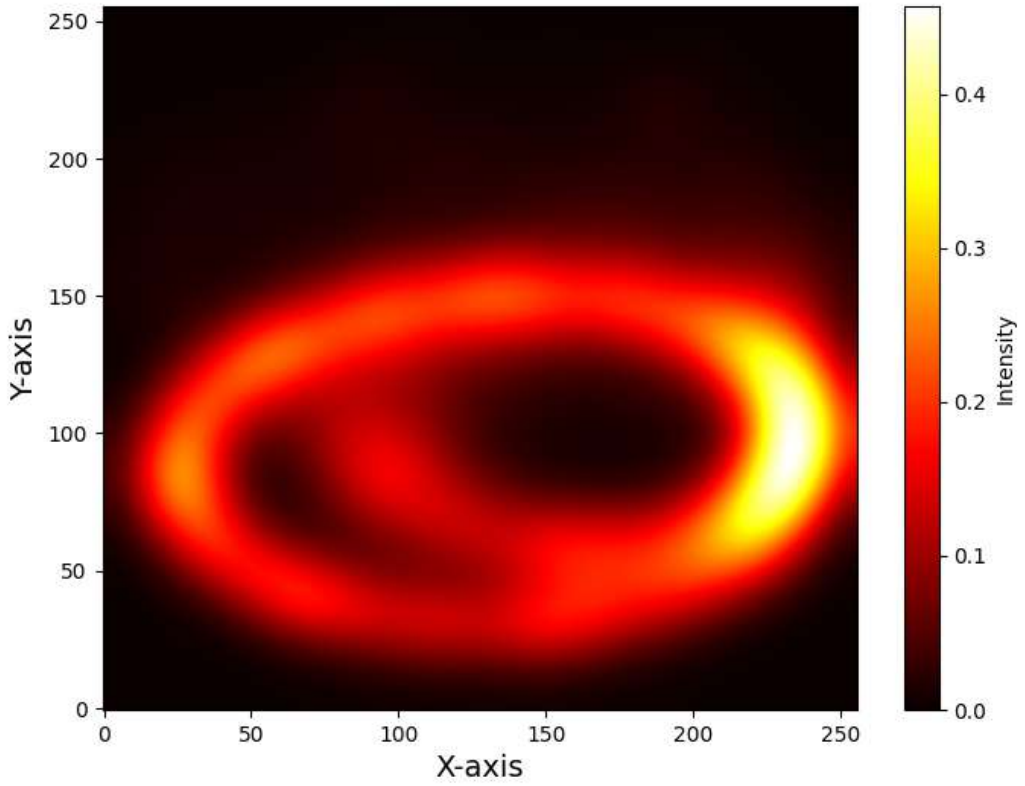


Figure 2.21. Heat map of phase plot for a female between 20 and 30 years old and a voice pathology

Lastly, a heatmap of a healthy subject is presented in Figure 2.22. This heatmap shows a long closed phase, unlike the pathological cases in Figure 2.20, as the intensity is now concentrated on the left side of the heatmap. Additionally, this figure displays fewer vibratory irregularities, maintaining a circular figure with only one prominent rim, which is typical in healthy vocal function.

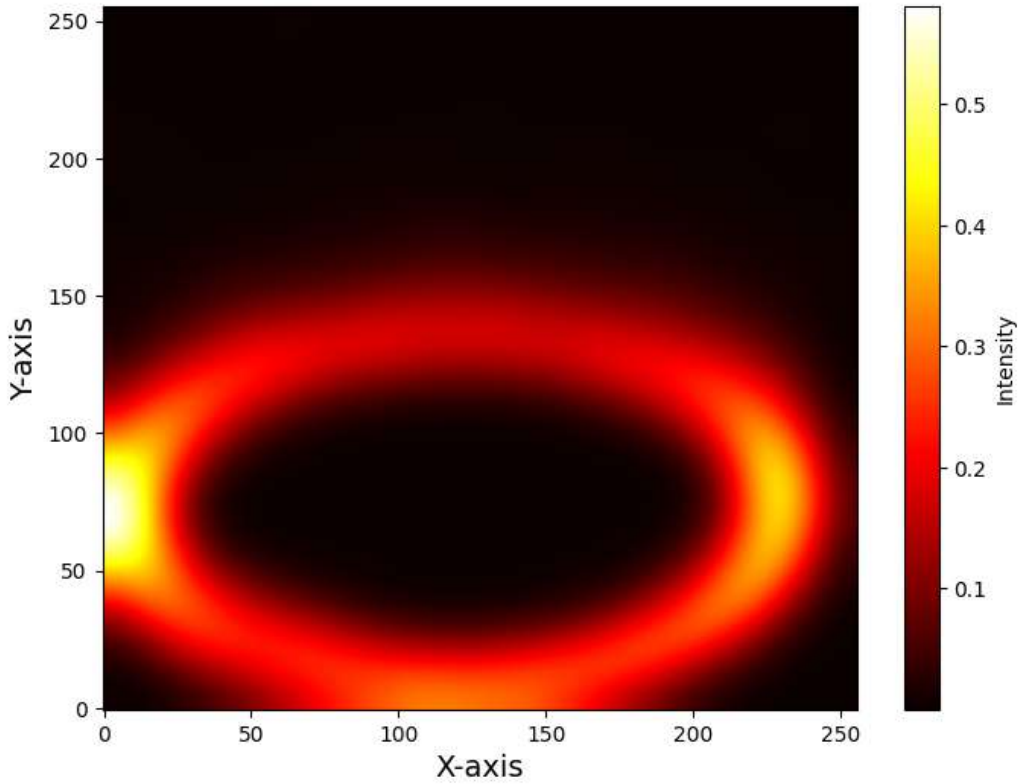


Figure 2.22. Heat map of phase plot for a healthy male between 20 and 30 years old

The quantitative analysis of the proposed phase plots is performed by training a CNN and then using embeddings of the last layers as a feature vector, no special system requirements are needed for the creation and calculation of the phase plots but a GPU is recommended for the training of the CNN.

The details of the architecture used in this study are given in Section 2.4.7.

2.4 Pattern recognition methods

2.4.1 Hard Margin Support Vector Machine

Support Vector Machines, a popular supervised learning method, are the classifiers in this research. SVMs excel at creating hyper-planes to separate data points.

Suppose we have a linear SVM; the main focus is to find a margin or so-called hyperplane that maximizes the distance between the data and itself.

For a binary classification problem, we have a subject S_i with a label y_i and a set of features x_i with size d, being d the number of features.

With this, we can select a decision limit, as shown in Figure 2.23 where:

$$< w, x_i > +b = 0$$

where $< w, x_i >$ is the dot product and w and b are the parameters of the SVM model.

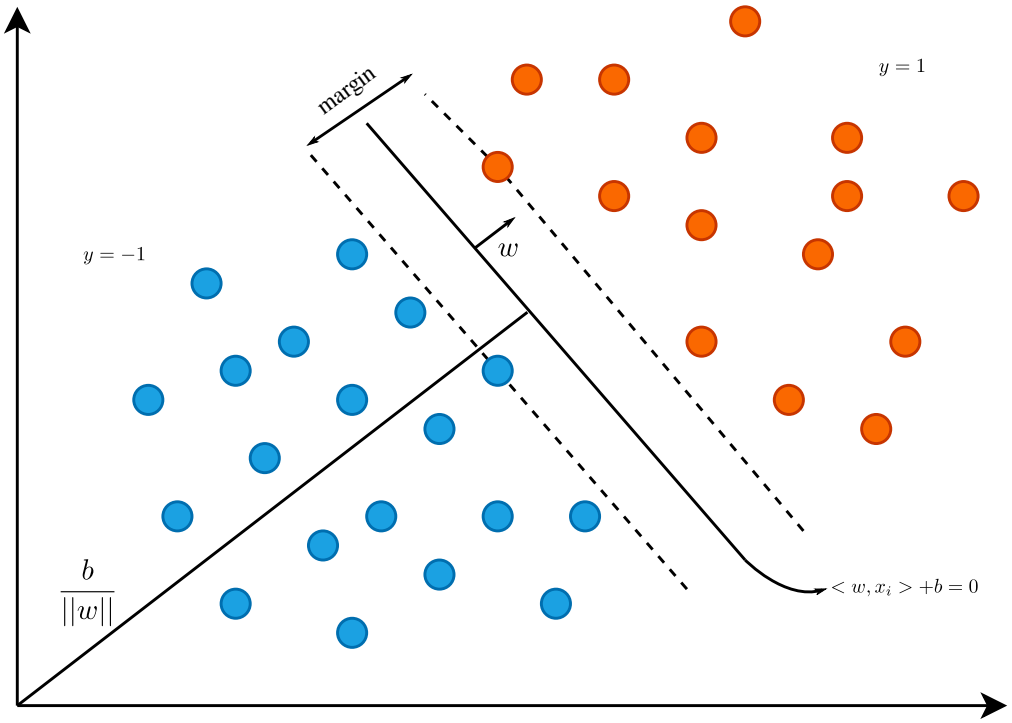


Figure 2.23. Hard-Margin SVM

For a set of data $(x_1, y_1), (x_2, y_2), (x_3, y_3), ...(x_i, y_i)$, with i being the number of subjects, we can linearly separate them with the following inequalities that will be two margins:

$$< w, x_i > +b \geq 1 \text{ if } y_i = 1 \tag{2.14}$$

$$< w, x_i > +b \leq -1 \text{ if } y_i = -1 \tag{2.15}$$

That can be grouped in inequality 2.16:

$$y_i(< w, x_i > +b) \geq 1, \forall i \tag{2.16}$$

Suppose we have two elements, x_a over one margin and x_b over the other as shown in Figure 2.24. These elements over each margin are called support vectors. If we want to maximize the distance between 2.14 and 2.15, we should subtract the equations to find the distance between the two:

$$(< w, x_a > +b) - (< w, x_b > +b) = 2 \tag{2.17}$$

$$< w, x_a - x_b >= 2$$

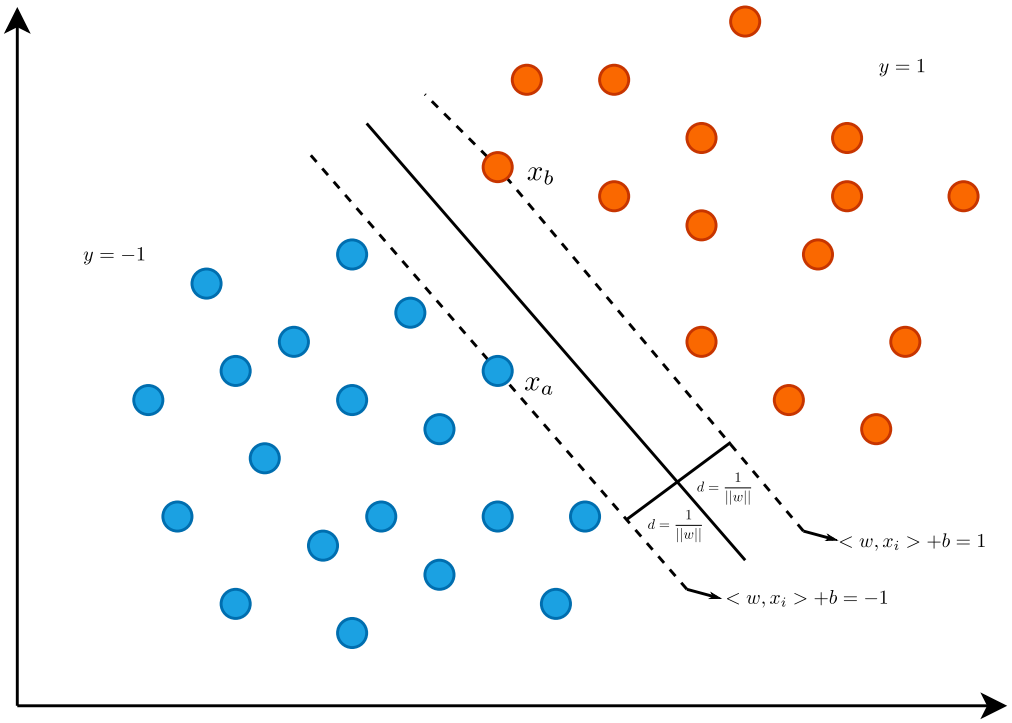


Figure 2.24. Support vectors x_a and x_b

The dot product between two vectors can be represented as the product of their magnitudes by the cosine of the angles between them.

$$|w||(x_a - x_b)|cos(\theta) = 2$$

From the Figure 2.24 we notice that the distance d between this two elements can be express as $|(x_a - x_b)|cos(\theta) = d$, rewriting the equations we have that:

$$d = \frac{2}{|w|}$$

This means that if our goal is to find the best hyperplane, which means the one where the distance of the points to this hyperplane is the maximum, we need to maximize the previous equation, or:

$$\min_w \frac{|w|^2}{2}$$

Subject to $y_i(< w, x_i > +b) \geq 1$. This optimization problem can be solved using Lagrange's multipliers, creating a new equation:

$$L_p(w, b, \lambda_i) = \frac{|w|}{2} - \left[\sum_{i=1}^{N} \lambda_i (y_i(< w, x_i > +b)) - \sum_{i=1}^{N} \lambda_i\right] \tag{2.18}$$

With $\lambda_i \geq 0$ and $i = 1, 2, 3, ..., N$. The optimal points can be found by derivating the 2.18 equation for w and b and making it equal to 0 as can be seen in 2.19 and 2.20.

$$\frac{\partial L_p}{\partial w} = w - \sum_{i=1}^{N} \lambda_i y_i x_i = 0 \tag{2.19}$$

$$\frac{\partial L_p}{\partial b} = -\sum_{i=1}^{N} \lambda_i y_i = 0 \tag{2.20}$$

Because the optimization problem is subject to an inequation, we need to make sure that our best hyperplane fulfills the Karush-Kuhin-Tucker (KKT) conditions that can be expressed as:

1. Primal feasibility or primal constraint

$$y_i(< w, x_i > +b) - 1 \geq 0$$

2. Stationarity

$$\nabla L_p = 0$$

$$\frac{\partial L_p}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{N} \lambda_i y_i x_i$$

$$\frac{\partial L_p}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N} \lambda_i y_i = 0$$

3. Complementary slackness

$$\lambda_i \left[y_i (< w, x_i > + b) - 1 \right] = 0, \forall i$$

4. Dual feasibility

$$\lambda_i \geq 0, \forall i$$

Replacing all these restrictions in 2.18, we get a Lagrange Dual problem:

$$L_p(w, b, \lambda_i) = \sum_{i=1}^{N} \lambda_i - \left[\sum_{i=1}^{N} \lambda_i \left(y_i \left(\sum_{j=1}^{N} \lambda_j y_j x_j \right) \cdot x_i + b \right) \right] + \frac{1}{2} \left(\sum_{i=1}^{N} \lambda_i y_i x_i \right)^T \left(\sum_{j=1}^{N} \lambda_j y_j x_j \right)$$

$$L_D(w, b, \lambda_i) = \sum_{i=1}^{N} \lambda_i - \left[\underbrace{\sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j < x_i, x_j >}_{A} + \underbrace{\sum_{i=1}^{N} b \lambda_i y_i}_{B} \right] + \underbrace{\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j < x_i, x_j >}_{C} \tag{2.21}$$

From the 2^{nd} condition of KKT we noticed that $B = 0$ and $C = \frac{1}{2}A$, so replacing we are left with:

$$L_D(w, b, \lambda_i) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j < x_i, x_j >$$

2.4.2 Soft Margin Support Vector Machine

For the soft margin, we are working with a similar problem as the hard margin, but we need to add a threshold or a tolerance; this tolerance allows data that is outside the margin to be included and added, as shown in Figure 2.25 to Equation 2.14 as follows:

$$yi(< w, x_i > +b) \geq 1 - \xi_i, \forall i \tag{2.22}$$

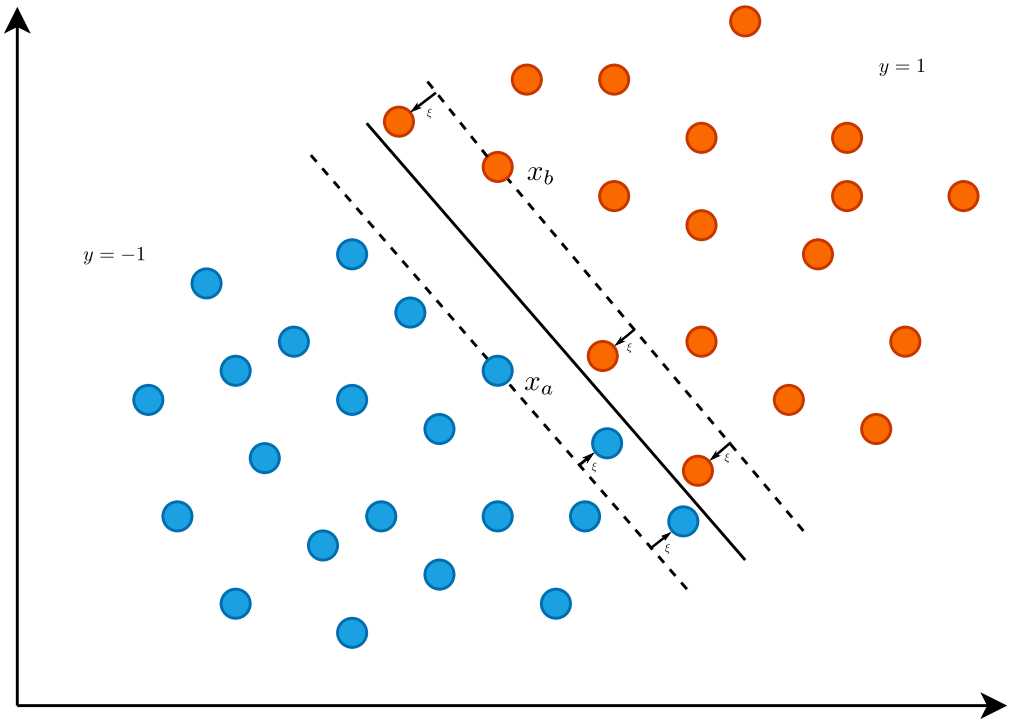


Figure 2.25. Soft-Margin SVM

This introduces a new set of values to optimize, modifying the equation to:

$$\min_{w,\xi} \frac{|w|^2}{2} + \sum_{i=1}^{N} \xi_i \tag{2.23}$$

Subject to $y_i(< w, x_i > +b) \geq 1 - \xi_i$. Because all values of ξ_i need to be greater than zero (values outside of the margins), we need to add a new restriction:

$$\xi_i \geq 0$$

We add a new variable, C, used as a parameter of the importance of this new tolerance ξ_i. This new parameter also modifies the 4^{th} KKT condition to $\lambda \geq 0$ to $0 \geq \lambda_i \geq C$.

The parameter C represents how many errors we accept in return for a more general classifier:

- A high value of C means an SVM with hard margin.
- A value of C close to 0 creates a broader margin in return for errors in the classification
- A value of $C = 0$ generates a hyperplane that doesn't classify

2.4.3 Kernel trick

All the previous mathematics was done, supposing the data was linearly separable. Still, in the case of data that cannot be separated linearly, we need to find ways of using the SVM for classification. To solve this, the data can be transformed into other dimensions that can probably be more separable.

An example of the advantages of the kernel trick can be seen in Figure 2.26

This transformation is feasible in the case of a few samples, but the difficulty of this transformation increases with the increments of the samples.

A trick can be used to solve this problem. If we take just one sample $< x_i, x_j >$, we can use a function $\mathcal{K}(x_i, x_j)$ that takes the points and finds the result of the operation in this new space. With this, we don't need to transform all the data into the new space; we see the result of the operation in the new dimension.

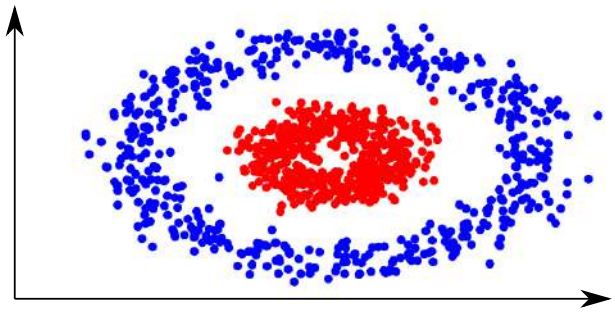
The kernel function for a linear kernel can be represented as:

$$\mathcal{K}(x_i, x_j) =< x_i, x_j >= x_i \cdot x_j$$

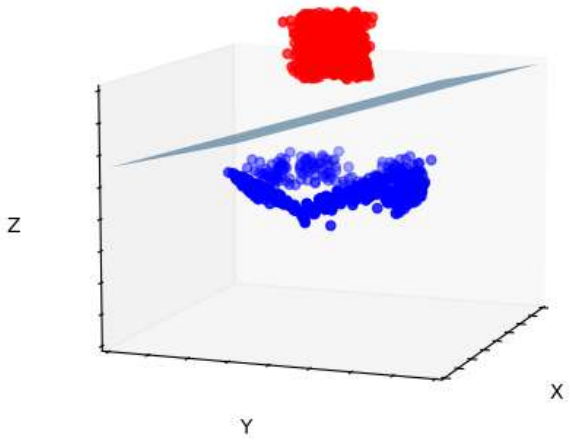
Another commonly used kernel is the Radial Basis Function (rbf) kernel:

$$\mathcal{K}(x_i, x_j) = e^{-\gamma|x_i - x_j|^2}$$

This last kernel type introduces a new parameter γ; this controls the behavior of the hyperplane, and a small value of γ makes the model behave as a linear model.



(a) Data non-linearly separable



(b) Hyperplane after kernel trick

Figure 2.26. Kernel trick example

2.4.4 Decision tree

Decision trees are based on constructing trees that can be used to group the data; the main focus is optimizing the generalization error. These groups can be done recursively, grouping the data in partitions with the same label.

The trees are based on impurity criteria between the features to know which attribute is the best for the split. The best "score" attribute is commonly used for the split criteria. After the split is done, the idea is to reduce the impurity; thus, we need to find a tree that always performs this reduction in the impurity, making it another optimization problem.

For this problem, suppose we have a feature vector x_i representing the features of the subject i with a label y_i.

Suppose we have a node t, which we represent as $\mathcal{N}_t$. This node will be divided into two parts. For this division, a feature is selected, and a range of this feature is used to know the side a value will belong if we define a pair $\beta = (i, th_t)$ where i is the node and th_t is the threshold defined for this node, then we can say that the values for two subsequent nodes will be:

$$\mathcal{N}_t^{left} = \{(x, y) | x_i \leq th_t\}$$
$$\mathcal{N}_t^{right} = \mathcal{N}_t / \mathcal{N}_t^{left}$$

Nevertheless, we need to select the features that will be used and the order; for this, we use the impurity criteria defined as I, the number of samples defined as S and C as the variable to optimize, such as:

$$C(\mathcal{N}_t, \beta) = \frac{S_t^{left}}{S_t} I(\mathcal{N}_t^{left}(\beta)) + \frac{S_t^{right}}{S_t} I(\mathcal{N}_t^{right}(\beta))$$

At the end, the optimal candidate β will be:

$$\beta = \min C(\mathcal{N}_t, \beta)$$

The algorithm continues to generate splits or branches until one of the following conditions is fulfilled:

- One sample is obtained at the end.
- The value of samples in a node is less than the minimum value of samples allowed.
- The maximum depth is reached.

For the decision trees, a maximum depth and a minimum number of samples in a node are recommended so the algorithm doesn't generate too many ramifications to solve the data's variability; this is called "pruning" and avoids overfitting when the model is being created.

2.4.5 Random forest

Similarly to decision trees, random forest is a classifier that uses impurity criteria to create branches and trees. In the case of the random forest, multiple decision trees are created, each performing the classification process. In the end, the final prediction is decided either by averaging the decisions of each tree or by the majority of the votes.

This combination of decision trees allows the random forest to be a relatively good classifier against overfitting problems and works well with high amounts of data. However, it increases the complexity of a decision tree, and we lose interpretability.

In this work, we used Information Gain (IG) and the Gini Index (GI) in decision trees and random forests out of all impurity criteria.

Information Gain

To apply this impurity test, we first need to talk about entropy. The entropy measures the impurity and randomness within a dataset; it ranges between 0 and 1, being 0 in a pure homogenous dataset [74]. Entropy E, can be calculated by adding the probability to obtain each possible event multiplied by the base 2 logarithms of the same probability as shown in 2.24

$$\text{Entropy} = -\sum_{i=1}^{N} p_i log_2(p_i) \tag{2.24}$$

Information can then be defined as the difference between the entropy of one class and the conditional entropy of one class with a set of features [75]

$$IG(x) = E(D) - E(x) \tag{2.25}$$

With D being the dataset, x the feature and E being the entropy.

Gini Index

It is considered another impurity measurement and is related to the Gini value, which is the probability of two samples having labels different from their original labels [75]. This Gini value can be calculated as shown in Equation 2.26:

$$\text{Gini value} = \sum_{i=1}^{N} \sum_{j \neq i} p_i p_j = 1 - \sum_{i=1}^{N} p_i^2 \tag{2.26}$$

Being p_i, the probability of the label or event i being in the dataset. With this, we can define the Gini index as the impurity of a subset S^n vs the dataset, as shown in Equation 2.27.

$$\text{Gini index} = \sum_{n=1}^{N} \frac{S^n}{D} Gini(S^n) \tag{2.27}$$

2.4.6 Deep Neural Network

Deep neural networks (DNNs) or feed-forward networks are architectures that allow computers to learn from patterns found in data they have seen before. These networks are inspired by how our brains work. The most basic element of a neural network is the neuron, a unit that receives one or more inputs, applies a mathematical operation, and gets an output.

A neuron in a DNN assigns a weight to each input, and then these weighted inputs are summed along a term called bias; this bias acts as a threshold or a shift. Afterward, this operation's result is passed through an activation function such as *sigmoid, tanh or ReLU* that induces non-linearity. Equation 2.28 shows the output of a neural network where l is the layer number, m is the number of units in the layer, w is the weight, b is the bias, g is the activation function, and x and y are the input and the output of each layer, respectively. This same concept can be seen in Figure 2.27.

$$\begin{aligned} z_i^{(l)} &= \left(w_{i,1}^{(l)} x_1^{(l)} + w_{i,2}^{(l)} x_2^{(l)} + ... + w_{i,m}^{(l)} x_m^{(l)} + b_i^{(l)} \right) \\ y^{(l)} &= g\left(z_i^{(l)} \right) = g\left(\mathbf{w}_i^{(l)} \mathbf{x}^{(l)} + b_i^{(l)} \right) \end{aligned} \tag{2.28}$$

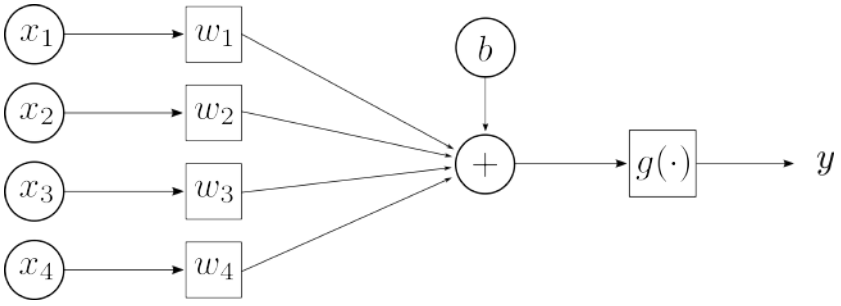


Figure 2.27. Graphical description of neuron with 4 inputs

The weights and biases can start randomly, from zero or an initialization criteria. Then, after each iteration, they will be updated to minimize a loss function using an optimization algorithm, allowing the neuron to learn patterns to classify the inputs.

However, one network is not enough; more complex data requires combinations of neurons. This is when DNNs are used; these architectures interconnect multiple neurons in a chained structure where the output of one function is the input of the next one, creating a more robust model. The DNNs are based on an input, hidden, and output layer, as shown in Figure 2.28. Hidden layers are the intermediate layers of the network and consist of a set of neurons, where all the outputs are inputs to the next layer [76].

DNNs are highly customizable because their architecture is based on the number of inputs, hidden layers, neurons inside the hidden layers, interconnectedness, and outputs.

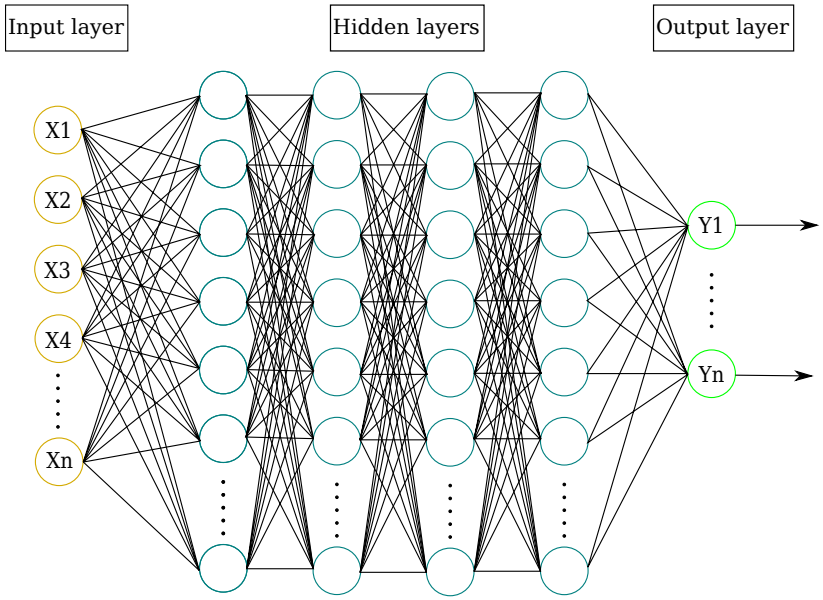


Figure 2.28. Visual representation of fully connected neural network

Loss function

When training our models, we aim to obtain the combination of parameters that yields the best results for our data. This "best results" statement means there needs to be a way to compare multiple results to find the "best." This is done using a loss or cost function, which measures the differences between the predicted values and targets. A loss function is selected depending on the task, classification, or regression; the loss value calculated is the one we aim to minimize by updating the weights and biases in the network [77]. In the case of classification, the cross-entropy function is the most used because it compares the probability distributions of the actual targets versus the predicted distribution and applies a logarithmic penalization, which is more drastic than a linear one. Equation 2.29 shows the cross-entropy loss, where N is the number of samples in training, $\mathbf{y}$ is the vector of actual targets and $\dot{\mathbf{y}}$

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[ylog(\dot{y}) + (1-y)log(1-\dot{y})] \tag{2.29}$$

Backpropagation

Backpropagation is a method used in neural networks to update the weights and biases based on the values in the loss function after each feedforward pass. Adjusting the parameters is the network's "learning" process, making this the core of neural networks.

As mentioned before, the backpropagation is applied after a feedforward. In the end, the lost function is calculated, which gives us an idea of how much we need to adjust the parameters to reduce the loss. To update the parameters, we need to find how much the change of weight or bias will affect the cost function. This can be calculated using the cost function's derivative with respect to the influence of weight. Because the cost function doesn't have any direct relation with the weights in Equation 2.29, we applied the chain rule when doing the derivative as shown in Equation 2.30

$$\frac{\partial \mathcal{L}}{\partial w^{(l)}} = \frac{\partial \mathcal{L}}{\partial \dot{y}} \frac{\partial \dot{y}}{\partial z^{(l)}} \frac{\partial z^{(l)}}{\partial w^{(l)}} \tag{2.30}$$

Three derivatives appear with the chain rule:

- The first one represents the influence of the output on the loss function. This partial derivative depends on the loss function selected.
- The second one shows how much the output changes depending on the activation function selected; this partial derivative consists of derivating the activation function. This adds the constraint of being derivable to the activation functions.
- The last one represents how the activation function is affected by the change of the weights. This consists on derivating $\boldsymbol{w}_i^{(l)}\boldsymbol{x}^{(l)} + b_i^{(l)}$

2.4.7 Convolutional Neural Network

Convolutional neural networks are feed-forward networks widely employed in computer vision and video recognition. Like DNNs, CNNs are based on neurons and are self-optimized using the knowledge learned from the data [78]. CNN's main advantage is that it can extract features from raw images.

This network is based on how humans use our visual organs or perceptions. Because CNN focuses mainly on images, the neurons must be adapted

for the inputs. For this, multiple layers are interconnected; these layers are the so-called kernels.

The input image goes through those kernels where image segments are convolved with the kernels following Equation 2.31, where S is an image segment and K is the kernel. Note that the kernel has a size of nxn in this case, but this size is considered a hyper-parameter, so it can be optimized based on the problem. A bigger kernel generates fewer operations, but smaller details can be overlooked. This convolution operation is the core operation of the CNN as it creates feature maps as shown in Figure 2.29. These features highlight borders or big contrasts in the image on the first layer and more abstract features in the following layers [79].

$$S * K(x,y) = \sum_{i=0}^{n}\sum_{j=0}^{n} \boldsymbol{S}(x+i, y+j) \cdot \boldsymbol{K}(i,j) \tag{2.31}$$

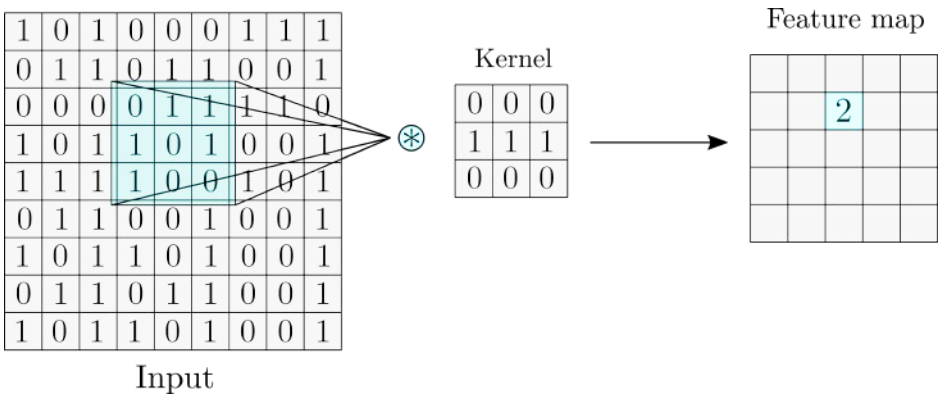


Figure 2.29. Convolutional operation between inputs and kernel

To generate the remaining values of the feature map, the kernel slides into different segments to perform the convolutional operation, and the number of pixels by which the kernel is moved is called stride. Sometimes, because of the image, segment, or kernel size, the operation cannot be performed in the borders because of missing information. Padding is applied to avoid losing information; this operation includes zeros near the borders when necessary to complete segments. Figure 2.30 shows an example of an input image with zero padding of 1 and a stride of 2.

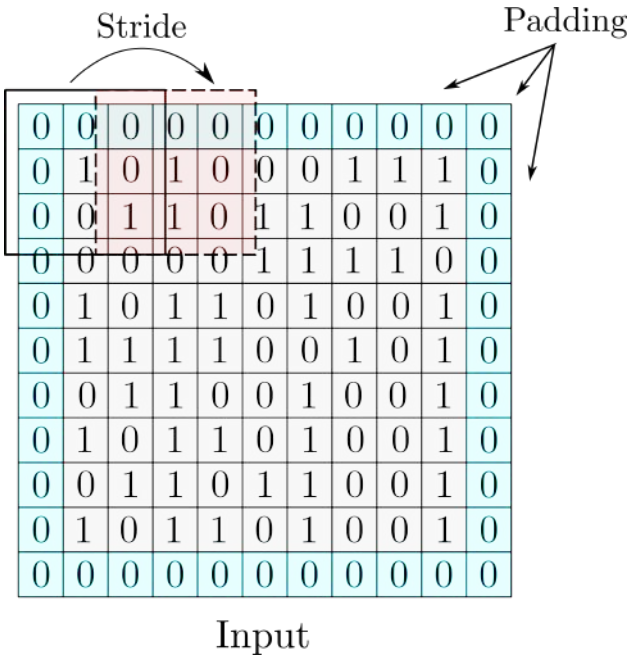


Figure 2.30. Input image with zero padding and a stride of 2

In the same fashion as DNN, CNN uses activation functions at the end of each convolutional layer. Section 2.4.6 shows the importance of this activation function. The most common activation function for the convolutional layers is the ReLU based on the good results compared to other activation functions like hyperbolic tangent [80].

Lastly, CNN's main focus is to reduce the image to a set of features that allow us to compare it with other images. The convolution operation is applied to each pixel so the remaining image will be the same size. For this, a pooling operation is done, reducing the dimensionality of the data in terms of width and height while keeping information. The two most common pooling operations are max pooling, which selects the maximum value from the convolution operation within a rectangle, and average pooling, which takes the mean of the resulting pixels. This pooling operation makes it more computationally efficient and allows the network to learn robust features against small variants in the input [76].

2.4.8 Regularization

A common problem that happens not only in deep learning but also in machine learning is how to make the algorithm perform better for samples that

the model hasn't seen. Usually, these models generate great results when training but low performance for new inputs. This commonly happens because the model fits too much to the training data, reducing his generalization capabilities. This phenomenon is called overfitting; different strategies have been created to reduce errors during tests, even if it means increasing errors during training [76]. Dropout and early stopping are two methods used in this work for the regularization process.

Dropout

The multiple connections between neurons in a DNN make this architecture good when learning difficult patterns from the data. Increasing the number of connections can improve the performance of the network but at the same time can lead to a more expensive model, computational speaking, and also a model prune to overfit [81].

Dropout is a regularization method used to reduce the model's complexity and the risk of overfitting; this is done by dropping some neurons from the network based on fixed probability [82]. Dropping a neuron means temporarily removing all its input and output connections; this process forces the network to learn more general representations of the data and avoid co-adapting some neurons to certain samples.

Equation 2.28 is modified when using dropout. Now, the output vector of a layer is multiplied by a binary mask from a Bernoulli distribution as shown in Equation 2.32, generating a new output vector where some outputs will be 0, similarly as if the neuron was "turned off".

Dropouts have not also shown good results in feed-forward networks but in probabilistic models or other types of networks like recurrent neural networks [76] or CNNs [83].

$$\begin{aligned} \boldsymbol{d}^{(l)} &\sim Bernouli(p) \\ \hat{\boldsymbol{x}}^{(l)} &= \boldsymbol{d}^{(l)} \odot \boldsymbol{x}^{(l)} \\ \hat{y}_i^{(l)} = g\left(\hat{z}_i^{(l)}\right) &= g\left(\boldsymbol{w}_i^{(l)} \hat{\boldsymbol{x}}^{(l)} + b_i^{(l)}\right) \end{aligned} \tag{2.32}$$

Early stopping

Most neural networks update their internal parameters, such as the weight and biases, to minimize the loss function. This process is done iteratively in

epochs, where an epoch is a complete pass from the training set. After one run, we can test the network with the current parameters, so two loss values are obtained at the end, one for the training and the test set.

Sometimes, the test or validation loss can increase after some iterations while the training loss keeps decreasing. This means that the model is adapting well to the information from training but cannot correctly classify new data, which is the concept of overfitting.

If the validation loss keeps increasing, that means that at one point, this loss was at its lowest peak, meaning that we can obtain a model with a better validation performance if we "stop" and use the parameters at that point [76]. This process is called early stopping and consists of monitoring the validation loss during training. Let's assume $\mathcal{L}_{val}(n)$ as the validation loss calculated in epoch n; the early stopping algorithm consists of stopping the training process when $\mathcal{L}_{val}(n) \geq \mathcal{L}_{val}(n-1)$, but because sometimes there are fluctuations in the loss values because of the iterative process of finding the best parameters a patience p is defined as a wait criterion. Now, the algorithm will stop training if the validation loss of the n^{th} epoch is higher or equal to previous epochs based on the patient value, as shown in Equation 2.33

$$\mathcal{L}_{val}(n) \geq max\{\mathcal{L}_{val}(n-1), ..., \mathcal{L}_{val}(n-1-p)\} \tag{2.33}$$

2.5 Multi-modal approach

Modality fusion aims to combine and include information from two different signals that capture the same phenomena. This approach has gained considerable attention in the research area due to the multiple benefits of combining different signals. The common method is to take a signal that describes a phenomenon and extract information from it. This information is called features, and its main use is to create a representation of this sample. The set of features is then introduced to a classifier, which gives us a discrete result if we are working with classes (i.e., 1, 2, HC, VP) or a continuous result in the case of a regression problem.

Combining different modalities can give classifiers new information that can be complementary and improve the results obtained [84]. However, it is crucial to note that analyzing multiple modalities comes with challenges depending on the topic and the signals. Synchronization is one of them;

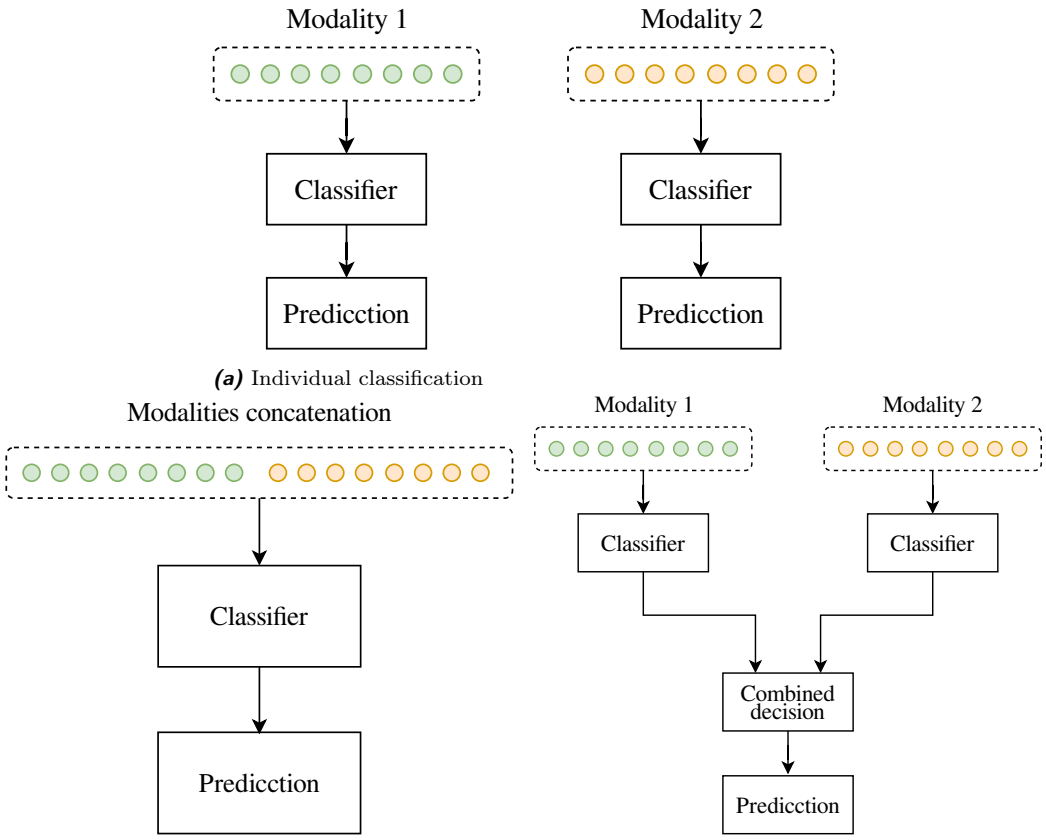
capturing different signals usually comes with different rates [85], more so when the signals are not captured simultaneously; this causes an alignment to be required to combine both signals correctly.

Another challenge is the cost associated with the acquisition of multiple signals, involving financial and computational costs [85]. Therefore, an analysis of the available information is essential to prevent investments in expensive equipment for issues that could be addressed more cost-effectively. For example, in the context of medical image screening, Krones et al. [86] identify various imaging modalities, including X-rays, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Ultrasound. Each of these methods presents distinct drawbacks, such as high costs (as seen with MRI) and potential health risks associated with radiation exposure (as in the case of CT). Given that each imaging modality typically requires separate procedures, the cumulative costs can escalate significantly. Additionally, the synchronization of signals requires special equipment's that can increase the cost. In contrast, integrating different forms of information, such as time-series, tabular, or text data, can provide complementary insights at a lower overall cost.

Lastly, the question of "how to fuse?" remains. Multiple techniques have been developed with some variations, but they depend on the types of signals and the way they are obtained. We can use two types of fusion, a feature level fusion (early fusion) or a more decision level fusion (late fusion) [87] as shown in Figure 2.31.

The researchers have already solved most of these challenges for the SVD database. Both signals (EGG and speech) were recorded simultaneously, removing the synchronization problem. The signals recorded analyze the same phenomena but from a different point of view. EGG analyzes the glottal opening. Meanwhile, speech measures sound pressures a microphone captures, making them worth the evaluation and analysis.

A systematic analysis is done on "how to fuse?" Both early and late fusion are tested to see which one is best for classifying a healthy subject as a subject with a voice pathology.



(a) Individual classification

(b) Early fusion

(c) Late fusion

Figure 2.31. Different approaches using two modalities

2.6 Performance metrics

As mentioned, multiple metrics are commonly used to evaluate the models' performance. These metrics give a numerical value to the model's performance. The accuracy is the most commonly used, usually represented in percentages, and provides us with a notion of the correctly classified classes. However, accuracy cannot determine whether the model performs well, leading to misleading results. The other performance metrics used in this work are described in the following subsections.

2.6.1 Confusion matrix

It is one of the most popular performance metrics used in classification problems. It can be used both in a binary or a multi-label classification problem; for the case of a binary problem, it is based on a two by two matrix that compares the values predicted with the actual values. It takes each prediction and counts the number of values that fall into the following categories:

- **True Negative (TN):** Number of negative samples that were correctly classified as the negative class
- **True Positive (TP):** Number of positive samples that were correctly classified as the positive class
- **False Negative (FN):** Number of positive samples that were incorrectly classified as the negative class
- **False Positive (FP):** Number of negative samples that were incorrectly classified as the positive class

Table 2.1 shows an example of the confusion matrix. From this matrix, multiple metrics can be extracted.

Table 2.1. Example of a confusion matrix

		Predicted values	
		0	1
Original values	0	TN	FP
	1	FN	TP

2.6.2 Accuracy

It is considered one of the main criteria or at least one of the most used criteria to define the success or failure of a machine or deep learning model. Accuracy measures correctly classified or identified subjects out of all the predictions, which can be calculated by checking each prediction one by one, summing all the correct values and dividing by the total amount, or using the results of the confusion matrix as shown in Equation 2.34

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.34}$$

Accuracy tends to be a good metric when dealing with balanced classes. However, using other metrics alongside accuracy can be beneficial in the case of imbalances.

2.6.3 Sensitivity

Equation 2.35 shows the sensitivity, also known as recall or True Positive Rate (TPR). It is a metric that shows the ratio of True Positive out of all possible positive values, which is the ability of the classifier to find the positive class when predicting. Table 2.2 shows the values that will be used from the confusion matrix to calculate the sensitivity.

$$Sensitivity\ or\ recall = \frac{TP}{TP + FN} \tag{2.35}$$

Table 2.2. Values in confusion matrix for the sensitivity

		Predicted values	
		0	1
Original values	0	TN	FP
	1	FN	TP

2.6.4 Specificity

Equation 2.36 shows the specificity or True Negative Rate (TNR). Similar to sensitivity, specificity calculates the number of negative classes correctly classified out of all possible negative samples. This metric shows the model's

capability to predict the negative class. Table 2.3 shows the values that will be used from the confusion matrix to calculate the sensitivity

$$Specificity = \frac{TN}{TN + FP} \tag{2.36}$$

Table 2.3. Values in confusion matrix for the specificity

		Predicted values	
		0	1
Original values	0	TN	FP
	1	FN	TP

2.6.5 F1-Score

Equation 2.37 shows the F1-score, also known as the harmonic mean of precision and recall. This metric is similar to the accuracy, but because it uses recall and precision, it is better when dealing with imbalanced class datasets, which makes it a more robust metric to analyze the performance of a machine learning model. For this metric, all elements in the confusion matrix are used.

$$F1\text{-}Score = 2 * \frac{\frac{TP}{TP+FP} * \frac{TP}{TP+FN}}{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}} \tag{2.37}$$

Chapter 3

Data

The dataset that will be used is a public dataset called the Saarbruecken Voice Database. This German database has information on patients with a voice pathology and healthy controls performing tasks such as sustained vowels /a/, /i/, and /u/ for normal, high, low, and low-high pitch levels and a small phrase.

A basic analysis of the signals is performed. Table 3.1 shows general information about the subjects in the dataset, such as age or gender; also Table 3.2 shows all the voice pathologies that are present in the SVD dataset, as well as the number of subjects in each pathology. Still, the classes are imbalanced, and the table combines medical, voice, and speech diagnoses. The SVD, a public dataset, does not contain much information on how the clinical data was extracted. Still, the work mainly focuses on the automatic analysis with different biomarkers extracted from the signals, considering all these pathologies as one class so that the classification will be between healthy and pathological.

Table 3.1. Demographic information of subjects in SVD database.

	Subjects with a pathology		**Healthy subjects**	
	Male	Female	Male	Female
# of participants	681	839	201	348
Age [years]	53.08 ± 15.23	48.77 ± 15.33	32.31 ± 12.79	25.95 ± 11.97
Age range [years]	6 - 89	9 - 94	16 - 69	9 - 84

Table 3.2. Number of subjects for each pathology in SVD dataset

Subjects in each pathology			
Pathology	Subjects	Pathology	Subjects
Amyotrophic Lateral Sclerosis	2	Leukoplakia	41
Aryluxation	6	Medial Neck Cyst	1
Stuttering	20	Mesopharyngeal Tumor	1
Bulbar Paralysis	2	Monochorditis	3
Carcinoma in situ	1	Down Syndrome	1
Chondroma	1	Parkinson's Disease	1
Chordectomy	59	Mutation	2
Cyst	6	Mutational Fistula Voice	10
Diplophonia	5	Superior Laryngeal Nerve Lesion	3
Dish Syndrome	1	Superior Laryngeal Nerve Neuralgia	3
Dysarthrophonia	19	Non-fluency Syndrome	2
Dysodic	56	Orofacial Dyspraxia	1
Dysphonia	101	Papilloma	1
Dysplastic Dysphonia	1	Phonasthenia	10
Dysplastic Larynx	1	Phonation Nodules	17
Epiglottic Carcinoma	1	Polter Syndrome	3
Fibroma	2	Psychogenic Aphonia	1
Frontal Lateral Partial Resection	35	Psychogenic Dysphonia	91
Functional Dysphonia	112	Psychogenic Microphonia	1
GERD	3	Reinke's Edema	68
Singing Voice	17	Recurrent Nerve Paralysis	213
Granuloma	2	Open Rhinophonia	18
Hyperasthenia	1	Closed Rhinophonia	1
Hyperfunctional Dysphonia	213	Mixed Rhinophonia	1
Hypofunctional Dysphonia	16	Sigmatism	4
Hypopharyngeal Tumor	6	Spasmodic Dysphonia	64
Hypotonic Dysphonia	5	Vocal Cord Carcinoma	22
Internal Weakness	1	Vocal Cord Polyp	45
Intubation Granuloma	4	Synechia	2
Intubation Injury	3	Singer's Voice	2
Juvenile Dysphonia	1	Folds Hyperplasia	2
Laryngeal Tumor	5	Folds Voice	11
Contact Pachydermia	71	Velopharyngoplasty	2
Laryngitis	140	Senile Voice	40
Laryngeocele	3	Central Laryngeal Movement Disorder	14

When looking at the dataset, we notice a few imbalances. First, the amount of subjects per class is different (more than 200 subjects of difference); secondly, the number of female subjects is almost double that of the male subjects in the case of the healthy subject dataset; and lastly, the age distribution of the healthy subjects set is highly skewed to the left as shown in

Figure 3.1, where we can notice that there is a high number of subjects with ages in the range of 20s while the pathological dataset is more distributed, this imbalance can cause that the classifier creates a good separation between the two classes (healthy and pathological) but influence to the fact that most of the patterns are from younger voice and not from the actual disease.

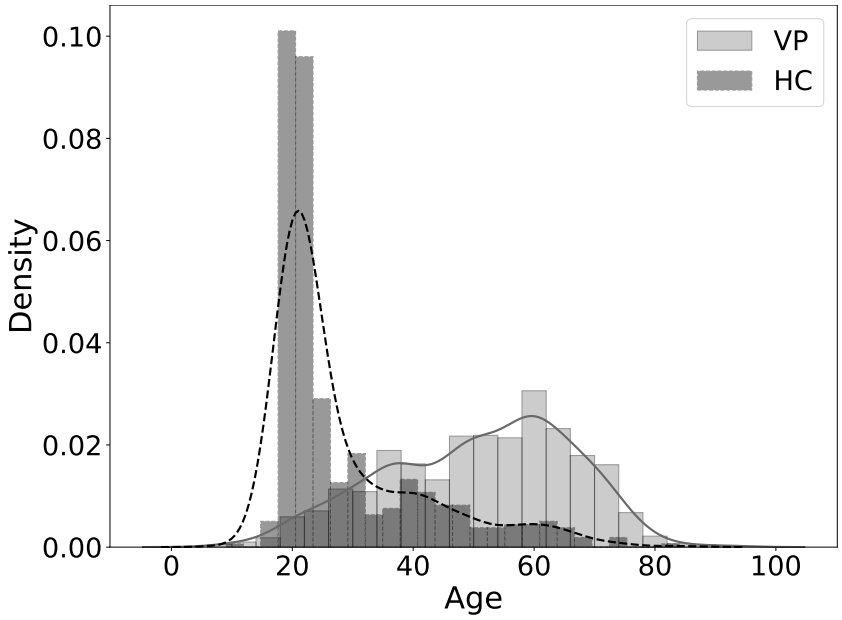


Figure 3.1. Age distribution

To solve those imbalances mentioned before, a subset of the database is selected; this subset will be balanced in age and gender. Subjects that don't have both EGG and speech signals are also removed. This will allow the comparison of multiple sets of features and ease the combination. The information of the final dataset after the balance can be seen in Table 3.3; also, Figure 3.2 shows the distribution of the ages in the data after the balance. We can notice that both groups follow a similar distribution. Finally, possible biases introduced by age or gender are discarded according to Welch's t-test ($p = 0.07$) and a chi-square test ($p = 0.89$), respectively.

Table 3.3. Demographic information of subjects in SVD database after balance in age and gender.

	Subjects with a pathology		**Healthy subjects**	
	Male	Female	Male	Female
# of participants	127	101	127	101
Age [years]	40.88 ± 12.11	41.75 ± 15.11	38.84 ± 11.92	39.82 ± 14.76
Age range [years]	18 - 67	18 - 79	24 - 69	24 - 84

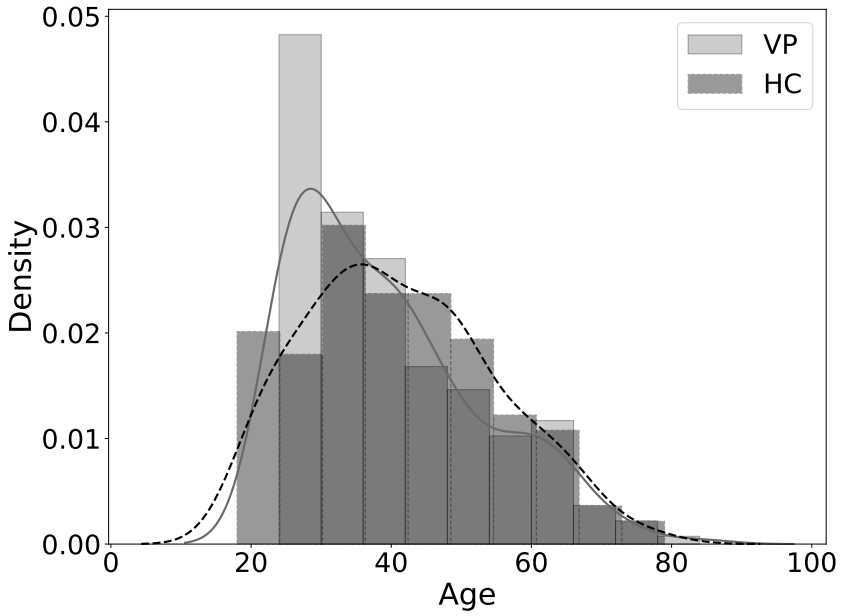


Figure 3.2. Age distribution balanced

Chapter 4

Experiments and results

This section compromises the results obtained during the development of the thesis. It's separated into three chapters; the first one shows the results obtained when we aim to classify a person with a voice pathology vs a healthy subject, and a comparison of the classification methods, features, and tasks is done. The second one shows how the classification changes when we combine the two modalities using different fusion methods. The third one compares the results obtained in uni-modality vs. multi-modality. Some parts of these results are published in [88].

Multiple experiments are performed to find the feature set that performs the best in classifying VP and HC. Each set of features is extracted from different tasks. These features are computed in the speech signal and the EGG signal. The classical approach uses SVM, DT, and RF as classifiers. All experiments were validated using a stratified k-fold cross-validation where 10 folds were selected; the stratified cross-validation tries to keep the same percentage of subjects for each class in the train and test set. A hyper-parameter optimization is performed for each classifier using a grid search, and the best set of parameters is selected in each fold based on the training accuracy.

The parameters optimized for the SVM were $C \in \{1e^{-5}, 1e^{-4}, ..., 10\}$, $\gamma \in \{1e^{-5}, 1e^{-4}, ..., 10\}$ and kernel $\in \{linear, rbf\}$. The parameters for the DT classifier were criterion $\in \{gini, entropy\}$, max depth $\in \{3, 5, 7, ..., 17\}$, min samples split $\in \{2, 4, 6, 8, 10\}$ and max leaf nodes $\in \{3, 5, 7, ..., 19\}$. Lastly, the parameters for the RF classifier were max depth $\in \{10, 20, 30, ...90\}$, min samples split $\in \{2, 4, 6, 8, 10\}$ and maximum features $\in \{log2, sqrt\}$. After selecting the best parameters, the model's performance is evaluated

using different performance metrics like accuracy, f1 - score, sensitivity, and specificity. More information on the metrics can be found in Section 2.6.

4.1 Uni-modal approach

4.1.1 Methodology

The methodology for the experiments can be seen in Figure 4.1. The experimental setup focuses on a multi-step process to assess the efficacy of classical and modern machine learning approaches in classifying signals for voice pathology detection. The methodology is divided into two key stages: feature extraction and classification.

The feature extraction process is crucial in this experiment, utilizing both EGG and speech signals. Different features are extracted from these signals and categorized into phonation, articulation, nonlinear, BFCC, and phase plots. Each type of feature serves a specific purpose in representing the signal characteristics that are vital for effective classification:

- Phonation features capture the vibratory characteristics of the vocal folds, which are essential in detecting abnormalities in vocal fold behavior, often associated with voice pathologies [89]–[91].
- Articulation features are critical for analyzing how sound is shaped by the vocal tract, which may reveal articulation issues that can indicate pathologies. [92]–[94].
- Nonlinear features help in identifying complex, chaotic behaviors in the voice signal that may not be captured by traditional linear methods, making them invaluable for detecting subtle variations in pathological voices [95]–[97].
- BFCC (Bark Frequency Cepstral Coefficients): This feature is closely related to MFCC but is more perceptually oriented, capturing how humans perceive speech sounds, which can lead to better characterization of voice disorders [98], [99].
- Phase plots represent a unique way of capturing dynamic system behaviors in the signal and are automatically extracted. This feature

allows for the analysis of the temporal evolution of the system, capturing additional nonlinear behaviors that other feature sets might miss [73], [100], [101].

Once the features are extracted, the next step involves classifying them using both classical and modern machine learning approaches. The classical approach includes algorithms such as Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (DT), which have been historically successful in voice pathology detection:

- Decision Trees provide an interpretable model that helps in understanding the decision-making process, making it suitable for medical applications where interpretability is important.
- Random Forest is an ensemble learning method combining multiple decision trees to improve classification accuracy and reduce overfitting, essential for complex datasets like voice signals.
- SVM is a robust classifier for high-dimensional spaces and is particularly effective when there is a clear margin of separation between classes (healthy vs pathological).

On the other hand, current state-of-the-art applications are based on deep learning in the form of neural networks, which learn complex feature representations from raw data. This approach tends to outperform classical methods when large amounts of data are available due to its ability to capture patterns within the data.

For the optimization process, a grid search technique is used to find the best hyper-parameters for each classifier to ensure the best performance. The idea here is that various combinations of parameters are tried out at every fold in multiple folds of cross-validation, resulting in a very robust model selection. Train a model for each fold and select the parameters that give the best performance by a voting mechanism across all folds.

The classifiers are then retrained on the whole training set, whose performance is evaluated on the test set after the best parameters are chosen. The average mean and standard deviation over all folds are reported as the evaluation metrics.

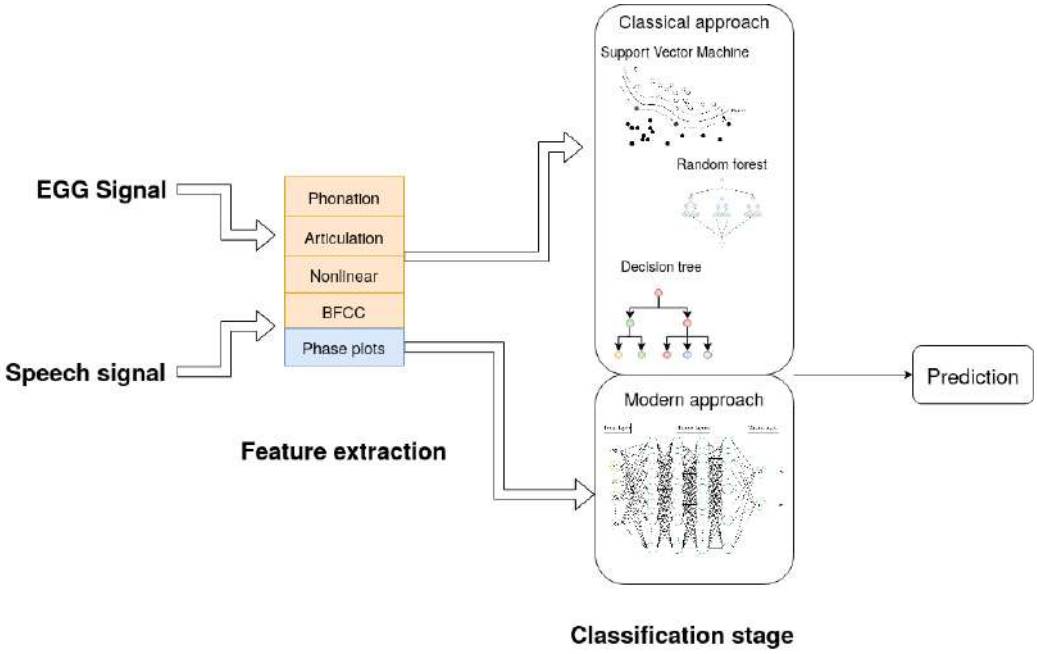


Figure 4.1. Base methodology for the uni-modal experiments

4.1.2 Experiments and results

Imbalance influence

Before performing the experiments shown in the methodology, it is important to showcase the influence of the imbalances. The main focus of the work is to test a methodology ensuring correct experimentation. One of the drawbacks mentioned in the state-of-the-art is the imbalances in the database that can induce biases in the model. To test this hypothesis we performed a simple experiment, we trained a model using the whole database. The features, models and task selection are not to much relevant to the experiment because we want to find whether the imbalances affect the results of the model, however we train an SVM with BFCC features, focusing on the vowel /a/ and the speech task.

Since the objective is to observe how the model performs with specific populations, we extracted a subset of the data, based on a condition such as fixed gender or a specific age range. This approach helps us analyze the model's behavior under controlled conditions and assess whether the imbalances affect its performance across different groups.

The results obtained with the whole dataset are shown in Table 4.1:

Table 4.1. Accuracies of the model when trained with the whole dataset and selecting different types of population

Population	ACC	F1-Score	Specificity	Sensitivity
Male	80.0%	0.80	0.65	0.85
Female	85.4%	0.85	0.86	0.84
Young (<24)	69.5%	0.59	0.97	0.04
Middle-age (24 – 40)	78.4%	0.75	0.26	0.94
Old (>40)	95.2%	0.92	0.0	1.0

On the other hand, when the database is balanced, we obtain the results shown in Table 4.2:

Table 4.2. Accuracies of the model when trained with the balanced dataset and selecting different types of population

Population	ACC	F1 Score	Specificity	Sensitivity
Male	60.0%	0.59	0.71	0.48
Female	54.1%	0.51	0.80	0.30
Young (<24)	53.8%	0.53	0.47	0.61
Middle-age (24 – 40)	55.1%	0.54	0.44	0.66
Old (>40)	60.7%	0.60	0.81	0.47

As shown in the table, when we use the entire database, we observe high accuracies across the age ranges. However, many of these models demonstrate either high sensitivity or high specificity, indicating that the models tend to classify most samples into one class. This occurs due to the imbalances present in the database for certain age ranges. For instance, in the case of younger individuals, the majority of samples are from healthy subjects, which results in high specificity (most samples are classified as healthy) and consequently high accuracy.

In contrast, when using the balanced database, the overall accuracies are lower, but the sensitivity and specificity values are more balanced. This indicates that the model is making an effort to correctly identify both classes, rather than favoring one, leading to a more reliable and fair classification. This shows that the imbalance in the database can introduce biases in the model, helping the model to achieve higher accuracies.

Graphs of the age vs the accuracies are shown in Figure 4.2 and Figure 4.3

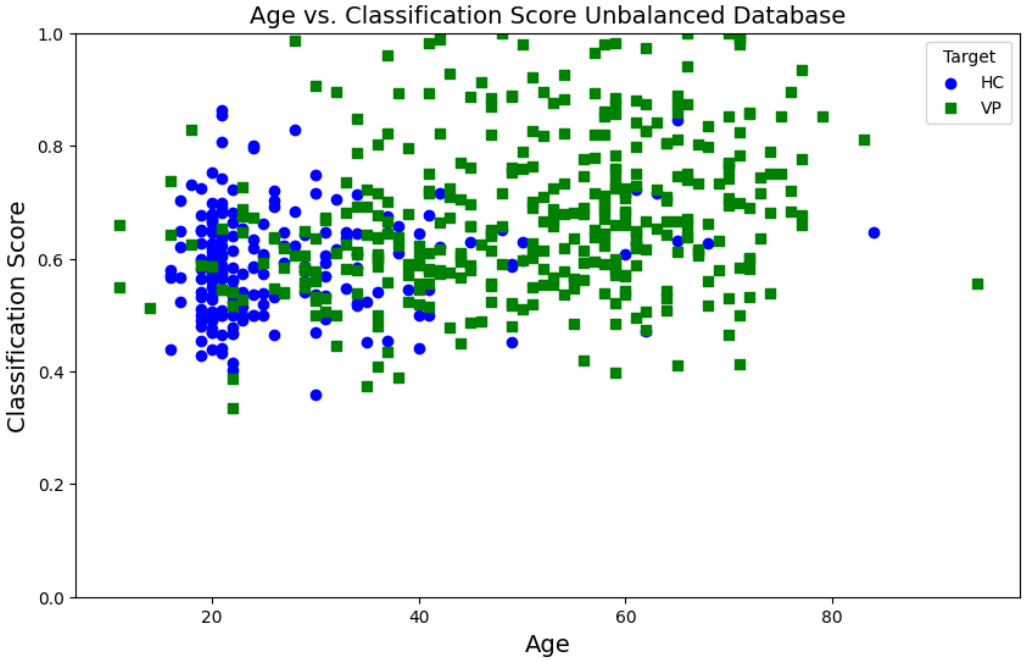


Figure 4.2. Age vs classification score with the unbalanced database

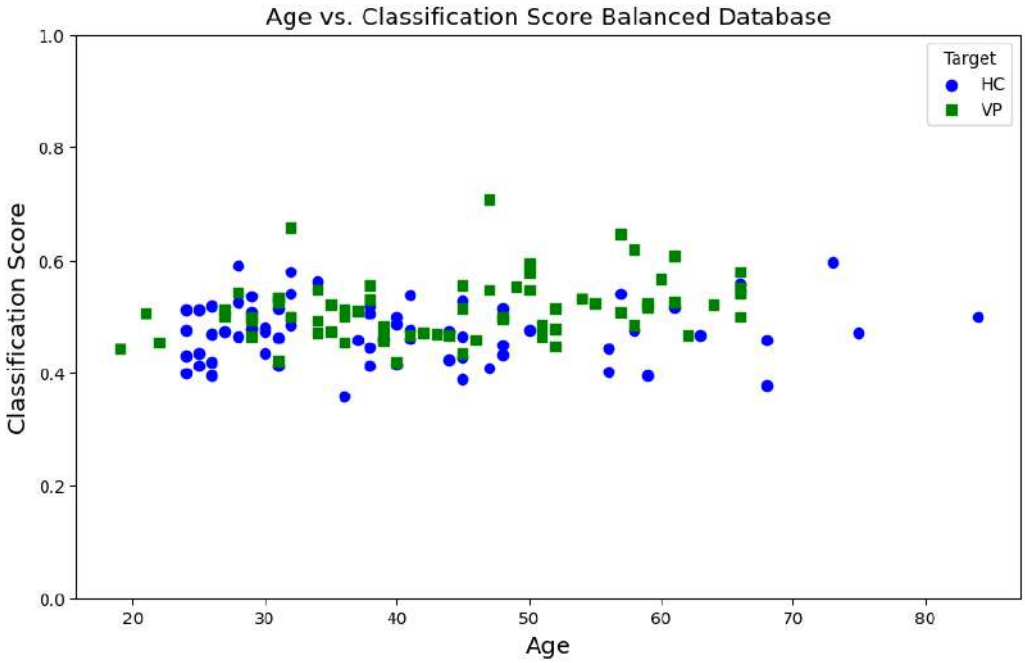


Figure 4.3. Age vs classification score with the balanced database

Classification with phonation features

Muscles like the vocal folds, diaphragm, or larynx play a crucial role in speech production. Abnormalities in these organs will induce changes in the subject speech. Features based on describing abnormal changes in these organs can be crucial in the early detection of a voice pathology. Phonation features measure changes in pitch and frequency, as well as amplitude variations of the speech signal.

Table 4.3 summarizes the accuracies obtained when using phonation features in all tasks for both signals and each classifier. The accuracies obtained for these experiments range from 52.7% up to 62.6%, which are not the greatest accuracies. Still, it is worth considering that the dataset selected from the SVD is balanced in terms of age and gender without worrying about the variety of pathologies in the final dataset. Table 4.3 shows that the best result had an accuracy of 62.6% using the vowel /u/ and the speech signal with a DT classifier. Table 5.3 shows that the experiment has a f1-score of 57.4%, a really good specificity of 91.6% but a low sensitivity of 33.1%. The EGG signal obtained close results for tasks like the vowel /a/ and the SVM classifier with an accuracy of 60.9%.

Table 4.3. Accuracies obtained for each task using phonation features in both signals. **EGG**: Electroglottography, **ACC DT**: Accuracy decision tree, **ACC RF**: Accuracy random forest, **ACC SVM**: Accuracy support vector machine. The mean ± standard deviation is reported.

	Speech			**EGG**		
Task	ACC DT [%]	ACC RF [%]	ACC SVM [%]	ACC DT [%]	ACC RF [%]	ACC SVM [%]
Vowel /a/	58.7 ± 9.3	61.1 ± 6.7	57.4 ± 9.2	58.9 ± 9.5	58.2 ± 11.8	**60.9 ± 12.8**
Vowel /i/	55.2 ± 13.3	54.4 ± 4.4	61.6 ± 14.2	58.1 ± 8.5	56.1 ± 10.0	52.8 ± 5.1
Vowel /u/	**62.6 ± 13.9**	54.7 ± 6.5	61.6 ± 10.2	55.4 ± 11.6	51.8 ± 3.4	52.7 ± 7.3
Phrase	61.0 ± 11.0	54.5 ± 7.0	58.8 ± 9.6	58.4 ± 11.4	58.0 ± 14.6	55.7 ± 5.2

Figure 4.4 shows the ROC curves of the three classifiers for the vowel /u/ using the speech signal and the vowel /a/ using the EGG signal. Also, the bottom right part of each plot shows the area under the curve.

More detailed classification metrics like the F1-score, sensitivity, and specificity using phonation features, as well as the results of the other tasks and the parameters used, can be found in chapter 5 from Table 5.1 to Table 5.4.

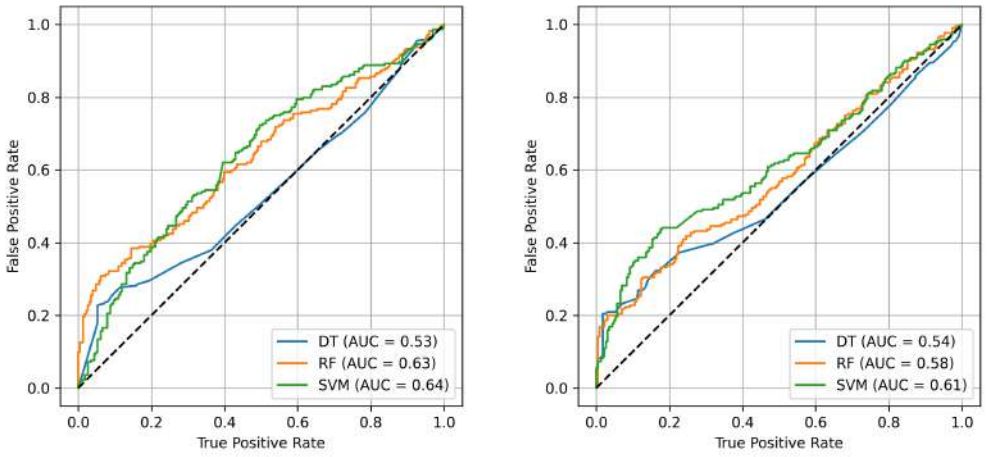


Figure 4.4. ROC curve for phonation features using vowel /u/ for the EGG signal (left) and vowel /a/ for the speech signal (right)

Classification with articulation features

After the vocal fold vibrates to generate sound waves, different muscles like the thong, the lips, and the palate, among others, coordinate to shape the sound waves in clear speech waves. Issues in some of those muscles will affect the quality of the voice. For instance, changes in the volume or a more breathy voice. Articulation features aim to capture patterns related to issues in the organs involved in speech production. An example is the format frequency that analyzes the changes in the resonate frequencies of the vocal tract caused by problems in some articulators.

Similar to the previous experiment, Table 4.4 summarize the accuracies of the different classifiers trained with the articulation features extracted from speech and EGG. Table 4.4 shows the best results were obtained using the EGG signal. Vowel /u/ yielded an accuracy of 62.2% with an f1-score of 57.5%, sensitivity of 37.5%, and specificity of 86.8%. Meanwhile, the vowel /i/ obtained a similar accuracy of 62.2 but with a higher standard deviation than the vowel /u/.

Table 4.4. Accuracies obtained for each task using articulation features for both signals. **EGG**: Electroglottography, **ACC DT**: Accuracy decision tree, **ACC RF**: Accuracy random forest, **ACC SVM**: Accuracy support vector machine. The mean ± standard deviation is reported.

	Speech			**EGG**		
Task	ACC DT [%]	ACC RF [%]	ACC SVM [%]	ACC DT [%]	ACC RF [%]	ACC SVM [%]
Vowel /a/	52.8 ± 7.8	56.7 ± 10.3	58.2 ± 9.0	54.1 ± 9.4	57.5 ± 10.1	57.8 ± 11.9
Vowel /i/	50.2 ± 6.2	58.0 ± 9.8	57.6 ± 7.3	**62.2 ± 13.6**	56.9 ± 8.7	60.3 ± 12.2
Vowel /u/	55.6 ± 4.9	53.4 ± 8.2	59.1 ± 7.0	**62.2 ± 12.5**	59.1 ± 10.3	59.5 ± 11.9
Phrase	50.8 ± 11.6	57.2 ± 8.8	59.3 ± 6.3	55.2 ± 6.8	58.9 ± 11.9	61.0 ± 10.9

A comparison of the performance for the three classifiers in both vowels/u/ and /i/ can be seen in the ROC curves shown in Figure 4.5. Information regarding the best parameters, as well as the results for the other tasks, can be found in chapter 5 from Table 5.5 to Table 5.8.

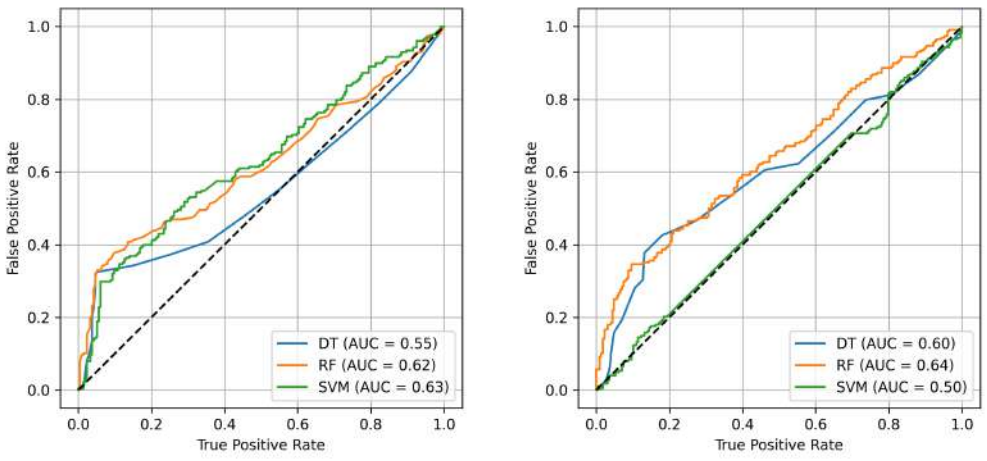


Figure 4.5. ROC curve for the best results obtained with articulation features using vowel /i/ (left) and vowel /u/ (right) for the EGG signal

Classification with BFCCs features

Bark frequency cepstral coefficient features provide information about the voice in the frequency domain. These features are usually robust against noise and are relevant for voice pathology detection due to their close relation to physiological processes.

A summary of the accuracy obtained using these features for all the tasks and signals can be found in Table 4.5. For the case of the BFCCs, the two best results are using the EGG signal. The best result was using the vowel /a/ alongside the DT classifier with an accuracy of 64.6%, a f1-score of 63.7%, a sensitivity of 54.8%, and a specificity of 74.5%. The second best was obtained using the phrase task and the SVM classifier, which obtained a 64.3% accuracy not too far from the vowel /a/.

Table 4.5. Accuracies obtained for each task using BFCCs features for both signals. **EGG**: Electroglottography, **ACC DT**: Accuracy decision tree, **ACC RF**: Accuracy random forest, **ACC SVM**: Accuracy support vector machine. The mean ± standard deviation is reported.

	Speech			**EGG**		
Task	ACC DT [%]	ACC RF [%]	ACC SVM [%]	ACC DT [%]	ACC RF [%]	ACC SVM [%]
Vowel /a/	55.0 ± 6.6	56.3 ± 8.4	58.2 ± 11.2	**64.6 ± 12.6**	59.3 ± 12.9	57.8 ± 9.4
Vowel /i/	45.7 ± 6.8	52.5 ± 10.0	57.8 ± 8.9	56.7 ± 11.2	56.8 ± 7.7	58.0 ± 11.3
Vowel /u/	43.1 ± 10.7	56.2 ± 15.7	57.4 ± 10.1	58.0 ± 13.1	59.7 ± 10.8	59.9 ± 14.6
Phrase	56.1 ± 10.6	61.3 ± 9.7	63.0 ± 8.0	57.3 ± 12.8	60.8 ± 13.1	**64.3 ± 10.2**

The ROC curve of the best results can be seen in Figure 4.6. Table 5.9 to Table 5.12 show more details on the classification using different tasks and all the metrics and best parameters.

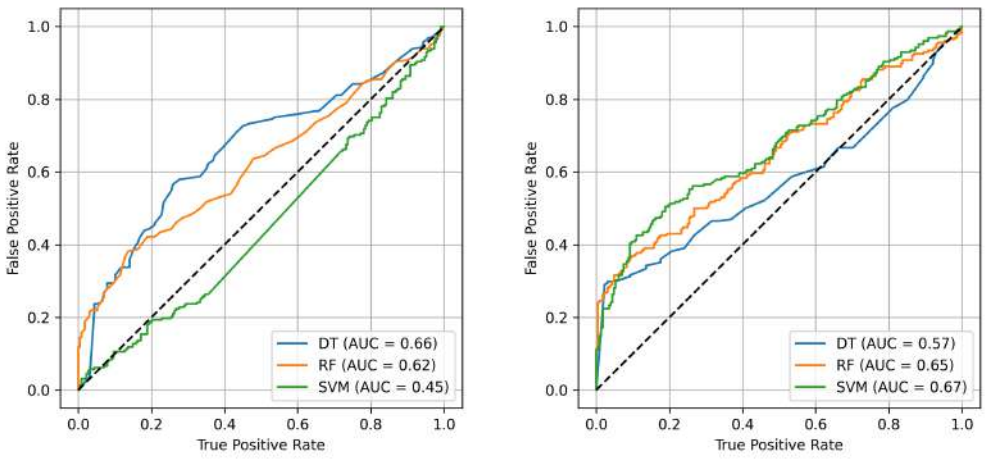


Figure 4.6. ROC curve for BFCC features using vowel /a/ (left) and phrase (right) for the EGG signal.

Classification with nonlinear features

Studies have shown the nonlinear dynamics during the voice production process [35]. The involvement of multiple systems, like the vocal folds and the vocal tract, among others, makes voice production a highly complex and nonlinear task. The analysis of the chaotic behavior, like roughness in the voice, is an indicator of voice pathologies or at least problems in the voice, so features that analyze and characterize this chaos during voice production can be relevant in the analysis of voice pathologies.

Table 4.6 summarizes the accuracies obtained when the nonlinear features are extracted from the EGG and speech signal and used to classify voice pathologies. The best result was obtained using the vowel /a/ with the speech signal and the SVM classifier; this combination obtained an accuracy of 60.6%, a f1-score of 59.3%, a sensitivity of 51.9%, and a specificity of 69.3%. The second-best result was obtained with the vowel /i/ using the EGG signal and the DT classifier with an accuracy of 60.5%

Table 4.6. Accuracies obtained for each task using non-linear features for both signals. **EGG**: Electroglottography, **ACC DT**: Accuracy decision tree, **ACC RF**: Accuracy random forest, **ACC SVM**: Accuracy support vector machine. The mean ± standard deviation is reported.

	Speech			**EGG**		
Task	ACC DT [%]	ACC RF [%]	ACC SVM [%]	ACC DT [%]	ACC RF [%]	ACC SVM [%]
Vowel /a/	58.7 ± 9.2	55.4 ± 10.8	**60.6 ± 9.8**	56.2 ± 11.1	57.5 ± 10.6	60.2 ± 13.4
Vowel /i/	51.2 ± 7.5	51.2 ± 7.5	54.9 ± 7.8	**60.5 ± 7.8**	58.2 ± 7.2	58.3 ± 8.5
Vowel /u/	48.1 ± 10.5	49.2 ± 9.3	51.2 ± 4.7	59.0 ± 13.7	55.8 ± 8.8	54.3 ± 6.4
Phrase	54.7 ± 8.5	54.3 ± 9.4	58.9 ± 9.4	56.2 ± 14.3	59.1 ± 10.3	59.3 ± 9.5

Figure 4.7 shows the ROC curves of the three classifiers for the vowel /u/ using the speech signal and the vowel /a/ using the EGG signal. Also, the bottom right part of each plot shows the area under the curve.

More detailed classification metrics like the F1-score, sensitivity, and specificity using phonation features, as well as the results of the other tasks and the parameters used, can be found from Table 5.13 to Table 5.16.

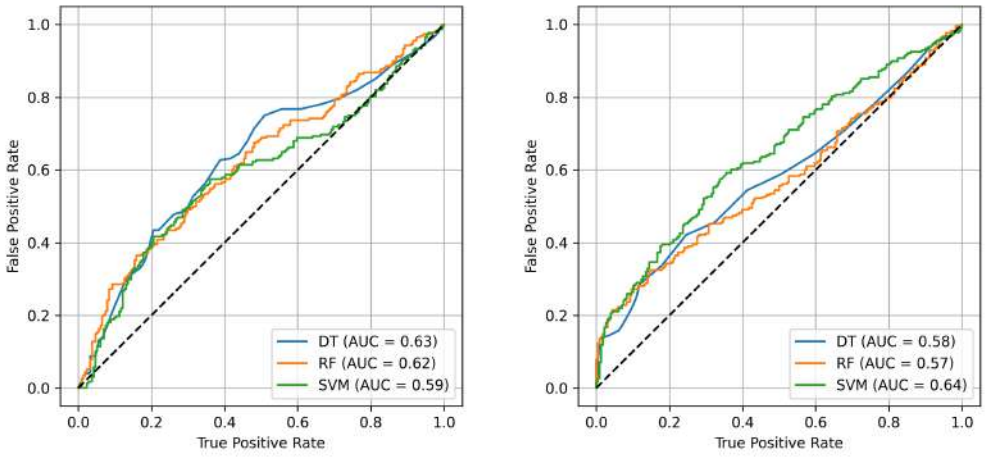


Figure 4.7. ROC curve for nonlinear features using vowel /a/ for the EGG signal (left) and speech (right)

Classification with phase plots features

All previous experiments were performed using a classical feature extraction and classification approach. This classification method grants us more control and interpretation of the features used and relates these features to mathematically represented physiological processes. However, more modern techniques have recently been used. This allows automatic feature extraction and a more customizable classifier. Both the feature extractor and the classifier are trained to learn how to characterize the problem and are highly used because they can adapt to multiple problems based on the data used for the training.

Images are the most common input signal where features are extracted using modern techniques. Images have high dimensionality and require an analysis of the spatial information; this is when automatic feature extraction can exceed handcrafted ones.

Phase plots are 2D representations of the glottal cycles during speech production. Because it is a 2D representation (image), features are extracted using a convolutional neural network combined with a fully connected, Figure 4.8 shows a graphical representation of the network used for this process. The convolutional layers have 64, 32, and 16 kernels respectively, with a kernel size of 3x3 except for the second layer. Rectified Linear Unit (ReLU) ac-

tivation functions are employed in all layers. Max pooling with a pool size of 2x2 is applied after each convolution. The network is trained using an ADAM optimizer [102] with a batch size of 16 and 100 epochs. To prevent overfitting, an early stopping algorithm with a patience of 5 is implemented. Two hyper-parameters of the network are optimized; the first one is the learning rate (Lr) of the network that can take values from $Lr \in \{1e-6, 1e-5, 1e-4\}$ and the size of the embeddings generated after the flattened layer, the values of the last ranges from $Ls \in \{128, 256, 512\}$. This optimization is done for each combination of task and signal; the parameters that yielded the best accuracy in the test are selected. Similarly to the classical approach, these parameters are fixed, and the network is trained again. Performance metrics like accuracy, f1 - score, sensitivity, and specificity are calculated for each fold and the mean and standard deviation are reported.

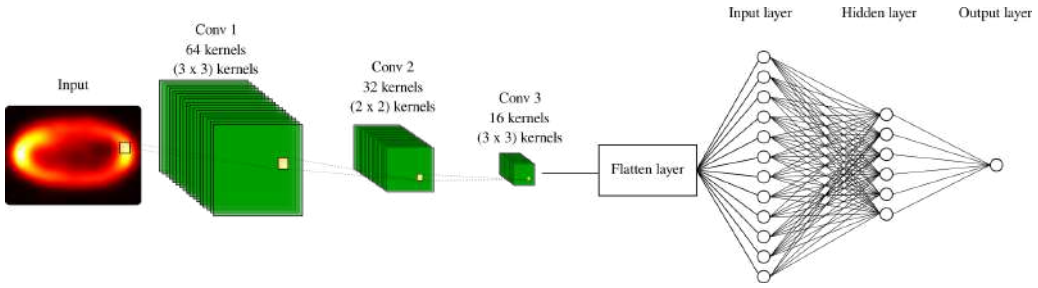


Figure 4.8. CNN architecture combined with a fully connected for the classification of voice pathologies

The convolutional layers are in charge of detection patterns inside the image and reduce the input image's dimensionality without losing too much information. Lastly, after the features are extracted, a fully connected layer uses these features to discriminate a patient with a voice pathology from a healthy subject.

Table 4.7. Results obtained by classifying voice pathologies using phase plots. **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity, **Lr**: Learning rate, **Ls**: Layer size

			Speech		
Task	**Parameters**	**ACC [%]**	**F1 [%]**	**SEN [%]**	**SPE [%]**
Vowel /a/	Lr= 0.0001 Ls= 512	56.1 ± 8.3	55.2 ± 8.2	57.6 ± 20.6	62.0 ± 12.0
Vowel /i/	Lr= 0.001 Ls= 128	54.4 ± 6.5	53.5 ± 7.2	38.3 ± 29.7	59.3 ± 30.5
Vowel /u/	Lr= 1e-05 Ls= 256	56.8 ± 8.0	56.2 ± 8.5	66.1 ± 46.8	33.3 ± 47.1
Phrase	Lr= 0.0001 Ls= 128	58.3 ± 9.8	57.3 ± 10.1	63.2 ± 45.6	35.7 ± 45.9
			EGG		
Task	**Parameters**	**ACC [%]**	**F1 [%]**	**SEN [%]**	**SPE [%]**
Vowel /a/	Lr= 1e-05 Ls= 512	**59.0 ± 11.1**	57.3 ± 12.1	56.5 ± 42.9	53.0 ± 47.6
Vowel /i/	Lr= 0.0001 Ls= 256	55.4 ± 8.5	52.4 ± 10.8	43.5 ± 30.5	81.2 ± 17.1
Vowel /u/	Lr= 0.001 Ls= 256	**58.3 ± 10.8**	55.3 ± 13.2	34.8 ± 29.9	88.8 ± 11.4
Phrase	Lr= 0.001 Ls= 512	57.3 ± 8.0	55.1 ± 8.3	63.8 ± 33.1	43.2 ± 29.4

Since we are using a single classifier, the DNN, the ROC curve illustrates the accuracies obtained for each task using EGG and speech. This figure can be found in Figure 4.9.

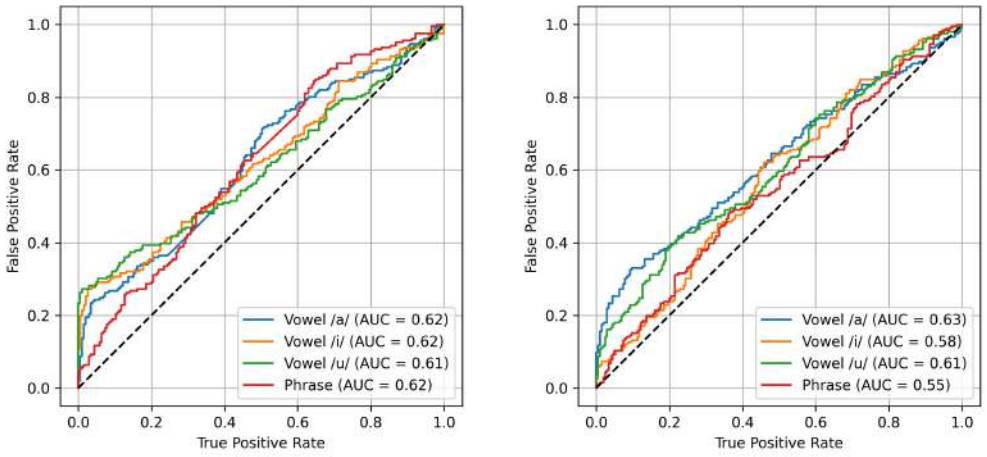


Figure 4.9. ROC curves for phase plot features derived from the EGG signal (left) and speech (right) across all tasks.

4.1.3 Discussion

The process done in the balance of the dataset generated a subset that does not consider the type of pathologies or the number of samples of each pathology. This creates a problem that can be more difficult to solve since some pathologies can have few subjects, the disease hasn't affected the patient too much, or simply the pathologies are difficult to differentiate between a healthy subject. However, this makes the experiments closer to where we don't know the patient's pathology; we aim to see if there are patterns that can be related to voice pathology.

Overall, the uni-modal experiments didn't show good results in accuracy; none of the experiments could surpass 70% accuracy, showing that the models had difficulties during the classification process. As detailed in the tables in the Appendix, most experiments exhibited low sensitivity values (below 30%) but high specificity, meaning that the model couldn't detect voice pathology. This causes the model to predict most samples as one class, generating accuracies close to 50 %.

Besides the low accuracy values, several observations emerged from the experiments. Most works in the literature focus exclusively on the vowel /a/, normally because it is the most common one. However, the experiments showed that other vowels can also obtain great results and work better de-

pending on the features. For example, the vowels/i/ and /u/ worked well with the articulation features, being the best two results with that feature set.

The best result overall was obtained by the BFCCs features; this set of features not only got the two best results with the vowel /a/ and the phrase but also generated more balanced models, obtaining sensitivities of 61.2%. These findings highlight the potential of the BFCCs features.

Lastly, the phase plots did not improve the results obtained by other classical approaches. Because the deep learning approach requires a large amount of data for the training process, the model is not able to learn and generalize the phenomena before being stooped by the early stopping algorithm to avoid overfitting. However, other techniques are used in further experiments to showcase the capabilities of these novel features.

4.2 Multi-modal approach

The uni-modal approach showed promising results in some combinations. Still, it is worth noticing that the sensitivity in most of them is low (below 30%), generating a good model for detecting the negative class. Still, VP is not the best for the positive class.

The introduction of complementary information can greatly enhance the classifier's performance. This is where the fusion of techniques can be useful. The main hypothesis is that combining the information of both modalities can improve the classification process. There are two ways of combining the information; the first consists of concatenating the two features and creating a new vector. The second one consists of taking the decision probabilities obtained by the classifier at the end and, with a conjoint probability, finding a new decision. By using these fusion techniques, we aim to introduce information from different sources that can potentially improve the accuracy and robustness of the classifier.

4.2.1 Experiments and results

Early fusion with classical features

Classical features are commonly created based on previous knowledge of the problem we want to analyze; this allows us to control and have an interpretation of the information we are introducing to a classifier. Also, these features

have proven to be effective in previous scenarios. The results obtained using these classical features were not the best, and a possible solution to this low accuracy can be to introduce additional information from other sources.

The hypothesis is that combining EGG and speech signals can improve the classification process and generate a more robust model. Due to the many combinations that can result from the fusion, only the three best results are showcased for each task.Table 4.8 to Table 4.11 shows the accuracy, f1-score, sensitivity and specificity for each task. The best result was obtained using the phrase task and combining the BFCC features extracted from both signals. The SVM obtained an accuracy of 66.8% and helped to increase the sensitivity values obtained so far up to 56.4%. Figure 4.10 shows the ROC curve with the best result obtained in each task.

Table 4.8. Best results obtained in the classification of voice pathologies using the vowel /a/ and early fusion. **DT**: Decision Tree, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity, **BFCC**: Bark Frequency Cepstral Coefficients

Feature 1	Feature 2	Classifier	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Articulation EGG	Nonlinear speech	SVM	63.0 ± 8.2	62.4 ± 8.5	59.0 ± 16.3	67.0 ± 11.0
BFCC EGG	Nonlinear speech	SVM	62.8 ± 8.2	62.4 ± 8.3	62.1 ± 17.5	63.5 ± 4.7
Nonlinear EGG	BFCC speech	DT	62.8 ± 8.4	61.6 ± 9.5	57.7 ± 22.1	67.9 ± 12.9

Table 4.9. Best results obtained by classifying voice pathologies using the vowel /i/ and early fusion. **DT**: Decision Tree, **RF**: Random Forest, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity, **BFCC**: Bark Frequency Cepstral Coefficients

Feature 1	Feature 2	Classifier	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Nonlinear EGG	Nonlinear speech	RF	63.8 ± 8.5	63.2 ± 8.8	61.7 ± 19.7	66.6 ± 10.1
Nonlinear EGG	Nonlinear speech	DT	62.7 ± 10.7	60.4 ± 12.5	49.8 ± 27.1	77.0 ± 10.2
Articulation EGG	BFCC speech	DT	62.2 ± 13.6	56.7 ± 17.7	33.1 ± 24.3	94.1 ± 4.3

Table 4.10. Best results obtained by classifying voice pathologies using the vowel /u/ and early fusion. **DT**: Decision Tree, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity, **BFCC**: Bark Frequency Cepstral Coefficients

Feature 1	Feature 2	Classifier	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Articulation EGG	Phonation speech	SVM	64.6 ± 10.0	62.6 ± 11.5	49.4 ± 24.3	79.8 ± 8.0
Articulation EGG	Nonlinear speech	DT	64.4 ± 10.0	61.8 ± 12.5	47.5 ± 23.8	81.0 ± 11.2
Phonation EGG	BFCC speech	DT	63.8 ± 11.4	62.9 ± 11.7	53.9 ± 18.5	73.6 ± 14.2

Table 4.11. Best results obtained by classifying voice pathologies using the phrase and early fusion. **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity, **BFCC**: Bark Frequency Cepstral Coefficients

Feature 1	Feature 2	Classifier	ACC [%]	F1 [%]	SEN [%]	SPE [%]
BFCC EGG	BFCC Speech	SVM	66.8 ± 6.2	65.8 ± 6.7	56.4 ± 14.2	77.2 ± 14.4
Articulation EGG	BFCC Speech	SVM	65.2 ± 9.3	64.2 ± 10.2	63.8 ± 21.3	66.7 ± 14.3
BFCC EGG	Articulation Speech	SVM	65.0 ± 8.3	64.2 ± 8.7	58.5 ± 18.0	71.5 ± 12.8

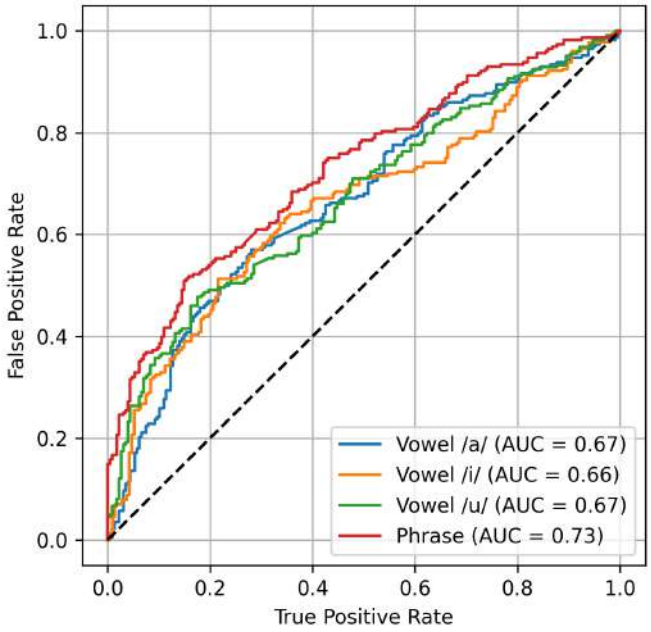


Figure 4.10. ROC curves for the best results for each task using early fusion in the classical features.

Early fusion with phase plots

Some classical features that were extracted require a window frame for their calculation. Framing the audio can lead to different feature lengths depending on the audio length. One way to solve this issue is to crop all the audios to the same length or to calculate the features for all the windows and use statistical values such as mean, standard deviation, kurtosis, and skewness of the features as the feature set, creating a static vector because the time constrain is removed.

These static vectors can be useful for the classification of voice pathologies as seen in previous results. However, some information is lost because of the compression nature, leading to misclassification and poor performance. On the other hand, phase plots are features that take into account the dynamics over time. Because they are images, the most common way is to extract the features automatically. The combination of classical features that use static feature sets with the phase plots that take into account the dynamics of the audio can lead to improvements in the classification process, and generate a more robust model.

One problem arises when attempting this combination of features because the phase plots are features extracted automatically and then used in a DNN to perform the classification while the classical features are an array of numbers, the concatenation cannot be done so easily as previous experiments. Either the classical features need to be introduced into the DNN alongside the phase plots or the output generated by the CNN for an image is taken as a feature set and used in the classical models. The low number of samples and increased number of inputs can generate that a DNN model overfits, so a better approach is to extract the features generated for the CNN and train classical classifiers.

For the feature extraction process, the best networks found in the unimodal approach with phase plots are used, after the network was trained and all the weights were adjusted, each sample was introduced to the network, and the array obtained after the flattened layer is taken as the feature embedding of that input.

Similar to previous experiments, just the three best results per task are reported. Table 4.12 to Table 4.11 shows the best results obtained for each task,

Table 4.12. Best results obtained in the classification of voice pathologies using the vowel /a/ and early fusion with phase plots. **DT**: Decision Tree, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity

Feature 1	Feature 2	Classifier	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Phase plot EGG	Nonlinear speech	SVM	84.5 ± 13.8	82.5 ± 18.1	79.3 ± 31.3	89.5 ± 15.8
Phase plot EGG	Phonation speech	SVM	83.1 ± 13.5	81.0 ± 17.9	78.4 ± 32.6	88.2 ± 14.5
Articulation EGG	Phase plot speech	SVM	83.0 ± 13.3	80.9 ± 17.6	76.3 ± 30.5	89.5 ± 14.0

Table 4.13. Best results obtained in the classification of voice pathologies using the vowel /i/ and early fusion with phase plots. **DT**: Decision Tree, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity

Feature 1	Feature 2	Classifier	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Phase plot EGG	Phonation speech	SVM	87.7 ± 13.5	85.9 ± 18.1	82.8 ± 29.5	93.0 ± 11.5
Phonation EGG	Phase plot speech	RF	87.0 ± 15.6	85.2 ± 19.8	84.5 ± 31.8	89.5 ± 11.6
Nonlinear EGG	Phase plot speech	SVM	86.7 ± 14.4	84.7 ± 18.9	84.6 ± 30.5	88.6 ± 17.0

Table 4.14. Best results obtained in the classification of voice pathologies using the vowel /u/ and early fusion with phase plots. **DT**: Decision Tree, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity

Feature 1	Feature 2	Classifier	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Phase plot EGG	Phonation speech	SVM	86.2 ± 14.9	84.1 ± 19.3	82.8 ± 31.2	89.5 ± 17.5
Nonlinear EGG	Phase plot speech	SVM	85.6 ± 15.0	83.4 ± 19.5	82.3 ± 31.2	88.6 ± 19.2
Phase plot EGG	Articulation speech	RF	85.4 ± 15.7	83.3 ± 20.0	81.4 ± 32.7	89.1 ± 15.8

Table 4.15. Best results obtained in the classification of voice pathologies using the phrase and early fusion with phase plots. **DT**: Decision Tree, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity, **BFCC**: Bark Frequency Cepstral Coefficients

Feature 1	Feature 2	Classifier	ACC [%]	F1 [%]	SEN [%]	SPE [%]
BFCC EGG	Phase plot speech	RF	88.5 ± 14.9	86.8 ± 18.6	86.3 ± 29.8	90.4 ± 16.1
Phonation EGG	Phase plot speech	RF	87.7 ± 16.8	85.9 ± 20.6	85.4 ± 31.5	90.4 ± 16.2
Nonlinear EGG	Phase plot speech	RF	87.2 ± 16.3	85.6 ± 19.6	86.3 ± 29.8	87.8 ± 16.5

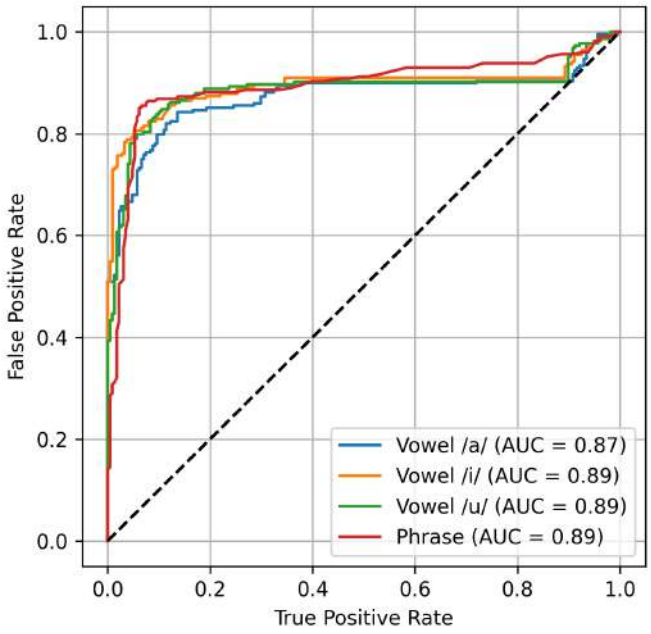


Figure 4.11. ROC curves for the best results for each task using early fusion.

Late fusion classification

Another method of combining information is to perform this combination after the classification process is performed. Each classifier at the end gives each sample a probability of belonging to each class. Depending on the feature, the task, the signal, and the classifier, the probability can be stronger, meaning that the classifier is more sure about the decision. Late fusion combines the decision of two classifiers and gives a new probability. The decision of the new probability can be made by selecting the maximum probability out of the two or averaging the probabilities. The second approach is followed for this experiment.

Late fusion has the advantage that does not require the training of new models as in the early fusion. The three best results for each task are shown from Table 4.16 to Table 4.19. The best result overall was obtained using the phrase task, combining the probabilities obtained by two SVM trained with BFCCs features. Not only did the experiment yield an accuracy of 67.0% improving the best accuracy obtained in the uni-modal experiments, but it also generated a more balanced model with a sensitivity of 61.1% and

a specificity of 72.8%. Figure 4.12 shows the ROC curve for the best result in each task.

Table 4.16. Best results obtained in the classification of voice pathologies using the vowel /a/ and late fusion. **DT**: Decision Tree, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity, **BFCC**: Bark Frequency Cepstral Coefficients

Feature 1	Classifier 1	Feature 2	Classifier 2	ACC [%]	F1 [%]	SEN [%]	SPE [%]
BFCC EGG	DT	BFCC speech	RF	65.4 ± 12.3	64.8 ± 13.0	55.6 ± 18.3	75.3 ± 7.4
BFCC EGG	DT	BFCC speech	SVM	65.2 ± 13.2	64.4 ± 14.3	56.5 ± 22.0	74.0 ± 6.4
Phonation EGG	DT	Phonation speech	SVM	62.0 ± 15.0	57.5 ± 17.9	37.6 ± 30.1	86.3 ± 7.3

Table 4.17. Best results obtained in the classification of voice pathologies using the vowel /i/ and late fusion. **DT**: Decision Tree, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity, **BFCC**: Bark Frequency Cepstral Coefficients

Feature 1	Classifier 1	Feature 2	Classifier 2	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Phonation EGG	SVM	Phonation speech	RF	62.7 ± 12.2	61.0 ± 13.5	51.9 ± 26.6	75.3 ± 13.8
Phonation EGG	SVM	Phonation speech	DT	62.7 ± 11.1	59.9 ± 13.2	44.6 ± 27.0	82.9 ± 10.1
Articulation EGG	DT	Articulation speech	SVM	62.7 ± 10.5	60.8 ± 12.1	50.1 ± 24.3	76.8 ± 11.2

Table 4.18. Best results obtained in the classification of voice pathologies using the vowel /u/ and late fusion. **DT**: Decision Tree, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity, **BFCC**: Bark Frequency Cepstral Coefficients

Feature 1	Classifier 1	Feature 2	Classifier 2	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Phonation EGG	DT	Phonation speech	RF	64.2 ± 12.4	61.4 ± 14.1	44.9 ± 26.6	83.7 ± 11.7
Phonation EGG	SVM	Phonation speech	SVM	63.3 ± 12.1	61.7 ± 12.9	51.8 ± 25.6	75.4 ± 12.1
Phonation EGG	DT	Phonation speech	SVM	62.2 ± 13.1	59.1 ± 15.1	41.4 ± 25.2	82.8 ± 9.9

Table 4.19. Best results obtained in the classification of voice pathologies using the phrase and late fusion. **DT**: Decision Tree, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity, **BFCC**: Bark Frequency Cepstral Coefficients

Feature 1	Classifier 1	Feature 2	Classifier 2	ACC [%]	F1 [%]	SEN [%]	SPE [%]
BFCC EGG	SVM	BFCC speech	SVM	67.0 ± 9.7	65.9 ± 10.3	61.1 ± 21.0	72.8 ± 14.2
BFCC EGG	SVM	BFCC speech	RF	63.7 ± 11.8	62.4 ± 12.6	56.7 ± 22.8	70.6 ± 15.6
BFCC EGG	RF	BFCC speech	SVM	63.5 ± 9.6	62.2 ± 10.3	56.3 ± 21.5	70.6 ± 14.8

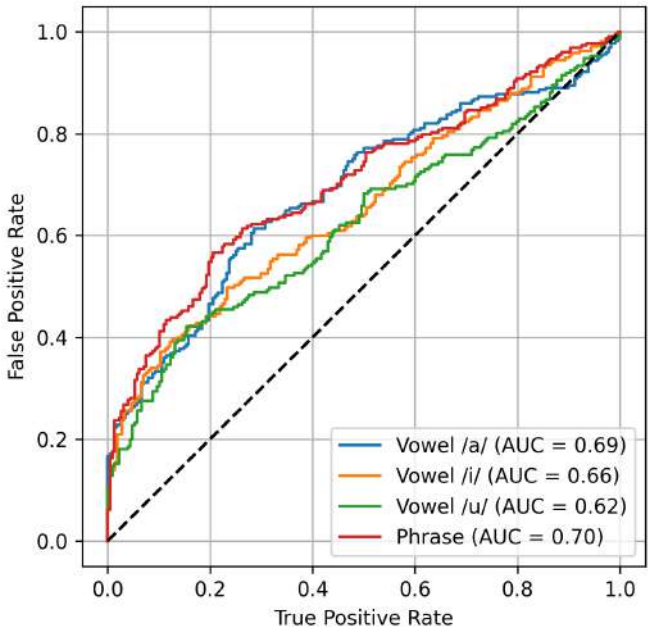


Figure 4.12. ROC curves for the best results for each task using late fusion.

4.2.2 Discussion

The results obtained from the fusion process demonstrate the potential effectiveness of these methods. Uni-modal experiments faced challenges during the classification process, yielding low sensitivity and, in some cases, specificity. This resulted in models with accuracies lower than 60%. To address these issues, an early fusion of classical features was employed.

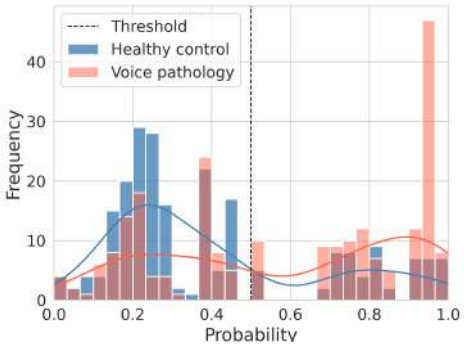
When combining classical features, nonlinear and BFCC features frequently ranked among the top three best combinations. BFCCs, which previously achieved the best results in the uni-modal experiments, were expected to perform well in the fusion process. On the other hand, the nonlinear features proved to be a crucial element for the classification process, providing complementary information that enhanced overall performance.

The best accuracies were obtained when the phase plots were included in the early fusion. BFCC features, when combined with the phase plots in the phrase task, showed a model with 88.5% accuracy. The combination of static information from the classical features and dynamic information from the phase plots yielded accuracies above 80%, generating more robust models

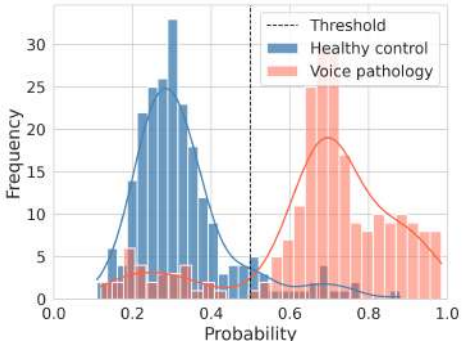
with sensitivities and specificities, in some cases, higher than 85%. This was particularly evident in the best result obtained in this study. SVM and RF classifiers most commonly appeared in the top three results, while DT was left behind in these early fusion results. This likely occurs because early fusion increases the number of features introduced to the model, necessitating the use of more complex models.

Lastly, late fusion did not improve upon the results obtained from early fusion with phase plots but still showed interesting outcomes. Features like phonation proved important when performing this type of fusion. Late fusion can be a valuable approach when we have models that are already trained, and we want to combine their information without needing to retrain them.

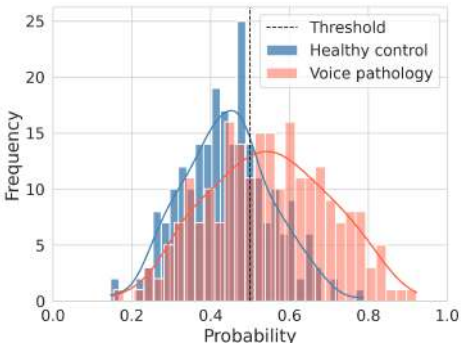
Figure 4.13 shows a comparison of the distribution score for the best result in each main experiment. 4.13a is the result obtained when using the vowel /a/ and BFCC features extracted from the EGG signal and a DT classifier, we can notice that the distributions are separated but a big portion of one class is always being misclassified. 4.13b shows the distribution obtained when the phase plots and the BFCCs features are concatenated, the plots show that the combination of the features improved the distribution and the separation of the classes. Lastly, 4.13c shows the score distribution when the probabilities of two SVM trained with BFCC features are combined, the plot shows an improvement compared with the uni-modal approach but the means of the two distributions are close to the threshold, generating multiple misclassified values reducing the accuracy of the model.



(a) Score distribution for the vowel /a/ using the BFCCs features extracted from the EGG signal and a DT classifier.



(b) Score distribution for the phrase task concatenating the BFCCs features extracted from the EGG signal and the phase plots extracted from the speech signal alongside a RF classifier.



(c) Score distribution for the phrase task combining the probabilities obtained by two SVM classifiers trained with BFCCs features extracted from EGG and speech.

Figure 4.13. Score distribution for the best results in each main set of experiments.

Chapter 5

Conclusions and future work

This work evaluated different strategies for the automatic classification of voice pathologies, with a focus on analyzing the SVD database, a German database that contains EGG and speech signals from healthy subjects and patients with various voice pathologies performing sustained vowels and a phrase task. The imbalances in both age and gender required the selection of a subset of the database to remove these imbalances and avoid biases in future experiments. Multiple handcrafted features were extracted based on the physiological aspects of the phenomena. Additionally, various CNN architectures were trained for the automatic feature extraction of a novel feature called phase plots. Finally, early and late fusion techniques were employed to combine the information from the two modalities.

Features such as BFCC, articulation, phonation, and nonlinear characteristics were extracted as handcrafted features, while phase plots utilized CNNs for automatic feature extraction. Initial experiments did not yield promising results, with models achieving accuracies close to 50% and sensitivity values as low as 0.0% in some cases. However, BFCC showed good accuracy and produced a more balanced model among the uni-modal approaches. Uni-modal experiments also highlighted the importance of other vowel tasks, such as vowel /i/, which achieved similar or even better results compared to vowel /a/, depending on the feature set used.

In the modern approach, a CNN with a fully connected layer was trained to classify voice pathology using phase plots, a novel feature set representing glottal cycles over time in a 2D plot. The CNN kernels were optimized to extract features from the phase plots, which were then used to optimize the weights of the fully connected layer. The low accuracies observed can be

attributed to the need for large amounts of data to train DNNs to avoid overfitting.

Two fusion methods were implemented to combine the information from EGG and speech. Early fusion concatenated the feature set before classification, while late fusion used the probabilities obtained by the classifier to perform a new classification. Since the phase plots feature was automatically extracted during network training, an embedding for each subject was generated from the CNN network post-training. Early fusion with phase plots yielded the best results, with an 88.5% accuracy when combined with BFCC features and using the phrase task. This feature set combined static information from BFCC and dynamic information from phase plots.

The results observed in this work demonstrate that the fusion method using novel features complements information obtained from other features and shows significant potential in discriminating voice pathologies from healthy subjects. From a clinical perspective, the proposed approach presents a significant step towards enhancing the diagnostic process for voice pathologies. Traditional diagnostic methods often rely heavily on subjective assessments, such as auditory-perceptual evaluations by clinicians, or invasive techniques like laryngoscopy. These methods, while valuable, can be time-consuming, expensive, and dependent on the clinician's expertise. The introduction of automated classification systems, particularly those leveraging both EGG and speech signals, offers a non-invasive, objective, and potentially more accessible solution for early diagnosis and monitoring of voice disorders. However, because our database contains patients who have already been diagnosed, we do not have enough information to determine whether the models are effective for pre-assessment or confirmatory evaluation.

Another advantages of this work lies in its potential to overcome the challenge of imbalanced datasets in medical diagnostics. Voice pathologies are often underrepresented in datasets compared to healthy controls, leading to biased models. The method employed in this study to mitigate these imbalances ensures that the developed system remains reliable across diverse patient populations, making it more generalizable and effective in real-world clinical applications.

Future work could explore the dynamic analysis of other classical features, like features related to the nature of the EGG signal or information from the phase plots, to verify the influence of these dynamic features on classification. Additionally, applying this pipeline and the phase plots features to other

research fields, such as swallowing detection, could be highly valuable.

Appendix A

5.1 Christian-Albrechts-Universität zu Kiel time

During my master's program, I undertook a six-month internship in Germany, where the primary objective was to analyze and automatically detect swallowing events. Although my master's degree is not directly related to this specific topic, the project utilized a pipeline similar to the one I employed during my master's research. This time, however, the pipeline was specifically adapted and focused on the detection of swallowing.

Swallowing implies the transit of food, liquids, and saliva from the mouth to the stomach in a well-synchronized way [103]. This process can be characterized by different phases, i.e., oral, pharyngeal, and esophageal. Dysphagia is a syndrome that affects the typical sequence of phases, which appear secondary to neurogenic or neuromuscular disorders - namely as functional dysphagia -, or to structural conditions [104], [105]. Dysphagia leads to malnutrition, dehydration, aspiration pneumonia, and even death.

The instrumental diagnosis of dysphagia is performed by two reference methods, i.e. Videofluoroscopy Swallowing Studies (VFSS) and Fiberoptic Evaluation of Swallowing (FEES) [106]. While VFSS allows the real-time visualization of the bolus transit and the detection of penetration or aspiration, it is invasive due to the X-ray exposure; this limits the number of evaluations for follow-up, and it has associated radiation risks, e.g., cancer or damage in the lens of the eyes [107]. Otherwise, FEES is uncomfortable and sometimes painful with anesthesia requirements because it implies the presence of a strange body in the nasopharynx [108].

Information obtained from less invasive sensors has been analyzed to find alternatives for detecting swallowing problems. For this, it is important to consider the capability of swallowing detection to establish whether a signal or sensor can be used for swallowing evaluation. The swallowing process can

be analyzed from different perspectives -*dimensions* -, such as mechanical and acoustic. Different sensors and sources of information can address each of them. The acoustic dimension of swallowing has been assessed with headset microphones for automatic speech analysis using machine learning models to determine the dysphagia screening capability [109], [110]. In addition, throat microphones have been used in combination with accelerometers for swallow detection [111] and swallowing-related event detection [112], mainly using VFSS as validation and machine learning/deep learning schemes. Other works have used throat microphones for -silent- aspiration detection under machine learning schemes in patients with dysphagia [113], [114].

During the internship, magnetic sensors and a throat microphone were tested for their capabilities in detecting and evaluating swallowing. A small database of healthy subjects was recorded performing different swallowing tasks.

5.1.1 Sensors description

Throat microphone

A commercial throat microphone from the brand IASUS, specifically the GP3 model, was used. Its frequency response ranges from 20Hz to 20kHz. Throat microphones like these are commonly used in military or firefighting contexts due to their ability to transmit sound effectively in noisy environments.

TM captures voice signals using a piezoelectric transducer that senses the vibrations in the throat area when a sound is produced. This makes TM an excellent sensor for detecting swallowing tasks, allowing us to have a suitable baseline sensor to compare with the magnetic sensor.

Magnetic arrangement

Three DJ-type magnetic sensors were used. These are MI (Magneto-Impedance) sensors, claimed to be capable of detecting changes in magnetic fields on the order of nanoteslas (nT).

These sensors measure changes in the magnetic field in a specific direction. Therefore, the configuration involves using two type-A sensors for measurements along the Y and Z axes and one type-B sensor for the X axis.

The sensors require a 15V power supply, two ground pins, and one output pin. However, since it's a multi-sensor configuration, a synchronization signal

is necessary in the last pin of the sensors. This signal is a sinusoidal signal with a frequency of 1MHz, which a standard MI-CB-1DJ-OSC oscillator or a waveform generator can generate.

Due to the sensor's sensitivity, abrupt movements cause desynchronization. To prevent this, a 3D box was designed to securely hold the sensors in their respective positions and control their movement.

A coil is placed around the neck near Adam's apple to induce changes in the magnetic field during swallowing. The movement of the throat during swallowing changes the magnetic field, which the magnetic sensors detect.

5.1.2 Data collection

The swallowing dataset was built with the help of people from the Digital Signal Processing and System Theory who participated voluntarily in the study. All the signals were recorded in two sessions; the throat microphone was used as the only sensor in the first recording, and for the second one, both sensors were used to generate three sets of signals (throat microphone, magnetic sensor, and both sensors simultaneously). This approach allows us to have both signals individually and check if the throat microphone influences the signals measured by the magnetic sensors.

Participants signed a consent form in their first session, which allowed the recordings and data to be used. A total of 8 tasks are recorded, which are grouped depending on the aim of the task. The first group of tasks was the swallowing task; the participant was requested to swallow saliva, water, and yogurt three times for each liquid. The water was provided in a cup, and a spoon was used for the yogurt; in both cases, the liquid supplied was 10 ml. The second set of tasks are called noise tasks; these aim to be sounds that confuse the model, being sounds that can be produced in the throat area but are not related to swallowing; for these groups, the participant was requested to cough 10 times and clear their throats 10 times.

The two other tasks include a text that contains different swallowing and noise tasks in between the text; for the two acoustic tasks, the participant is asked to sustain the vowel A for three seconds or more and to say the phrase "Guten Morgen, wie geht es Ihnen?".

This recording session is done first with the throat microphone, then the magnetic sensor is placed, and the recordings are repeated. Finally, the throat microphone is removed, and the last recordings are performed. The

Kiel Real-time Application Toolkit, or KiRAT, records both sensors simultaneously. The software allows us to capture the signal from the throat microphone and the three signals (one for each axis) from the magnetic sensor, the camera, and the condenser microphone, store the signals under one folder, and display the signal values in real-time.

Five labels are selected for the labeling process: swallowing saliva (S), swallowing water (SW), eating yogurt (SY), coughing (C), and throat clearing (T). The labeling process is performed by checking the audio signals in conjunction with the camera signal to detect and select the time range of the events. With these labels, we can select parts of the signals that contain relevant information for the problem.

5.1.3 Pre-processing and feature extraction

Due to the sensor's nature, a Hilbert transformation is applied; when the absolute value of this transformation is taken, the result is the signal's envelope. Furthermore, a low-pass filter with a cutoff frequency of 40 Hz is applied to remove the power supply's hum and any higher frequency used in the modulation process.

From the pre-processed signal, spikes can be seen at the signal's beginning and end. This is not caused by the swallowing phenomena but by the convolution process that is performed in the Hilbert transform because the signal doesn't have values before $t = 0$ for the convolution; a circular or periodic padding is applied internally where the left part of the signal joins the right, creating this spikes. To remove this, 2400 samples (440 ms) are removed from the transformed audio (1200 at the beginning and 1200 at the end).

A set of features is required to analyze the signals using a machine-learning model. These features aim to represent the feature in different characteristics that can be used to check for patterns that allow us to classify between different classes.

5.1.4 Learning schemes

The same pipeline used in this work is applied here for swallowing detection. Three machine learning methods are used to classify the swallowing events. A support vector machine, a decision tree, and a random forest classifier. SVMs excel at creating hyper-planes to separate data points; these are made

using three hyper-parameters: C, γ, and the kernel type. On the other hand, DT classifiers infer decision rules to predict the target label; these rules are the different "leaves" of the tree, and usually, to get to a specific leaf, the features for the signal needed to pass certain thresholds defined by the classifier. Lastly, RF classifiers combine multiple trees to improve the prediction and prevent the overfitting that the DT classifiers can cause. Still, these classifiers are slower, and we lose the interpretability of the decision trees.

Each classifier requires a hyper-parameter optimization to find the best classifier for our data. Two cross-validation methods are used: Leave One Subject Out (LOSO) and a Stratified Group K-Fold. Both methods split the data into train and test, always keeping all subject samples in one set. LOSO selects one subject for the test set and leaves the rest for the training, while the Stratified K-Fold selects a particular group of subjects for testing depending on the number of folds required; for this work, a total of 5 folds is used.

Hyper-parameter optimization is conducted via grid search to identify the optimal parameter set, with the mode (most frequently occurring value) used to select the final parameters. After the grid search, a model is created with the hyper-parameters fixed and tested in each cross-validation fold.

Each sensor is employed independently for the classification task, obtaining unimodal results. Afterward, early and late fusion techniques merge and use both sensors' information simultaneously. Early fusion concatenates the features of both sensors and uses them as input for the classifiers. On the other hand, late fusion uses the probabilities given by the classifiers trained with each sensor and performs the classification based on the joint probability.

This is a work in progress and the results aim to find relevant information in the magnetic sensors that can lead to more research in the usage of these new types of sensors.

5.2 Tables from uni-modal experiments

Table 5.1. Performance of classical classifiers using phonation features for the vowel /a/ task. **DT**: Decision Tree, **RF**: Random Forest, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity

Signal	Classifier	Parameters*	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Speech	DT	criterion = gini max depth = 3 min samples split = 10 max leaf nodes = 9	58.7 ± 9.3	54.7 ± 11.4	32.9 ± 16.9	83.8 ± 12.8
	RF	max features = sqrt max depth = 90 min samples split = 4	61.1 ± 6.7	60.3 ± 6.9	52.4 ± 13.2	69.7 ± 11.1
	SVM	C = 10 γ = 0.001 kernel = rbf	57.4 ± 9.2	55.8 ± 10.2	46.1 ± 22.4	70.3 ± 11.1
EGG	DT	criterion = entropy max depth = 3 min samples split = 2 max leaf nodes = 5	58.9 ± 9.5	58.1 ± 9.9	50.5 ± 18.9	67.5 ± 7.7
	RF	max features = sqrt max depth = 40 min samples split = 2	58.2 ± 11.8	52.2 ± 15.1	29.6 ± 22.7	85.3 ± 18.2
	SVM	C = 1 $\gamma = 1e^{-5}$ kernel = rbf	60.9 ± 12.8	54.7 ± 16.4	30.0 ± 26.0	90.8 ± 7.8

Table 5.2. Performance of classical classifiers using phonation features for the vowel /i/ task. **DT**: Decision Tree, **RF**: Random Forest, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity

Signal	Classifier	Parameters*	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Speech	DT	criterion = entropy max depth = 3 min samples split = 2 max leaf nodes = 3	55.2 ± 13.3	54.5 ± 13.7	45.5 ± 18.4	64.9 ± 11.8
	RF	max features = log2 max depth = 70 min samples split = 2	54.4 ± 4.4	52.5 ± 4.1	43.4 ± 18.5	65.3 ± 16.2
	SVM	C = 1 $\gamma = 1e^{-5}$ kernel = linear	61.6 ± 14.2	54.9 ± 18.2	30.4 ± 28.2	95.0 ± 5.9
EGG	DT	criterion = entropy max depth = 3 min samples split = 2 max leaf nodes = 3	58.1 ± 8.5	56.5 ± 9.2	44.8 ± 21.0	71.5 ± 7.2
	RF	max features = sqrt max depth = 40 min samples split = 6	56.1 ± 10.0	54.1 ± 11.2	42.2 ± 22.9	70.1 ± 9.8
	SVM	C = 0.001 $\gamma = 1e^{-5}$ kernel = linear	52.8 ± 5.1	49.0 ± 5.9	30.8 ± 17.6	74.7 ± 16.7

Table 5.3. Performance of classical classifiers using phonation features for the vowel /u/ task. **DT**: Decision Tree, **RF**: Random Forest, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity

Signal	Classifier	Parameters*	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Speech	DT	criterion = entropy max depth = 5 min samples split = 4 max leaf nodes = 11	62.6 ± 13.9	57.4 ± 17.5	33.1 ± 24.7	91.6 ± 5.0
	RF	max features = log2 max depth = 40 min samples split = 6	55.4 ± 11.6	51.6 ± 13.0	29.7 ± 15.5	80.6 ± 13.3
	SVM	C = 1 γ = 0.1 kernel = rbf	54.7 ± 6.5	53.7 ± 7.3	49.1 ± 18.3	60.5 ± 7.4
EGG	DT	criterion = gini max depth = 3 min samples split = 2 max leaf nodes = 3	58.2 ± 11.2	49.9 ± 17.4	46.0 ± 36.2	71.3 ± 36.2
	RF	max features = sqrt max depth = 50 min samples split = 2	43.1 ± 10.7	42.5 ± 11.0	40.1 ± 17.8	46.0 ± 9.4
	SVM	C = 1 γ = 0.01 kernel = rbf	61.3 ± 9.7	60.6 ± 9.6	57.3 ± 18.4	65.3 ± 13.9

Table 5.4. Performance of classical classifiers using phonation features for the phrase task. **DT**: Decision Tree, **RF**: Random Forest, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity

Signal	Classifier	Parameters*	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Speech	DT	criterion = gini max depth = 5 min samples split = 10 max leaf nodes = 9	61.0 ± 11.0	56.6 ± 13.4	33.0 ± 20.7	88.6 ± 6.2
	RF	max features = log2 max depth = 90 min samples split = 6	58.4 ± 11.4	57.3 ± 11.7	47.7 ± 18.8	69.2 ± 13.6
	SVM	C = 1 γ = 0.001 kernel = rbf	54.5 ± 7.0	47.9 ± 9.6	20.4 ± 12.9	88.1 ± 5.1
EGG	DT	criterion = entropy max depth = 9 min samples split = 2 max leaf nodes = 17	45.7 ± 6.8	42.2 ± 7.1	57.3 ± 23.8	34.1 ± 21.7
	RF	max features = sqrt max depth = 40 min samples split = 2	56.2 ± 15.7	55.1 ± 16.2	48.3 ± 22.7	64.1 ± 18.3
	SVM	C = 10 γ = 0.01 kernel = rbf	63.0 ± 8.0	62.2 ± 8.3	61.2 ± 18.3	65.0 ± 13.7

Table 5.5. Results using articulation features for the vowel /a/ task. **DT**: Decision Tree, **RF**: Random Forest, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity

Signal	Classifier	Parameters*	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Speech	DT	criterion = gini max depth = 3 min samples split = 2 max leaf nodes = 5	58.2 ± 11.8	52.2 ± 15.1	29.6 ± 22.7	85.3 ± 18.2
	RF	max features = sqrt max depth = 40 min samples split = 10	56.1 ± 10.0	54.1 ± 11.2	42.2 ± 22.9	70.1 ± 9.8
	SVM	C = 0.001 $\gamma = 1e^{-5}$ kernel = rbf	51.8 ± 3.4	35.5 ± 4.0	0.0 ± 0.0	100.0 ± 0.0
EGG	DT	criterion = gini max depth = 3 min samples split = 2 max leaf nodes = 9	64.6 ± 12.6	63.7 ± 13.7	54.8 ± 20.1	74.5 ± 6.5
	RF	max features = sqrt max depth = 10 min samples split = 4	59.3 ± 12.9	57.6 ± 14.1	47.1 ± 25.2	71.5 ± 6.4
	SVM	C = 0.1 $\gamma = 1e^{-5}$ kernel = rbf	57.8 ± 9.4	47.7 ± 15.9	39.1 ± 35.8	77.3 ± 38.8

Table 5.6. Results using articulation features for the vowel /i/ task. **DT**: Decision Tree, **RF**: Random Forest, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity

Signal	Classifier	Parameters*	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Speech	DT	criterion = gini max depth = 5 min samples split = 2 max leaf nodes = 9	58.0 ± 14.6	49.7 ± 19.6	25.3 ± 27.6	92.9 ± 9.5
	RF	max features = sqrt max depth = 90 min samples split = 2	57.4 ± 9.2	55.8 ± 10.2	46.1 ± 22.4	70.3 ± 11.1
	SVM	C = 0.01 $\gamma = 1e^{-5}$ kernel = linear	61.6 ± 14.2	54.9 ± 18.2	30.4 ± 28.2	95.0 ± 5.9
EGG	DT	criterion = entropy max depth = 3 min samples split = 2 max leaf nodes = 3	56.7 ± 11.2	50.2 ± 16.4	34.8 ± 33.3	80.3 ± 15.8
	RF	max features = sqrt max depth = 60 min samples split = 2	56.8 ± 7.7	55.6 ± 7.9	52.0 ± 22.2	62.9 ± 10.6
	SVM	C = 0.001 $\gamma = 1e^{-5}$ kernel = linear	58.0 ± 11.3	57.3 ± 11.6	54.2 ± 21.0	63.1 ± 12.7

Table 5.7. Results using articulation features for the vowel /u/ task. **DT**: Decision Tree, **RF**: Random Forest, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity

Signal	Classifier	Parameters*	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Speech	DT	criterion = entropy max depth = 3 min samples split = 2 max leaf nodes = 19	61.6 ± 10.2	57.1 ± 12.8	34.3 ± 21.5	88.1 ± 9.9
	RF	max features = sqrt max depth = 70 min samples split = 2	58.8 ± 9.6	57.1 ± 10.4	44.7 ± 20.3	72.8 ± 11.2
	SVM	C = 0.1 γ = 0.01 kernel = rbf	60.9 ± 12.8	54.7 ± 16.4	30.0 ± 26.0	90.8 ± 7.8
EGG	DT	criterion = entropy max depth = 3 min samples split = 2 max leaf nodes = 3	58.0 ± 13.1	49.7 ± 18.0	28.2 ± 30.4	87.8 ± 19.3
	RF	max features = log2 max depth = 10 min samples split = 6	59.7 ± 10.8	57.7 ± 11.8	46.7 ± 26.1	72.8 ± 9.5
	SVM	C = $1e^{-5}$ γ = 0.01 kernel = rbf	59.9 ± 14.6	51.7 ± 20.7	48.6 ± 37.8	72.1 ± 36.5

Table 5.8. Results using articulation features for the phrase task. **DT**: Decision Tree, **RF**: Random Forest, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity

Signal	Classifier	Parameters*	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Speech	DT	criterion = entropy max depth = 9 min samples split = 2 max leaf nodes = 17	52.8 ± 5.1	49.0 ± 5.9	30.8 ± 17.6	74.7 ± 16.7
	RF	max features = sqrt max depth = 40 min samples split = 2	52.7 ± 7.3	51.4 ± 7.7	41.9 ± 16.8	63.6 ± 12.9
	SVM	C = 10 γ = 0.01 kernel = rbf	55.7 ± 5.2	53.7 ± 6.6	40.6 ± 18.4	70.6 ± 11.8
EGG	DT	criterion = gini max depth = 3 min samples split = 2 max leaf nodes = 17	57.3 ± 12.8	55.8 ± 13.7	46.3 ± 24.0	68.4 ± 11.5
	RF	max features = log2 max depth = 20 min samples split = 4	60.8 ± 13.1	59.3 ± 13.7	49.8 ± 24.6	71.9 ± 13.0
	SVM	C = 10 γ = 0.001 kernel = rbf	64.3 ± 10.2	63.4 ± 10.7	55.0 ± 18.8	73.7 ± 11.6

Table 5.9. Results using BFCC features for the vowel /a/ task. **DT**: Decision Tree, **RF**: Random Forest, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity

Signal	Classifier	Parameters*	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Speech	DT	criterion = gini max depth = 3 min samples split = 2 max leaf nodes = 5	52.8 ± 7.8	49.0 ± 9.2	38.8 ± 25.0	66.8 ± 21.4
	RF	max features = sqrt max depth = 40 min samples split = 10	56.7 ± 10.3	55.6 ± 10.8	49.3 ± 21.0	64.1 ± 10.2
	SVM	C = 0.001 $\gamma = 1e^{-5}$ kernel = rbf	58.2 ± 9.0	50.4 ± 16.0	50.4 ± 34.9	66.9 ± 36.3
EGG	DT	criterion = entropy max depth = 3 min samples split = 2 max leaf nodes = 3	58.7 ± 9.2	55.0 ± 11.5	41.8 ± 24.9	75.5 ± 20.2
	RF	max features = sqrt max depth = 10 min samples split = 10	55.4 ± 10.8	54.1 ± 11.1	46.7 ± 21.9	64.0 ± 11.6
	SVM	C = 1 $\gamma = 0.1$ kernel = rbf	60.6 ± 9.8	59.3 ± 10.2	51.9 ± 21.9	69.3 ± 14.2

Table 5.10. Results using BFCC features for the vowel /i/ task. **DT**: Decision Tree, **RF**: Random Forest, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity

Signal	Classifier	Parameters*	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Speech	DT	criterion = gini max depth = 5 min samples split = 2 max leaf nodes = 9	50.2 ± 6.2	49.8 ± 6.1	50.8 ± 7.0	49.6 ± 12.7
	RF	max features = sqrt max depth = 90 min samples split = 2	58.0 ± 9.8	57.5 ± 9.8	55.5 ± 14.0	60.6 ± 15.6
	SVM	C = 0.01 $\gamma = 1e^{-5}$ kernel = linear	57.6 ± 7.3	57.2 ± 7.2	49.9 ± 9.3	65.5 ± 12.0
EGG	DT	criterion = entropy max depth = 5 min samples split = 2 max leaf nodes = 19	51.2 ± 7.5	47.1 ± 8.3	28.3 ± 15.1	74.0 ± 18.2
	RF	max features = sqrt max depth = 40 min samples split = 2	51.2 ± 7.5	51.0 ± 7.6	48.6 ± 12.3	53.8 ± 6.8
	SVM	C = 10 $\gamma = 1$ kernel = rbf	54.9 ± 7.8	53.1 ± 8.7	39.8 ± 16.0	70.0 ± 10.1

Table 5.11. Results using BFCC features for the vowel /u/ task. **DT**: Decision Tree, **RF**: Random Forest, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity

Signal	Classifier	Parameters*	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Speech	DT	criterion = entropy max depth = 3 min samples split = 2 max leaf nodes = 19	55.6 ± 4.9	51.5 ± 6.9	32.8 ± 17.1	78.4 ± 17.0
	RF	max features = sqrt max depth = 70 min samples split = 2	53.4 ± 8.2	52.5 ± 8.8	45.8 ± 17.2	60.9 ± 7.5
	SVM	C = 0.1 γ = 0.01 kernel = rbf	59.1 ± 7.0	56.9 ± 8.7	42.0 ± 17.6	76.3 ± 9.5
EGG	DT	criterion = entropy max depth = 5 min samples split = 2 max leaf nodes = 15	48.1 ± 10.5	44.9 ± 11.1	34.9 ± 21.0	61.2 ± 24.0
	RF	max features = sqrt max depth = 70 min samples split = 2	49.2 ± 9.3	48.8 ± 9.4	44.6 ± 13.3	54.0 ± 10.2
	SVM	C = $1e^{-5}$ γ = $1e^{-5}$ kernel = rbf	51.2 ± 4.7	36.0 ± 8.3	23.4 ± 39.2	80.0 ± 40.0

Table 5.12. Results using BFCC features for the phrase task. **DT**: Decision Tree, **RF**: Random Forest, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity

Signal	Classifier	Parameters*	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Speech	DT	criterion = entropy max depth = 5 min samples split = 8 max leaf nodes = 19	50.8 ± 11.6	49.6 ± 11.4	39.4 ± 12.8	62.2 ± 19.3
	RF	max features = sqrt max depth = 80 min samples split = 2	57.2 ± 8.8	56.8 ± 8.8	55.1 ± 15.1	59.2 ± 9.6
	SVM	C = 10 $\gamma = 1e^{-5}$ kernel = linear	59.3 ± 6.3	58.9 ± 6.2	62.1 ± 13.3	56.5 ± 9.1
EGG	DT	criterion = gini max depth = 3 min samples split = 2 max leaf nodes = 3	54.7 ± 8.5	52.5 ± 9.1	42.0 ± 21.9	67.6 ± 14.9
	RF	max features = log2 max depth = 60 min samples split = 2	54.3 ± 9.4	53.7 ± 9.5	51.7 ± 17.6	57.0 ± 11.9
	SVM	C = 10 $\gamma = 0.01$ kernel = rbf	58.9 ± 9.4	55.2 ± 11.1	34.1 ± 19.3	83.7 ± 6.7

Table 5.13. Results using nonlinear features for the vowel /a/ task. **DT**: Decision Tree, **RF**: Random Forest, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity

Signal	Classifier	Parameters*	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Speech	DT	criterion = gini max depth = 3 min samples split = 2 max leaf nodes = 9	54.1 ± 9.4	52.7 ± 10.7	47.5 ± 19.7	60.3 ± 13.2
	RF	max features = sqrt max depth = 10 min samples split = 4	57.5 ± 10.1	56.0 ± 10.6	50.6 ± 25.2	64.5 ± 12.9
	SVM	C = 0.1 $\gamma = 1e^{-5}$ kernel = rbf	57.8 ± 11.9	49.6 ± 18.0	47.3 ± 35.7	69.1 ± 36.5
EGG	DT	criterion = entropy max depth = 3 min samples split = 2 max leaf nodes = 3	56.2 ± 11.1	50.7 ± 13.7	32.7 ± 29.8	79.8 ± 16.6
	RF	max features = log2 max depth = 60 min samples split = 4	57.5 ± 10.6	56.5 ± 11.1	51.1 ± 22.8	63.9 ± 8.4
	SVM	C = 1 $\gamma = 1e^{-5}$ kernel = linear	60.2 ± 13.4	52.3 ± 18.2	24.8 ± 24.7	95.5 ± 3.9

Table 5.14. Results using nonlinear features for the vowel /i/ task. **DT**: Decision Tree, **RF**: Random Forest, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity

Signal	Classifier	Parameters*	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Speech	DT	criterion = entropy max depth = 3 min samples split = 2 max leaf nodes = 3	62.2 ± 13.6	56.3 ± 18.3	32.2 ± 24.3	95.1 ± 4.9
	RF	max features = sqrt max depth = 60 min samples split = 2	56.9 ± 8.7	54.9 ± 8.9	51.5 ± 26.7	63.8 ± 18.8
	SVM	C = 0.001 $\gamma = 1e^{-5}$ kernel = linear	60.3 ± 12.2	58.4 ± 13.4	48.9 ± 26.3	73.4 ± 11.8
EGG	DT	criterion = entropy max depth = 3 min samples split = 2 max leaf nodes = 5	60.5 ± 7.8	59.6 ± 8.2	63.0 ± 21.5	58.6 ± 12.4
	RF	max features = sqrt max depth = 10 min samples split = 10	58.2 ± 7.2	56.8 ± 7.6	55.9 ± 24.3	60.9 ± 13.8
	SVM	C = 1 $\gamma = 1$ kernel = rbf	58.3 ± 8.5	55.4 ± 10.1	42.3 ± 25.1	75.9 ± 11.9

Table 5.15. Results using nonlinear features for the vowel /u/ task. **DT**: Decision Tree, **RF**: Random Forest, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity

Signal	Classifier	Parameters*	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Speech	DT	criterion = entropy max depth = 3 min samples split = 2 max leaf nodes = 3	62.2 ± 12.5	57.5 ± 15.8	37.5 ± 27.7	86.8 ± 9.8
	RF	max features = log2 max depth = 10 min samples split = 6	59.1 ± 10.3	57.2 ± 11.3	47.1 ± 25.5	71.0 ± 8.2
	SVM	C = $1e^{-5}$ γ = 0.01 kernel = rbf	59.5 ± 11.9	52.4 ± 18.5	57.3 ± 36.1	62.6 ± 34.3
EGG	DT	criterion = entropy max depth = 3 min samples split = 2 max leaf nodes = 3	59.0 ± 13.7	53.7 ± 17.3	33.1 ± 29.0	85.0 ± 8.4
	RF	max features = log2 max depth = 10 min samples split = 2	55.8 ± 8.8	54.8 ± 9.2	50.7 ± 19.7	61.0 ± 12.9
	SVM	C = 10 γ = $1e^{-5}$ kernel = rbf	54.3 ± 6.4	41.8 ± 11.4	30.0 ± 36.9	79.5 ± 39.8

Table 5.16. Results using nonlinear features for the phrase task. **DT**: Decision Tree, **RF**: Random Forest, **SVM**: Support Vector Machine, **ACC**: Accuracy, **F1**: F1-score, **SEN**: Sensitivity, **SPE**: Specificity

Signal	Classifier	Parameters*	ACC [%]	F1 [%]	SEN [%]	SPE [%]
Speech	DT	criterion = gini max depth = 7 min samples split = 2 max leaf nodes = 9	55.2 ± 6.8	52.7 ± 8.1	43.8 ± 20.9	66.8 ± 18.6
	RF	max features = log2 max depth = 10 min samples split = 8	58.9 ± 11.9	57.8 ± 12.6	49.4 ± 21.5	68.4 ± 9.5
	SVM	C = $1e^{-5}$ γ = 0.01 kernel = rbf	61.0 ± 10.9	54.0 ± 18.5	59.5 ± 34.4	63.4 ± 34.4
EGG	DT	criterion = gini max depth = 7 min samples split = 2 max leaf nodes = 9	56.2 ± 14.3	54.0 ± 14.7	39.3 ± 20.2	73.1 ± 19.1
	RF	max features = sqrt max depth = 30 min samples split = 10	59.1 ± 10.3	58.4 ± 10.5	54.7 ± 17.6	63.5 ± 13.2
	SVM	C = 10 γ = 10 kernel = rbf	59.3 ± 9.5	59.1 ± 9.6	60.8 ± 11.7	57.7 ± 10.8

List of Figures

List of Tables

Bibliography

[1] E. Vilkman, "Voice problems at work: A challenge for occupational safety and health arrangement," *Folia Phoniatrica et Logopaedica*, vol. 52, no. 1-3, pp. 120–125, Aug. 1999. DOI: 10.1159/000021519. [Online]. Available: https://doi.org/10.1159/000021519.

[2] A. Rameau, R. S. Foltz, K. Wagner, and K. B. Zur, "Multidisciplinary approach to vocal cord dysfunction diagnosis and treatment in one session: A single institutional outcome study," *International journal of pediatric otorhinolaryngology*, vol. 76, no. 1, pp. 31–35, 2012.

[3] M. Eye and E. Infirmary, "Voice disorders database, version. 1.03 (cd-rom)," *Lincoln Park, NJ: Kay Elemetrics Corporation*, 1994.

[4] T. A. Mesallam, M. Farahat, K. H. Malki, *et al.*, "Development of the arabic voice pathology database and its evaluation by using speech features and machine learning algorithms," *Journal of healthcare engineering*, vol. 2017, 2017.

[5] B. Woldert-Jokisz, "Saarbruecken voice database," 2007.

[6] N. Saenz-Lechon, J. I. Godino-Llorente, V. Osma-Ruiz, and P. Gomez-Vilda, "Methodological issues in the development of automatic systems for voice pathology detection," *Biomedical Signal Processing and Control*, vol. 1, no. 2, pp. 120–128, 2006.

[7] 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 Vocology*, vol. 36, no. 2, pp. 60–69, 2011.

[8] M. M. Hakkesteegt, M. P. Brocaar, M. H. Wieringa, and L. Feenstra, "The relationship between perceptual evaluation and objective multiparametric evaluation of dysphonia severity," *Journal of Voice*, vol. 22, no. 2, pp. 138–145, 2008.

[9] N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, M. Blanco-Velasco, and F. Cruz-Roldán, "Automatic assessment of voice quality according to the grbas scale," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2006, pp. 2478–2481.

[10] M. Markaki and Y. Stylianou, "Using modulation spectra for voice pathology detection and classification," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2009, pp. 2514–2517.

[11] D. Martínez, E. Lleida, A. Ortega, A. Miguel, and J. Villalba, "Voice pathology detection on the saarbrücken voice database with calibration and fusion of scores using multifocal toolkit," in *Advances in speech and language technologies for Iberian Languages*, Springer, 2012, pp. 99–109.

[12] N. Souissi and A. Cherif, "Speech recognition system based on short-term cepstral parameters, feature reduction method and artificial neural networks," in *2016 2nd international conference on advanced technologies for signal and image processing (ATSIP)*, IEEE, 2016, pp. 667–671.

[13] L. Verde, G. De Pietro, and G. Sannino, "Voice disorder identification by using machine learning techniques," *IEEE access*, vol. 6, pp. 16 246–16 255, 2018.

[14] P. Harar, J. B. Alonso-Hernandezy, J. Mekyska, Z. Galaz, R. Burget, and Z. Smekal, "Voice pathology detection using deep learning: A preliminary study," in *2017 international conference and workshop on bioinspired intelligence (IWOBI)*, IEEE, 2017, pp. 1–4.

[15] M. Alhussein and G. Muhammad, "Voice pathology detection using deep learning on mobile healthcare framework," *IEEE Access*, vol. 6, pp. 41 034–41 041, 2018.

[16] D. Korzekwa, R. Barra-Chicote, B. Kostek, T. Drugman, and M. Lajszczak, "Interpretable deep learning model for the detection and reconstruction of dysarthric speech," *arXiv preprint arXiv:1907.04743*, 2019.

[17] J.-Y. Lee, "Experimental evaluation of deep learning methods for an intelligent pathological voice detection system using the saarbruecken voice database," *Applied Sciences*, vol. 11, no. 15, p. 7149, 2021.

[18] M. S. Hossain, G. Muhammad, and A. Alamri, "Smart healthcare monitoring: A voice pathology detection paradigm for smart cities," *Multimedia Systems*, vol. 25, no. 5, pp. 565–575, 2019.

[19] J.-N. Lee and J.-Y. Lee, "An efficient smote-based deep learning model for voice pathology detection," *Applied Sciences*, vol. 13, no. 6, p. 3571, 2023.

[20] G. S. Liu, J. M. Hodges, J. Yu, C. K. Sung, E. Erickson-DiRenzo, and P. C. Doyle, "End-to-end deep learning classification of vocal pathology using stacked vowels," *Laryngoscope Investigative Otolaryngology*, vol. 8, no. 5, pp. 1312–1318, 2023.

[21] A. N. Omeroglu, H. M. Mohammed, and E. A. Oral, "Multi-modal voice pathology detection architecture based on deep and handcrafted feature fusion," *Engineering Science and Technology, an International Journal*, vol. 36, p. 101 148, 2022.

[22] G. Ras, N. Xie, M. Van Gerven, and D. Doran, "Explainable deep learning: A field guide for the uninitiated," *Journal of Artificial Intelligence Research*, vol. 73, pp. 329–396, 2022.

[23] L. O. Ramig and K. Verdolini, "Treatment efficacy: Voice disorders," *Journal of Speech, Language, and Hearing Research*, vol. 41, no. 1, S101–S116, 1998.

[24] E. Seifert, "Stress and distress in non-organic voice disorder," *Swiss medical weekly*, vol. 135, no. 2728, pp. 387–397, 2005.

[25] J. I. Godino-Llorente, P. Gómez-Vilda, F. Cruz-Roldán, M. Blanco-Velasco, and R. Fraile, "Pathological likelihood index as a measurement of the degree of voice normality and perceived hoarseness," *Journal of Voice*, vol. 24, no. 6, pp. 667–677, 2010.

[26] S. L. Smith and I. R. Titze, "Vocal fold contact patterns based on normal modes of vibration," *Journal of biomechanics*, vol. 73, pp. 177–184, 2018.

[27] P. Kitzing, "Electroglottography," in *Diseases of the Larynx*, Arnold, 2000, pp. 127–138.

[28] K. Kuligowska, P. Kisielewicz, and A. Włodarz, "Speech synthesis systems: Disadvantages and limitations," *Int J Res Eng Technol (UAE)*, vol. 7, pp. 234–239, 2018.

[29] J. Benesty, M. M. Sondhi, Y. Huang, *et al.*, *Springer handbook of speech processing*. Springer, 2008, vol. 1.

[30] I. Cobeta, F. Núñez, and S. Fernández, *Patología de la voz*. Marge books, 2013.

[31] J. Van den Berg, "Myoelastic-aerodynamic theory of voice production," *Journal of speech and hearing research*, vol. 1, no. 3, pp. 227–244, 1958.

[32] C. A. Rosen and T. Murry, "Diagnostic laryngeal endoscopy," *Otolaryngologic Clinics of North America*, vol. 33, no. 4, pp. 751–757, 2000.

[33] M. Södersten and P. Lindestad, "A comparison of vocal fold closure in rigid telescopic and flexible fiberoptic laryngostroboscopy," *Acta oto-laryngologica*, vol. 112, no. 1, pp. 144–150, 1992.

[34] D. B. Hawkins and R. W. Clark, "Flexible laryngoscopy in neonates, infants, and young children," *Annals of Otology, Rhinology & Laryngology*, vol. 96, no. 1, pp. 81–85, 1987.

[35] J. F. Kaiser, "Some observations on vocal tract operation from a fluid flow point of view," *Vocal Fold Physiology: Biomechanics, Acoustics, and Phonatory Control*, pp. 358–386, 1983.

[36] H. Teager and S. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," *Speech production and speech modelling*, pp. 241–261, 1990.

[37] J. J. Jiang, Y. Zhang, and C. McGilligan, "Chaos in voice, from modeling to measurement," *Journal of Voice*, vol. 20, no. 1, pp. 2–17, 2006.

[38] S. S. Narayanan and A. A. Alwan, "A nonlinear dynamical systems analysis of fricative consonants," *The Journal of the Acoustical Society of America*, vol. 97, no. 4, pp. 2511–2524, 1995.

[39] B. Goswami, "A brief introduction to nonlinear time series analysis and recurrence plots," *Vibration*, vol. 2, no. 4, pp. 332–368, 2019.

[40] 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 Systems with Applications*, vol. 82, pp. 184–195, 2017.

[41] H. Kantz and T. Schreiber, *Nonlinear time series analysis*. Cambridge university press, 2004, vol. 7.

[42] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence, Warwick 1980: proceedings of a symposium held at the University of Warwick 1979/80*, Springer, 2006, pp. 366–381.

[43] A. Wolf, J. B. Swift, H. L. Swinney, and J. A. Vastano, "Determining lyapunov exponents from a time series," *Physica D: nonlinear phenomena*, vol. 16, no. 3, pp. 285–317, 1985.

[44] R. Taylor, "Attractors: Nonstrange to chaotic," *Society for industrial and applied mathematics, undergraduate research online*, pp. 72–80, 2010.

[45] A. Delgado-Bonal and A. Marshak, "Approximate entropy and sample entropy: A comprehensive tutorial," *Entropy*, vol. 21, no. 6, p. 541, 2019.

[46] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American journal of physiology-heart and circulatory physiology*, vol. 278, no. 6, H2039–H2049, 2000.

[47] J. S. Richman, D. E. Lake, and J. R. Moorman, "Sample entropy," in *Methods in enzymology*, vol. 384, Elsevier, 2004, pp. 172–184.

[48] H. E. Hurst, "A suggested statistical model of some time series which occur in nature," *Nature*, vol. 180, no. 4584, pp. 494–494, 1957.

[49] H. E. Hurst, "The problem of long-term storage in reservoirs," *Hydrological Sciences Journal*, vol. 1, no. 3, pp. 13–27, 1956.

[50] M. J. Owren, "Human voice in evolutionary perspective," *Acoust. Today*, vol. 7, no. 24, pp. 10–1121, 2011.

[51] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[52] J. C. Stemple, N. Roy, and B. K. Klaben, *Clinical voice pathology: Theory and management*. Plural Publishing, 2018.

[53] A. Chou, C. Schrof, E. Polce, M. Braden, J. McMurray, and J. Jiang, "Comparing the nonlinear dynamic acoustic parameters of healthy adult and pediatric voices," *Annals of Otology, Rhinology & Laryngology*, vol. 127, no. 12, pp. 937–945, 2018.

[54] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis–jitter, shimmer and hnr parameters," *Procedia Technology*, vol. 9, pp. 1112–1122, 2013.

[55] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition," in *8th Annual Conference of the International Speech Communication Association; 2007 Aug. 27-31; Antwerp (Belgium).[place unknown]: ISCA; 2007. p. 778-81.*, International Speech Communication Association (ISCA), 2007.

[56] B. H. Story, "An overview of the physiology, physics and modeling of the sound source for vowels," *Acoustical Science and Technology*, vol. 23, no. 4, pp. 195–206, 2002.

[57] B. Gick, I. Wilson, and D. Derrick, *Articulatory phonetics*. John Wiley & Sons, 2013.

[58] Y. Ji, J. Wei, J. Zhang, *et al.*, "Speech behavior analysis by articulatory observations," *Procedia computer science*, vol. 111, pp. 463–470, 2017.

[59] T. Murry, "Spasmodic dysphonia: Let's look at that again," *Journal of Voice*, vol. 28, no. 6, pp. 694–699, 2014.

[60] C. L. Ludlow, "Treatment for spasmodic dysphonia: Limitations of current approaches," *Current opinion in otolaryngology & head and neck surgery*, vol. 17, no. 3, pp. 160–165, 2009.

[61] T. Havas, D. Lowinger, and J. Priestley, "Unilateral vocal fold paralysis: Causes, options and outcomes," *Australian and New Zealand journal of surgery*, vol. 69, no. 7, pp. 509–513, 1999.

[62] S. Sapir, L. Ramig, and C. Fox, "Speech and swallowing disorders in parkinson disease," *Current opinion in otolaryngology & head and neck surgery*, vol. 16, no. 3, pp. 205–210, 2008.

[63] C. Moreau, C. Ozsancak, J.-L. Blatt, P. Derambure, A. Destee, and L. Defebvre, "Oral festination in parkinson's disease: Biomechanical analysis and correlation with festination and freezing of gait," *Movement disorders: official journal of the Movement Disorder Society*, vol. 22, no. 10, pp. 1503–1506, 2007.

[64] J. C. Vásquez-Correa, J. Orozco-Arroyave, T. Bocklet, and E. Nöth, "Towards an automatic evaluation of the dysarthria level of patients with parkinson's disease," *Journal of communication disorders*, vol. 76, pp. 21–36, 2018.

[65] L. R. Rabiner, R. W. Schafer, *et al.*, "Introduction to digital speech processing," *Foundations and Trends® in Signal Processing*, vol. 1, no. 1–2, pp. 1–194, 2007.

[66] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[67] R. Vergin and D. O'Shaughnessy, "Pre-emphasis and speech recognition," in *Proceedings 1995 Canadian Conference on Electrical and Computer Engineering*, IEEE, vol. 2, 1995, pp. 1062–1065.

[68] B. J. Shannon and K. K. Paliwal, "A comparative study of filter bank spacing for speech recognition," in *Microelectronic engineering research conference*, vol. 41, 2003, pp. 310–12.

[69] C. Kumar, F. Ur Rehman, S. Kumar, A. Mehmood, and G. Shabir, "Analysis of mfcc and bfcc in a speaker identification system," in *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, IEEE, 2018, pp. 1–5.

[70] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[71] T. Arias-Vergara, M. Döllinger, T. Schraut, K. A. M. Khairuddin, and A. Schützenberger, "Nyquist plot parametrization for quantitative analysis of vibration of the vocal folds," *Journal of Voice*, 2023.

[72] I. J. Moon and S. H. Hong, "What is temporal fine structure and why is it important?" *Korean journal of audiology*, vol. 18, no. 1, p. 1, 2014.

[73] K. A. M. Khairuddin, K. Ahmad, H. M. Ibrahim, and Y. Yan, "Description of the features and vibratory behaviors of the nyquist plot analyzed from laryngeal high-speed videoendoscopy images," *Journal of Voice*, vol. 36, no. 4, 582–e11, 2022.

[74] S. Tangirala, "Evaluating the impact of gini index and information gain on classification using decision tree classifier algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, pp. 612–619, 2020.

[75] Y. Yuan, L. Wu, and X. Zhang, "Gini-impurity index analysis," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 3154–3169, 2021.

[76] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[77] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE access*, vol. 8, pp. 4806–4813, 2019.

[78] K. O'shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[79] J. Gu, Z. Wang, J. Kuen, *et al.*, "Recent advances in convolutional neural networks," *Pattern recognition*, vol. 77, pp. 354–377, 2018.

[80] A. L. Maas, A. Y. Hannun, A. Y. Ng, *et al.*, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, Atlanta, GA, vol. 30, 2013, p. 3.

[81] N. Srivastava, "Improving neural networks with dropout," *University of Toronto*, vol. 182, no. 566, p. 7, 2013.

[82] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[83] A. Poernomo and D.-K. Kang, "Biased dropout and crossmap dropout: Learning towards effective dropout regularization in convolutional neural network," *Neural networks*, vol. 104, pp. 60–67, 2018.

[84] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[85] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia systems*, vol. 16, pp. 345–379, 2010.

[86] F. Krones, U. Marikkar, G. Parsons, A. Szmul, and A. Mahdi, "Review of multimodal machine learning approaches in healthcare," *arXiv preprint arXiv:2402.02460*, 2024.

[87] L. Wu, S. L. Oviatt, and P. R. Cohen, "Multimodal integration-a statistical view," *IEEE Transactions on Multimedia*, vol. 1, no. 4, pp. 334–341, 1999.

[88] N. Calvo-Ariza, T. Arias-Vergara, and J. Orozco-Arroyave, "Automatic assessment of voice disorders using phase plots," in *Workshop on Engineering Applications*, Springer, 2023, pp. 127–138.

[89] S. R. Kadiri and P. Alku, "Analysis and detection of pathological voice using glottal source features," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 367–379, 2019.

[90] M. Markaki and Y. Stylianou, "Voice pathology detection and discrimination based on modulation spectral features," *IEEE Transactions on audio, speech, and language processing*, vol. 19, no. 7, pp. 1938–1948, 2011.

[91] Y. Maryn, N. Roy, M. De Bodt, P. Van Cauwenberge, and P. Corthals, "Acoustic measurement of overall voice quality: A meta-analysis," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2619–2634, 2009.

[92] J. Hillenbrand and R. A. Houde, "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," *Journal of Speech, Language, and Hearing Research*, vol. 39, no. 2, pp. 311–321, 1996.

[93] L. Gavidia-Ceballos and J. H. Hansen, "Direct speech feature estimation using an iterative em algorithm for vocal fold pathology detection," *IEEE Transactions on Biomedical Engineering*, vol. 43, no. 4, pp. 373–383, 1996.

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

[95] D. A. Rahn III, M. Chou, J. J. Jiang, and Y. Zhang, "Phonatory impairment in parkinson's disease: Evidence from nonlinear dynamic analysis and perturbation analysis," *Journal of voice*, vol. 21, no. 1, pp. 64–71, 2007.

[96] Y. Zhang and J. J. Jiang, "Nonlinear dynamic mechanism of vocal tremor from voice analysis and model simulations," *Journal of sound and vibration*, vol. 316, no. 1-5, pp. 248–262, 2008.

[97] H. Herzel, D. Berry, I. R. Titze, and M. Saleh, "Analysis of vocal disorders with methods from nonlinear dynamics," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 5, pp. 1008–1019, 1994.

[98] P. Thaine and G. Penn, "Extracting mel-frequency and bark-frequency cepstral coefficients from encrypted signals.," in *Interspeech*, 2019, pp. 3715–3719.

[99] T. Villa-Canas, E. Belalcazar-Bolaños, S. Bedoya-Jaramillo, *et al.*, "Automatic detection of laryngeal pathologies using cepstral analysis in mel and bark scales," in *2012 XVII Symposium of Image, Signal Processing, and Artificial Vision (STSIVA)*, IEEE, 2012, pp. 116–121.

[100] I. R. Titze, "Nonlinear source–filter coupling in phonation: Theory," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 2733–2749, 2008.

[101] K. A. M. Khairuddin, K. Ahmad, S. C. Proehoeman, H. M. Ibrahim, and Y. Yan, "Preliminary findings of vocal fold vibratory characteristics of singers analyzed by laryngeal high-speed videoendoscopy," *Journal of Voice*, 2024.

[102] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[103] Q. Li, Y. Minagi, T. Ono, *et al.*, "The biomechanical coordination during oropharyngeal swallowing: An evaluation with a non-invasive sensing system," *Scientific reports*, vol. 7, no. 1, p. 15 165, 2017.

[104] R. F. Pfeiffer, "Neurogenic dysphagia," in *Bradley's Neurology in Clinical Practice*, R. B. Daroff, J. Jankovic, J. C. Mazziotta, and S. L. Pomeroy, Eds., Seventh Edition, Elsevier Health Sciences, 2016, 148–157.e2, ISBN: 978-0-323-28783-8. DOI: http://dx.doi.org/10.1016/B978-0-323-28783-8.00015-6.

[105] L. R. Carucci and M. A. Turner, "Dysphagia revisited: Common and unusual causes," *Radiographics*, vol. 35, no. 1, pp. 105–122, 2015.

[106] S. E. Langmore, "Evaluation of oropharyngeal dysphagia: Which diagnostic tool is superior?" *Current opinion in otolaryngology & head and neck surgery*, vol. 11, no. 6, pp. 485–489, 2003.

[107] Y. Morishima, K. Chida, and H. Watanabe, "Estimation of the dose of radiation received by patient and physician during a videofluoroscopic swallowing study," *Dysphagia*, vol. 31, pp. 574–578, 2016.

[108] A. Nacci, F. Ursino, R. La Vela, F. Matteucci, V. Mallardi, and B. Fattori, "Fiberoptic endoscopic evaluation of swallowing (fees): Proposal for informed consent," *Acta Otorhinolaryngologica Italica*, vol. 28, no. 4, p. 206, 2008.

[109] S. Roldan-Vasco, A. Orozco-Duque, J. C. Suarez-Escudero, and J. R. Orozco-Arroyave, "Machine learning based analysis of speech dimensions in functional oropharyngeal dysphagia," *Computer Methods and Programs in Biomedicine*, vol. 208, p. 106 248, 2021.

[110] H.-Y. Park, D. Park, H. S. Kang, H. Kim, S. Lee, and S. Im, "Post-stroke respiratory complications using machine learning with voice features from mobile devices," *Scientific Reports*, vol. 12, no. 1, p. 16 682, 2022.

[111] Y. Khalifa, J. L. Coyle, and E. Sejdić, "Non-invasive identification of swallows via deep learning in high resolution cervical auscultation recordings," *Scientific Reports*, vol. 10, no. 1, p. 8704, 2020.

[112] S. Mao, A. Sabry, Y. Khalifa, J. L. Coyle, and E. Sejdic, "Estimation of laryngeal closure duration during swallowing without invasive x-rays," *Future Generation Computer Systems*, vol. 115, pp. 610–618, 2021.

[113] S. Sarraf Shirazi, A. H. Birjandi, and Z. Moussavi, "Noninvasive and automatic diagnosis of patients at high risk of swallowing aspiration," *Medical & biological engineering & computing*, vol. 52, pp. 459–465, 2014.

[114] T. T. Frakking, A. B. Chang, C. Carty, *et al.*, "Using an automated speech recognition approach to differentiate between normal and aspirating swallowing sounds recorded from digital cervical auscultation in children," *Dysphagia*, vol. 37, no. 6, pp. 1482–1492, 2022.