POLITECNICO DI TORINO

Master's Degree in Ingegneria del Cinema
e dei Mezzi di comunicazione

Master's Degree Thesis

Interpretable acoustic features for depression detection: a comparative study of healthy & Parkinson's disease individuals

Supervisors

Prof. ANTONIO SERVETTI

Dr. MATHEW MAGIMAI DOSS

Candidate

BARBARA RUVOLO

ACADEMIC YEAR 2022-2023

A Giulia

Acknowledgements

I want to begin by expressing my gratitude to Dr. Mathew Magimai Doss, my project supervisor, for giving me the opportunity to undertake this exciting project. I am deeply thankful for his trust in me and the numerous ideas and teachings he has generously shared with me over the past months. I would also like to thank my supervisors, Tilak Purohit and Bogdan Vlasenko, for their patience, empathy, invaluable insights and thoughtful suggestions. Their help has significantly enhanced the quality of this work. Inoltre ringrazio il Prof. Antonio Servetti che ha riposto fiducia in me e mi ha accompagnato lungo questo percorso.

Ringrazio profondamente la mia famiglia tutta al femminile: nonna Peppina, mamma, Giulia, zia Carla e Licia; tutti i risultati raggiunti fino ad oggi, compreso questo, sono il frutto della dedizione, del sacrificio, della passione e dell'amore che mi avete dato e che ho appreso imitandovi e guardandovi quotidianamente lottare e andare avanti. Giulia, a te dedico le mie vittorie e le mie sconfitte, perché le ho sempre affrontate sapendoti vicina. Durante tutto il lavoro di tesi sei stato il mio punto di riferimento e il mio unico e dedito pensiero. Odino e Nello, forse voi non lo potete capire, ma vi ringrazio di averci salvato la vita.

Ringrazio dal profondo del cuore Giovanni; tu hai creduto in me quando ancora ero solo una bambina, mi hai cresciuta nei valori dello sport e del rispetto altrui e poi mi hai guardata da lontano senza mai lasciarmi sola.

Ai miei amici da una vita Alessia, Davide, Fabiola, Gianluca e Sandro; voi siete i pochi che hanno visto e vissuto tutte le versioni di me e mi avete sempre rispettata e stimata, senza mai giudicarmi; vi ringrazio perché per me quella che mi avete dimostrato è forma d'amore più sincera che esista.

Ringrazio Cecilia, Isabella e Francesca, colleghe e amiche per l'affetto che avete avuto nei miei confronti sin dal primo giorno e l'accoglienza calorosa nelle vostre case e famiglie. Francesca ti ringrazio per ogni singolo momento, risata e pianto che ci ha portate qui insieme.

Doveroso ringraziare i miei amici di Losanna Linda, Camilla, Francesco, Antonello, Roberto M., Matteo, Filippo, Tommaso, Chiara, Elena, Roberto C. e Arturo. Sento che questo risultato l'ho raggiunto al vostro fianco; non dimenticherò mai la gioia nei vostri occhi nel sapere che sarei rimasta più tempo in

Svizzera, e vi ringrazio per ogni parola, feedback, pausa caffè, cena, trekking e risata condivisa insieme. In particolar modo ringrazio il supporto di Camilla, imprescindibile, inestimabile e costantemente presente durante questi mesi, che mi è stata quotidianamente accanto come amica, collega e confidente.

Un pensiero speciale va a Giacomo e Jordi; sono profondamente grata alla vita per averci fatto incontrare, perché è accanto a voi che ho imparato a credere in me stessa ed è ispirandomi a voi che ho imparato a sognare in grande. Grazie per avermi regalato l'anno migliore della mia vita.

Lucia, Daniele, in silenzio vi siete fatti pian piano spazio nel mio cuore, con un affetto e una premura che ha lasciato il segno. Daniele ti ringrazio per il costante supporto, ma ancor di più per avermi ricordata quanto meravigliosa e leggera possa essere la vita se affrontata con le persone giuste. Lucia, ti ringrazio per avermi presa per mano ogni singolo istante in cui ne ho avuto di bisogno, per avermi incoraggiata ad affrontare quello che non pensavo di poter fare in momenti in cui non ho creduto abbastanza in me, e per avermi ascoltata e dato sempre i giusti consigli quando non ho saputo prendere le decisioni da sola.

Abstract

This research investigates using speech analysis to detect depression non-invasively and reliably by examining acoustic traits in healthy speech associated with depression and distinguishing them from pathological speech patterns.

The study compares two methods for depression detection: the first method involves extracting handcrafted features from the audio signal and utilizing diverse machine learning models for classification, while the second method involves using convolutional neural networks (CNNs) to model source, filter, and overall combined information from three different input signals.

The results suggest that depression in healthy patients can be better classified with features that carry information related to the vocal source of the signal. On the other hand, for Parkinson's disease patients, presumably due to the effects of the disease, retaining the full spectral content of the patient's voice is more effective in identifying depression.

Table of Contents

Chapter 1

Introduction

Depressive disorder (also known as depression) is a common mental disorder, affecting about 5% of the global adult population [1]. Depression is characterized by an individual's struggle to cope with challenging life circumstances, resulting in enduring emotions of sadness, pessimism, and an inability to effectively handle daily obligations. If not addressed timely, it often elevates the risk of an individual resorting to suicidal behaviors [2]. The World Health Organization (WHO) predicted depression to be the second most significant disability worldwide by 2030 [3].

Timely identification and assessment of depression play a pivotal role in effective treatment [4] [5]. Patient follow-ups and systematic symptom monitoring are essential for making informed treatment choices and evaluating treatment progress [6] [7].

Even today, depression diagnosis relies exclusively on the clinical assessment [8]. Hong et al.'s report [7] highlights that standardized scales in research and clinical settings depend on either self-reported or clinician-reported scores to improve the diagnosis and monitoring of symptoms. While these scales help minimize bias, there remains a potential for subjective variation in interpreting items during clinical interviews, leading to variability in diagnosis [9]. This bias further affects the assessment in longitudinal treatments. Furthermore, the shortage of resources and well-trained practitioners poses a significant challenge to effectively diagnosing and monitoring depression patients [2]. Currently, there exists no objective measure for the clinical detection of depression [4]. Therefore, bolstering the existing diagnostic approaches with a dependable, affordable, automated screening tool or framework for detecting and monitoring depression holds particular importance.

Recent research suggests significant potential in leveraging speech, specifically the non-verbal paralinguistic cues for automatically detecting depression [4] [5] [6] [7]. Speech stands out as a favourable option for incorporation into an automated system, given its inexpensive, remote, and non-invasive measurability. In fact, clinicians frequently rely on the verbal behaviour of a patient, noting reduced verbal activity, low arousal, altered prosody, and a 'lifeless' tone in speech as indicative signs of depression [10].

This thesis aims to investigate how speech can be utilized in a reliable manner for the detection of depression. While the primary focus centers on detecting depression through speech/acoustics, the thesis also aims to examine the acoustic traits that distinguish pathological speech associated with depression from regular speech.

Pathological speech in patients with *Parkinson's disease* (PD) was examined and analyzed. PD is a neurological condition that impacts basal ganglia functions, leading to the gradual reduction of dopamine-secreting neurons [11]. The diminished dopamine levels result in characteristic motor impairments such as tremors and hypokinetic dysarthria (articulation difficulties), while non-motor symptoms encompass complex behavioural alterations and depression. Pathological speech analysis can provide valuable insights into the presence of depression, even when patients may struggle to express their emotional state verbally. However, accurately characterizing changes and finding insights from speech signals can be challenging, as it typically requires separating the voice source and the vocal tract information accurately.

By exploiting different machine learning and deep learning methods, this research aims to:

- compare the performances of different classification methods for detecting depression;
- investigate the primary acoustic features for detecting depression from the speech of healthy and pathological patients and distinguish whether they are related to the source voice or the vocal tract system;
- understand how depression in Parkinson's patients manifests itself differently in comparison to healthy patients in terms of acoustic features.

1.1 State of the art

Depression represents a notable phenomenon which has gained increased attention in recent years, particularly in terms of its automated detection and severity assessment. To address this challenge, researchers have explored the automatic classification and severity prediction of depression using various modalities, including audio, video, and text, by extracting relevant parameters from sessions of patient clinical interviews [12] [13].

Numerous speech characteristics have been identified as potential indicators of depression: the influence of depression on human speech production extends to speech motor control, as documented in previous studies [14] [15], and is manifest through observable abnormalities in prosody, articulation, and phonetic accuracy; also, alterations in voice quality, encompassing features such as changes in glottal pulse shape, breathiness degree, jitter, and shimmer, have been reliably associated with depression [16] [17] [18]. Articulatory and phonetic errors have also been shown to be indicators of depression [19]. Since depression can sometimes display symptoms associated with negative emotions, some researchers have incorporated features inspired by speech emotion recognition studies [20]. However, it is crucial to recognize that the expression of negative emotions significantly differs from the experience of clinical depression. Several studies have employed statistical analyses of features known as low-level descriptors (LLD) (see Sec. 2.3.2) associated with both the vocal source and vocal tract to enhance existing systems [20] [21]. However, not every statistical property contributes to the improvements.

Despite the progress made in this field, there is still no agreement on a set of features that can be used to identify depression from speech signals reliably. Furthermore, these systems' performance could be limited by the features chosen and their statistical properties. In more recent developments, a shift toward deep learning methods has been observed. A notable example is utilizing neural networks incorporating convolutional and long short-term memory layers. These networks were employed to predict depression, using features such as log Mel filter-bank (LMFB) and magnitude-spectrogram data [22].

1.2 Outline

Chapter 2, *Background*, defines speech production and its related features. Also, it gives an overview of feature extraction and classification methods used in the study. Finally, it introduced the main knowledge of Convolutional Neural Networks in the context of speech.

Chapter 3, *Datasets, Protocols and Evaluation Metrics*, describes the dataset provided for the analysis and defines the experimental protocols and metrics used in the experiments.

Chapter 4, presents the Handcrafted features methodology and gives the relative results with comments.

Chapter 5, explains the end-to-end approach, details the used CNNs and describes results.

Chapter 6, reports comments and considerations on the results obtained.

4

Chapter 2

Background

2.1 Speech production

The speech signal is often represented in terms of a source-filter model and modelled as a two-stage process. The first process models the sound source originating at the glottis as a time-varying signal $e(t)$, as a periodic pulse train with a pulse spacing τ_p. The second stage works as a filter that amplifies and attenuates the signal with a continuous impulse response and a peak at a chosen resonance frequency, called formats. The filter represents the vocal tract system $\nu(t)$. The resulting speech signal $s(t)$ is obtained by the convolution of $e(t)$ and $\nu(t)$ in the time domain:

$$s(t) = e(t) * \nu(t).$$

Through Figure 2.1, it is possible to get an idea of the source-vocal filter model that has just been described. In the frequency domain, this involves multiplying the Fourier transform (FT) of the excitation signal and the FT of the vocal tract:

$$S(j\omega) = E(j\omega) \cdot V(j\omega).$$

The resulting waveform is also periodic with a period of τ_p, with a line spectrum with frequency of $1/\tau_p$ and an envelope determined by the vocal tract's frequency response [23].

The sound source (the glottis) The source of voiced speech sounds emanates from the vibration of the vocal folds, which are situated within the portion of the larynx referred to as the glottis. When air is forced to flow from the lungs through a closed glottis, the vocal folds enter a state of vibration. This vibratory motion serves as the primary sound source for most speech sounds.

It is worth noting, however, that not all speech sounds are generated by the glottal source wave. Voiceless speech sounds originating higher in the vocal tract

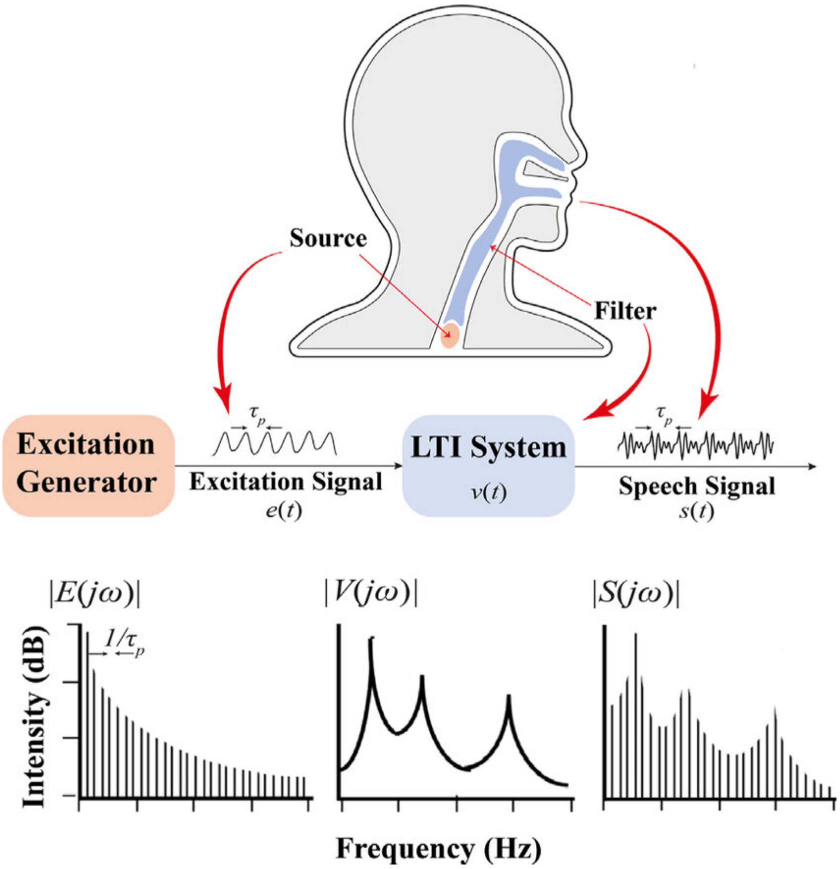


Figure 2.1: In the source-filter model of speech production, the glottis serves as the origin of the excitation signal, while the vocal tract, including the nasal and oral cavities, acts as the filtering element. The accompanying figure illustrates the temporal and spectral characteristics of the source, vocal tract, and the resulting speech signal.
Source: [24]

are instead produced by constriction within the vocal tract itself. For instance, in the case of the voiceless labiodental fricative [f], the sound source is air passage through the constriction between the lower lip and the upper teeth. The filter for this particular sound is relatively small since there is not much in front of these structures to alter the sound.

The filter (the vocal tract) The glottal source wave undergoes filtration within the vocal tract as it progresses towards the external environment. Numerous essential anatomical structures concerning speech production within the vocal tract come into play. These include the epiglottis, pharynx, velum, various tongue parts (tip, blade, body, and root), the alveolar ridge, hard palate, teeth, lips, and the nasal cavity. Each component serves as a potential filter for modifying the sound originating from the glottal source wave.

2.2 Prosodic and acoustic features

Speech features can be divided into four main groups: source, spectral, prosodic, and formant features.

Source related features: Source features convey information about the glottis during natural voice production. They can either parameterize this flow via glottal features or parameterize vocal fold movements via voice quality features. A limited body of research has delved into the impact of depression on source measures, with a predominant focus on voice quality attributes. Voice quality measures frequently employed in the analysis of speech affected by depression encompass jitter, which quantifies small cycle-to-cycle variations in glottal pulse timing during voicing; shimmer, which measures small cycle-to-cycle variations in glottal pulse amplitude in voiced segments; and harmonic-to-noise ratio (HNR), a ratio that gauges the presence of harmonics relative to inharmonic components. Depressed speech is often associated with breathy and tense voice qualities, indicating a decline in laryngeal coordination. In a study by Flint et al. (1993) [25], increased spirantization was observed in depressed individuals compared to healthy controls. Spirantization reflects aspirated leakage at the vocal folds and indicates disruptions in vocal fold behaviour.

Spectral features: Spectral features are utilized to characterize the speech spectrum, which represents the frequency distribution of the speech signal at a specific moment, typically in a high-dimensional representation. Among the commonly employed spectral features are the Power Spectral Density (PSD) and Mel Frequency Cepstral Coefficients (MFCCs). Spectral features are particularly effective in capturing a range of characteristics, including the decay of intensity, prosodic irregularities, and articulatory and phonetic errors associated with changes in speech motor control. Moreover, they offer detailed insights into vocal tract behaviour, potentially capturing information about muscle tension and control alterations.

However, it is worth noting that since these features encompass all the information present in speech, both linguistic and paralinguistic, this comprehensiveness may present challenges for the performance of classification or prediction systems relying solely on these features. In the literature, prominent spectral effects have been documented. It involves a relative shift in energy from lower to higher frequency bands or a decrease in energy variability within sub-bands. Notably, Tolkmitt et al. (1982), as documented in their study [26], were the first to observe a shift in spectral energy, specifically from frequencies below 500 Hz to the 500-1000 Hz range, in correlation with increasing severity of depression.

Prosodic features: Prosodic features represent the long-time (phoneme level) variations in perceived rhythm, stress, and speech intonation. Key examples include speaking rate, pitch (the auditory perception of tone), and loudness. In practical terms, fundamental frequency (F0, representing the rate of vocal fold vibration) and energy are the most commonly employed prosodic features, as they directly relate to the perceptual attributes of pitch and loudness. Hollien (1980) [27] has suggested that individuals experiencing depression exhibit distinct speech patterns, highlighting five potential characteristics: reduced speaking intensity, a narrower pitch range, slower speech rate, diminished intonation, and a lack of linguistic stress.

Curiously, even though clinical depictions of speech affected by depression often describe it as dull, monotonous, and lacking vitality (as documented by Hall et al., 1995), research has yielded divergent results concerning the influence of depression on fundamental frequency (F0) variables. It is, however, not unexpected that numerous studies have identified noteworthy correlations between a diminished F0 range and a reduced average F0 in tandem with escalating levels of depression severity. This contradiction in findings can be attributed to the diverse and varied nature of depression symptoms.

Formant features: Formants are the dominant components in the speech spectrum and contain significant amounts of information on the resonance properties of the vocal tract. Flint et al. (1993) [25] have identified significant disparities in the second formant location (F_2) associated with the phoneme /ai/ in individuals diagnosed with depression, in comparison to a control group that was carefully matched. They propose that this decrease in F_2 location may be attributed to a deceleration of tongue movement, specifically from a low-back to the high-front position. Formant-based characteristics are widely favoured in developing systems for classifying depressive speech. In their system architecture, Low et al. (2011) [28] observed that a combined set of the first three formant frequencies and their corresponding bandwidths exhibited noteworthy distinctions between individuals with depression and control subjects, with a statistical significance of $p<0.05$. Additionally, Helfer et al. (2013) [29] devised a binary classifier for distinguishing low and high depression cases using features derived from formant frequencies. These features included dynamic aspects, such as velocity and acceleration, resulting in reported classification accuracies of 70% with a Gaussian Mixture Model and 73% with a Support Vector Machine, respectively.

2.2.1 The impact of Parkinson's disease on speech

Parkinson's disease profoundly affects the acoustic features of speech. The progressive degeneration of the basal ganglia, resulting in dopamine insufficiency, limits the muscular control of the larynx, oral cavity, and other physiological support mechanisms for speech [30]. These limitations lead to speech abnormalities characterized by mono-pitch, mono loudness, altered speech rate, articulation difficulties, changes in nasality, voice quality issues, and disturbances in prosody. These acoustic abnormalities can substantially impact the ability of individuals with PD to communicate effectively.

2.2.2 The impact of depression on speech

Depression has been associated with a diverse range of alterations in speech, spanning prosodic, source, formant, and spectral attributes. These findings often vary across different feature sets, which is unsurprising given the complexity of speech production and the diversity of depression symptoms. Prosodic alterations, particularly decreased speech rate measures, provide evidence of slower articulatory muscle activity. Source-related changes suggest a decline in laryngeal coordination. Additionally, reductions in energy, both in the full frequency range and subbands, along with formant dynamics, point to increased articulatory effort and modifications in the vocal tract's resonance properties.

2.3 Short-time feature representations

The accurate extraction and interpretation of these acoustic features and classification of depression remain challenging, especially in speech difficulties such as those with Parkinson's disease. This section describes one of the traditional approaches to feature extraction from an audio signal.

2.3.1 Features extraction

In the classification process, the algorithm's primary objective is to identify specific characteristics or attributes that differentiate different classes, whether applied to image or waveform data. This is precisely what the feature extraction phase aims to achieve. It involves extracting relevant and informative characteristics from raw speech signals, which can be used for analysis, modelling, or classification tasks. The goal is to transform the raw audio data into numerical features that capture essential aspects of the speech signal.

Traditionally, speech assessment has conventionally relied on utilising handcrafted features known as low-level descriptors (LLDs).

2.3.2 Low Level Descriptors

LLDs are a generic set of features. openSMILE [31] is a standard software capable of extracting Low-Level Descriptors (LLD) and applying various filters, functionals, and transformations to these. It provides the eGeMAPS and ComPARE handcrafted feature set representations, used in the experimental studies.

eGeMAPS

The (extended) Geneva Minimalistic Acoustic Parameter Set [32] feature set was created to standardize affective computing research by generating the best collection of engineered features. It is mostly used in the field of speech analysis and emotion recognition [33].

The feature set includes 88 different descriptors in total. The features include the following Low-Level Descriptors:

- Frequency-related parameters (8)
- Energy/Amplitude related parameters (3)
- Spectral parameters (14)

These LLDs are extracted at every 10 ms within the speech. They encompass a range of short-term features related to the vocal source and vocal tract, as detailed in the provided Tab 2.1

Source-related	*System-related*
Loudness	Alpha ratio
F0 semitone from 27.5 Hz	Hammarberg index
Jitter	Spectral slopes (0-500, 500-1500)
Shimmer	Spectral flux
HNR (dB)	F1 (freq, bw, ampLogRelF0)
logRelF0-H1-H2	F2 (freq, ampLogRelF0)
logRelF0-H1-A3	F3 (freq, ampLogRelF0)
	MFCC (1-4)

Table 2.1: LLD features grouped by source and system (see [34] for detailed explanations)

The arithmetic mean and coefficient of variation (standard deviation normalized by the arithmetic mean) are applied as functionals to those 25 LLDs, yielding 50 parameters. Later, from loudness and fundamental frequency, eight functionals are applied and the mean of the slope of the rising and falling portion of the signal are computed. As a result, 70 parameters are generated. Finally, six temporal features are included: Rate of loudness peaks, mean length and standard deviation of voiced regions, mean length and standard deviation of unvoiced regions, and number of continuous voice regions per second. With other functionals on other parameters, we arrive at 88 features.

A detailed description of the LLDs provided by the openSMILE toolkit can be found in the Appendix.

ComPARE

The ComPARE [35] is the official baseline feature set introduced for the INTERSPEECH Computational Paralinguistics Challenge [36]. The feature set consists of 6373 different parameters. The different classes of used LLDs are:

- Energy-related LLDs (4)
- Spectral LLDs (54)
- Voicing related LLDs (6)

Many functionals are then applied to those parameters. In addition, the ComPARE set included 5 global temporal statistics based on voiced/unvoiced segments, computed using functionals applied to the fundamental frequency (F0) LLD. The statistics included the ratio of non-zero values (percentage of voiced frames out of the total frames) and segment length statistics (minimum, mean, maximum, and standard deviation of voiced segments, where $F0 > 0$. More information about ComPARE LLDs can be found in the Appendix.

2.3.3 Fixed-length feature representation

Since most features incorporate frame-level information, several mapping techniques can be employed for generating fixed-length utterances or speaker-level acoustic feature vectors.

Functionals Another approach to handle segments with variable lengths and eliminate the dependency of the feature vector dimensionality on the segment length is statistical functionals that can be applied to the time series of LLDs. A functional $\mathcal{F}$ maps a series of values $x(n)$ to a single value $X_{\mathcal{F}}$:

$$x(n) \rightarrow X_{\mathcal{F}} \tag{2.1}$$

Thus, the result is independent of the length of the input. Common functionals are the arithmetic mean, standard deviation, maximum, and minimum values. Typically, these functionals are applied to each LLD individually or to multiple descriptors simultaneously, such as the covariance or correlation between two descriptors.

Bag of Audio Words BoAW [37] is a sparse audio representation formed by quantising acoustic LLDs; each frame-level LLD vector is assigned to an audio word from a codebook learnt from some training data. Counting the number of assignments for each audio word, a fixed-length histogram (bag) representation of an audio chunk is generated. The histogram represents the frequency of each identified audio word in a given audio instance. For example, spectral envelope features associated with formants can result in phonemic audio representations, and their histogram representations can convey the occurrence frequencies of pronounced phonemes, pauses, and silences.

2.4 Classification methods

Classification is the procedure of identifying, comprehending, and organizing objects and concepts into predefined groups. Machine learning classification algorithms leverage training data to estimate and generate the probability or likelihood that incoming data will belong to one of the established categories or classes. In essence, a classifier is a model that, relying on input training information, assigns new observations to specified classes or clusters.

The process of selecting suitable classification methods posed a significant challenge. After thorough deliberation, the ultimate decision was made favouring a combination of fundamental techniques commonly used for classification tasks. These techniques include Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting (GB), in addition to the incorporation of Convolutional Neural Network (CNN).

2.4.1 Support Vector Machine

The aim of a Support Vector Machine (SVM) [38] is to detect the best hyper-plane in N-dimensional space (N the number of features) that properly classify the data. Of all the possible hyper-planes that could be chosen, the best is the one with the maximum margin (i.e. the distance between the closest data points from each class) and that minimises classification errors: in general the larger the margin, the lower the generalization error of the classifier.

SVMs can handle both linearly separable and non-linearly separable data by leveraging a technique known as the kernel trick. The kernel trick empowers the algorithm to covertly transform the input features into a higher-dimensional space, in which the data becomes linearly separable. Consequently, SVMs can effectively address intricate classification challenges that lack a linear decision boundary within the original feature space.

Support Vector Classification [39] is a classification method based on SVMs.

The following parameters are essential for the algorithm and will be fine-tuned to search for the best F1 score:

- Kernel: the purpose of this parameter is to accept the input data and convert it into the hyperplane defined by the mathematical function of the kernel. The kernels tested in our case are linear, polynomial, Radial Basis Function (RBF), and sigmoid.
- Regularization parameter (C): the 'smoothness' of the margins is controlled by C, which allows the SVM to tolerate a certain degree of classification error: a high value of C means the model is harder (less tolerant to misclassifications). Whereas a low value of C means that the model is softer (more tolerant to misclassifications).
- Gamma (γ): The kernel coefficient in question applies to the sigmoid, RBF, and polynomial kernels. It governs how much a single training point can influence the surrounding region. Lower gamma values (ranging from 0.008 to 0.01) signify a broader similarity radius, causing more points to be grouped together. In contrast, higher gamma values (ranging from 3.0 to 11.0) necessitate points to be in very close proximity to each other to be classified within the same category.

2.4.2 Random Forest

RF is based on decision trees and combines multiple trees to make predictions. To build each tree within the forest, a random subset of the original training data is chosen using bootstrap sampling, resulting in each tree being trained on a slightly varied dataset, thereby infusing diversity into the forest.

At each node of a decision tree, a random subset of features is considered to determine the best split. This approach ensures that each tree only assesses a subset of features, reducing the risk of any single feature dominating the decision-making process. The tree is constructed by iteratively dividing the data based on different features and thresholds, to minimize impurities in the resulting subsets.

Once all the trees are constructed, predictions are made by aggregating the outputs of individual trees through a voting mechanism. In classification tasks, the class that garners the most votes becomes the predicted class.

The following parameters are essential for the algorithm and will be fined tuned to search for the best F1 score:

- Number of estimators: represents the number of decision trees in the random forest. Typically, the greater the number of estimators, the better the performance of the random forest up to a certain threshold. Nonetheless, it's important to note that an excessive number of estimators can lead to computational complexity and prolonged training times.

- Maximum depth: A decision tree expands by separating data recursively based on characteristics and thresholds until a stopping criterion, which may be the maximum depth, is satisfied. The trees can recognise more intricate patterns in the data when the maximum depth is greater, although overfitting is also possible. To prevent overfitting and let the trees identify significant correlations in the data, it is critical to tweak this parameter properly.
- Minimum sample split: dictates the minimum number of samples needed further to split an internal node within a decision tree. When the number of samples at a node falls below the specified minimum, that node is designated as a leaf node, and any additional splitting is halted. This parameter serves the purpose of regulating the depth of the tree, ensuring that it does not excessively divide regions with inadequate data.
- Minimum samples leaf: establishes a minimum requirement for the number of samples that must be present in a leaf node. Should a split operation lead to a leaf node with fewer samples than the specified minimum, that split is abstained from. Like the minimum samples split parameter, this setting is essential for managing the size and depth of the decision tree and serves as a safeguard against overfitting.

Tuning these parameters is crucial to optimize the Random Forest model's performance and enhance its ability to generalize well on new, unseen data.

2.4.3 Gradient Boosting

GB leverages the power of decision trees to make predictions. In the construction of each tree within the ensemble, Gradient Boosting follows a sequential approach. It starts with a simple model, typically a shallow tree, and then builds additional trees, with each tree attempting to correct the errors of its predecessor.

At each stage of tree construction, Gradient Boosting assigns more weight to data points that were previously misclassified or had higher prediction errors. This adaptive weighting system ensures that the next trees focus on the challenging instances in the data. A group of trees is progressively put together, and each tree adds its own unique knowledge to the predictions. Gradient Boosting promotes variety across the different trees by considering a subset of characteristics at each node while deciding the optimum split, similar to Random Forest. These criteria were selected to reduce the possibility of any aspect predominating the decision-making process. Once all the trees are constructed and trained, predictions are made by combining the outputs of individual trees. In classification tasks, the class that garners the most weighted votes becomes the predicted class, leading to robust and accurate predictions.

The Gradient Boosting model has the same parameters listed in the previous paragraph and described for the Random Forest model (Number of estimators, Maximum depth, Minimum sample split, Minimum samples leaf) and includes a parameter called the "learning rate" that regulates the speed at which the model learns from errors.

2.4.4 Convolutional Neural Network

Over the past decade, notable demonstrations of the automatic acquisition of task-specific knowledge from raw waveforms have occurred. This is achieved by employing convolutional neural networks (CNN) rather than relying on hand-crafted features. This innovation has been particularly evident in the context of phoneme classification [40], speech recognition [41] [42], speaker recognition [43] and verification [44], gender recognition [45], emotion recognition [46].

CNNs are a type of neural network that uses the convolution layer to filter and decompose the input signal. These filters multiply local areas of the input data element-wise using small arrays of learnable parameters as their representation. After adding the findings, a single value known as a "convolutional feature" or "activation" is created. CNNs create feature maps that capture various data elements by swiping numerous filters across the entire input.

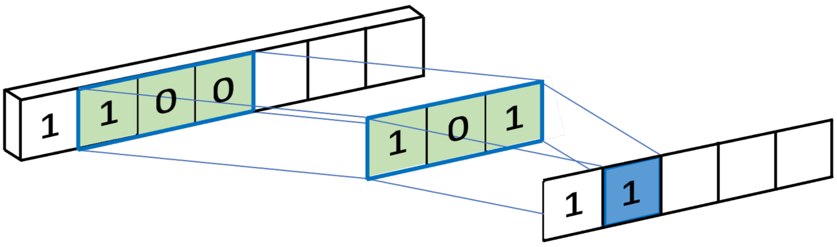


Figure 2.2: Example of a convolution operation

In a concise overview of CNN architecture, we observe that a non-linear activation function is applied element-wise after each convolution operation to introduce complexity into the network. Fig 2.2 represents the convolution operation with a 1D input, where the middle list is the filter. This process involves the multiplication of the filter with a specific part of the input, followed by summation, producing a singular value referred to as the convolutional feature or activation. By sliding the filters across the entire input, multiple convolutional features are generated, constructing feature maps that encapsulate various aspects of the input. Among the commonly used activation functions, Rectified Linear Unit (ReLU) stands out; another common activation function is the Hyperbolic tangent (Tanh). Additionally, a crucial component is the pooling layer, employed to reduce the spatial dimensions of feature maps while preserving essential information. Beyond these, we incorporate fully connected layers, akin to those found in traditional neural networks. These layers are responsible for capturing high-level representations and making predictions. Notably, each neuron in a fully connected layer connects with every neuron in the preceding layer, enabling the network to learn intricate relationships between features. Ultimately, the last fully connected layer yields the model's prediction.

2.5 Summary

This chapter explores the crucial steps in managing speech analysis, focusing on producing a voice sound, feature extraction and classification methods. Feature extraction is a vital process that transforms raw data into informative representations, enabling us to analyze the speech signals effectively. We also discuss several classification methods, including Support Vector Machines, Random Forests, Gradient Boostings, and deep learning. Each of these methods offers different approaches to categorizing and analyzing speech signals, each with its hyperparameters, strengths, and limitations.

Chapter 3

Datasets, protocols and evaluation metrics

In this chapter an overview of available datasets and their use is presented, followed by an in depth description of the dataset that the rest of the chapters focus on.

3.1 Depression in Parkinson's disease

The first dataset consists of speech data from 60 Spanish speakers from Colombia, including 25 Depressive PD patients (D-PD) and 35 Non-Depressive PD patients (ND-PD) [47]. The participants were asked to talk about their daily routines, and immediately after the task, they were evaluated by a neurologist. The transcripts of the speech data were obtained from audio recordings and were manually transcribed following a verbatim protocol. Additionally, the transcriptions were translated into English because the underlying language model was trained on English texts.

There is a total of 60 audio files for a total duration of about 4892 seconds. On average, the monologues of D-PD patients lasted for approximately 84±34 seconds, while those of ND-PD patients lasted for approximately 80±37 seconds.
For the study, silences between phrases were taken in since they could contain important information useful to the task.
From the audio file, some statistics can be extracted, presented in Tab 3.1, and Fig 3.1 shows the distribution of the duration grouping by class. The maximum duration of close to 172 s concerns only one audio belonging to the Not Depressive class, accompanied by a few speech audios which exceed 100 s. However, as depicted in 3.1, a significant proportion of audios from both classes fall within the range of 50 s and 120 ms. This observation is further supported by the median of 82.12.

Class	Mean duration	Max duration (s)	Min duration (s)	Median
Overall	81.5	171.85	23.37	82.12
Depressive	83.8	150.3	38.44	82.12
Not Depressive	81.12	171.85	23.37	84.24

Table 3.1: Statistics on PD-D

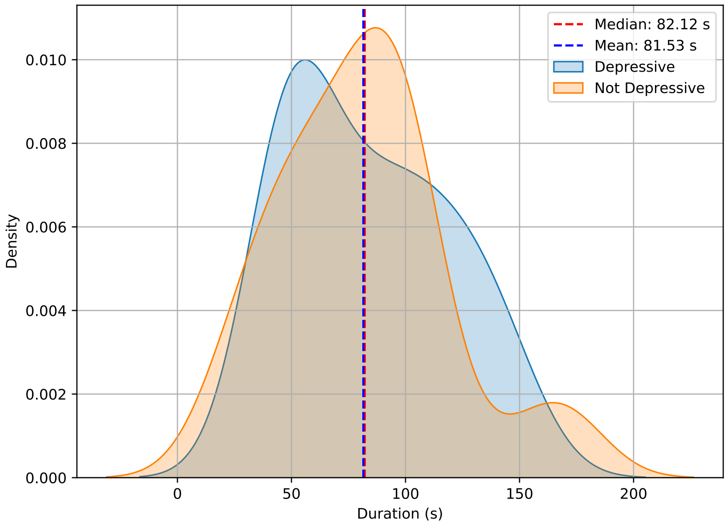


Figure 3.1: Distribution of the duration

3.2 The Distress analysis interview corpus - Wizard of Oz

The Distress analysis interview corpus Wizard of Oz (DAIC-WOZ) [48] database comprises audio-visual interviews of 189 participants, male and female, who underwent evaluation of psychological distress such as anxiety, depression, and post traumatic stress disorder. An animated virtual interviewer conducts the interviews called Ellie controlled by a human interviewer in another room, and the participants include both distressed and non-distressed individuals. Each participant was assigned a self-assessed depression score through the patient health questionnaire (PHQ-8) method [49].

This dataset consists of a total of 17 hours of audio data. The distributions shown in Fig 3.2, and the statistics extracted and reported in Tab 3.2 refer to the conversation pieces in which the patient speaks, then combined into a single audio file, without inserting silences.
It is evident that this dataset has considerably longer audio files than the previous dataset presenting a sample with a maximum length of 1215 s.
The durations of the speech audio file are mostly focused around the median and the mean, between 100 s and 700 s. For the classification part, the audio are segmented into small chunks of an average length of 2.7 s. In this way, the dataset is more robust, and training is performed on more audio segments.

Class	Mean duration (s)	Max duration (s)	Min duration (s)	Median (s)
Overall	450.36	1214.9	62.23	421.58
Depressive	478.71	1214.9	127.97	458.72
Not Depressive	438.45	1174.93	62.23	413.19

Table 3.2: Statistics on DAIC-WOZ

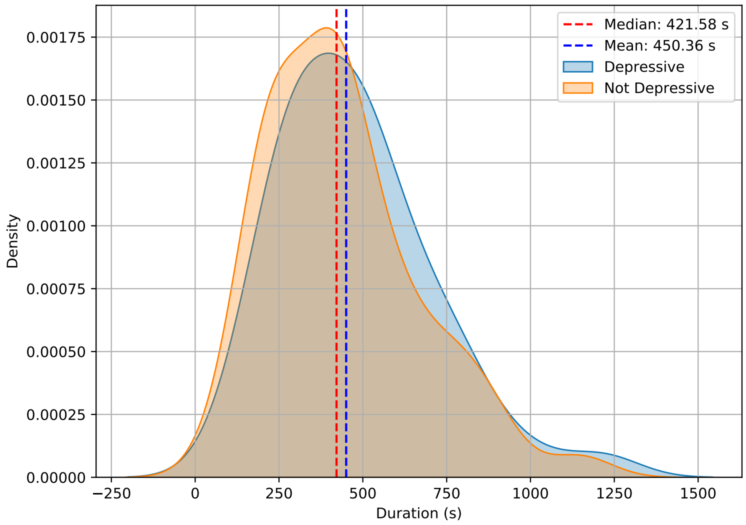


Figure 3.2: Distribution of duration

3.3 Experimental setup

This section aims to delve into the details of our experimental configurations.

3.3.1 PD protocol

To address potential challenges when dealing with a small and relatively imbalanced dataset, such as PD, a traditional train/validation/test split for model training and testing may not suffice. In such cases, it is essential to devise an appropriate approach for the train-test split. A highly effective solution to this issue is the implementation of k-fold cross-validation.

The procedure involves a single parameter called k, which refers to the number of groups a given data sample is split into. Then it is used to apply this procedure on the limited sample to assess how the model is expected to perform in general when we used to make predictions during the model's training. Every observation within the data sample is allocated to a specific group and remains within that group throughout the procedure. This ensures that each sample can be included in the holdout set once and utilized for training the model $k-1$ times.

This methodology entails randomly partitioning the observation set into k groups, or folds, of roughly equal sizes. The initial fold is treated as the validation set, while the method is fitted using the remaining $k-1$ folds.

On the PD-D dataset, cross-validation was applied by assigning k the value of 1, i.e. the Leave One Out (LOO) protocol.

In speech context, speakers often represent examples. In automatic speech recognition (ASR) tasks, a speaker refers to an individual who produces speech or vocal utterances. Leave One Speaker Out (LOSO), illustrated in Figure 3.3, is executed iteratively, with each iteration involving the following steps:

1. One session is selected as the test speaker for the current iteration;

2. The remaining session's data (excluding the test speaker) is combined to create the training set;

3. A speech processing model (e.g., classifier, recognizer) is trained using the training set;

4. The trained model is tested on the data from the test speaker to assess its performance.

The final performance estimate is the average of all the individual scores. For training purposes, speaker 52 was excluded due to recording errors.

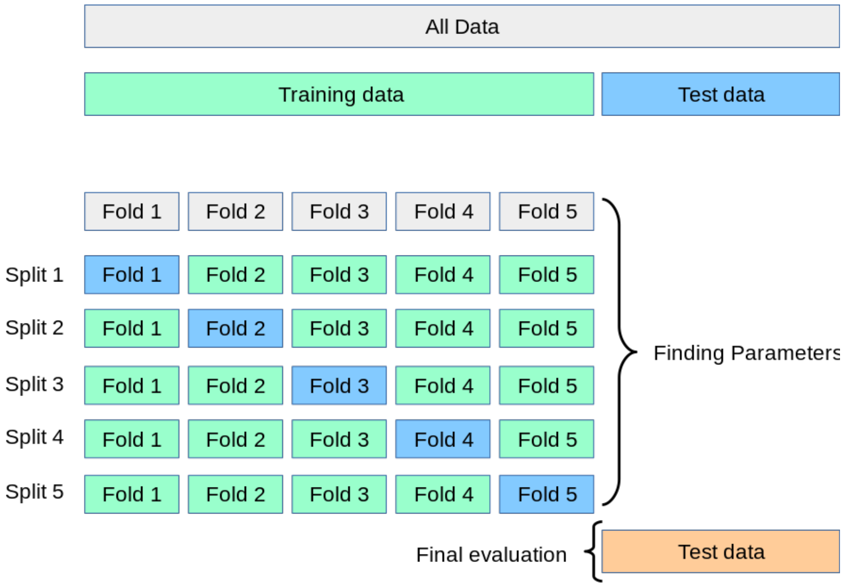


Figure 3.3: Overview of Leave One Speaker Out protocol[1]

3.3.2 DAIC protocol

The detection of depression was carried out exclusively using the speech modality from the DAIC-WOZ corpus, framing it as a binary classification problem at the speaker level. The time labels available in the dataset were utilized to isolate the speech recordings of the participants for experimentation. For training purposes, sessions 318, 321, 341, and 362 were excluded due to time-labeling errors. The techniques were evaluated on the dev set, as the test set had been reserved as a component of the AVEC 2016 challenge [50].

Data augmentation

Due to the naturally larger proportion of non-depressed people relative to those who are depressed, unbalanced datasets are widespread in many real-world contexts, including identifying depression. When creating and assessing machine learning models, this intrinsic class imbalance might cause problems. When confronted with imbalanced data, machine learning models tend to favour the majority class (non-depressed) during training.

It is noticeable that the proportion of depressive and not depressive patients in DAIC-WOZ is imbalanced: only 42 individuals were identified as having depression.

[1] https://scikit-learn.org/stable/modules/cross_validation.html

Thus, a data augmentation technique has been adopted. After splitting the data according to the protocol, only on training data, the following two-step approach has been applied:

1. By examining the distribution of labels in the set the algorithm identifies which category between depressed, labelled as 1, and not depressed, labelled as 0, needs additional samples for balancing and the necessary number of repetitions;

2. The script addresses class imbalance by creating additional instances of underrepresented data.

3.4 Performance metrics

The accuracy counts the number of times a model is correctly predicted over the full dataset. This measure can only be relied on if the dataset is class-balanced, or if each class in the dataset contains an equal number of samples. Thus, the F1 score, precision (P), and recall (R) have been used to assess the binary classification tasks of depression detection. To that end, frame-level results like true positive (TP), false positive (FP), false negative (FN), and false positive (FP) can be exploited to calculate the classification metrics. Precision measures the proportion of predicted positive samples that are actually true positives.

Recall is calculated as the ratio of true positives (TP) to the sum of true positives and false negatives (FN).

The harmonic mean of recall and precision is the F1 score.

They are computed as:

$$\mathrm{P} = \frac{TP}{TP + FP}$$

$$\mathrm{R} = \frac{TP}{TP + FN}$$

$$\text{F1 Score} = \frac{2}{\frac{1}{\mathrm{P}} + \frac{1}{\mathrm{R}}} = \frac{2 \times \mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}}$$

3.5 Summary

This chapter introduces two distinct datasets used in the studies. The first dataset, Depression in Parkinson’s Disease (PD-D), includes 60 speakers, of whom only 24 were validated as depressed. The second dataset, the Distress analysis interview corpus Wizard of Oz (DAIC-WOZ), contains 189 male and female patients. There is an imbalance between the classes, with non-depressed predominating. It is important to note that there is an imbalance between the classes, with non-depressed individuals being predominant. To mitigate any issues during training, we resorted to data augmentation. We also present the evaluation metrics used to assess the performance of the models.

Chapter 4

Handcrafted feature study

The traditional approach to speech processing tasks, like emotion recognition or depression detection, involves short-time feature extraction. These handcrafted features are extracted from speech by a feature extractor. They are aggregated to obtain fixed-length representations at the utterance or speaker level and finally used as the input of a classifier.

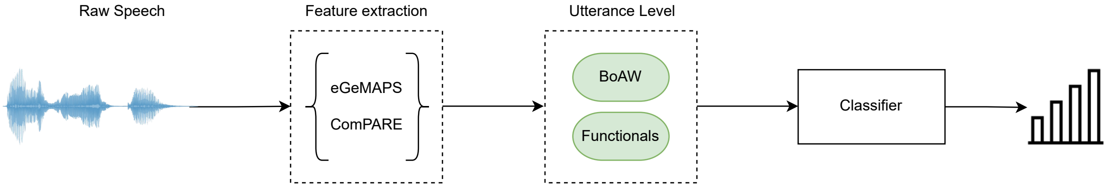


Figure 4.1: Handcrafted features pipeline

4.1 Experimental strategy

The image in Figure 4.1 depicts the main steps of the pipeline that have been used in the study.

First, as already used several times in literature for other research, eGeMAPS and ComPARE feature sets were extracted, analysed and compared. The features are then normalised with a scaler[1] that scales and translates each feature individually so that it lies in the range given on the training set; in this study between zero

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html`

and one. According to its classification algorithm, the classifier is fed with the normalized features and gets the most probable belonging class as output. This study applied Support Vector Machine, Random Forest, and Gradient Boosting algorithms to determine the optimal classifier and hyperparameters for the specific features and dataset used in the analysis.

At the last stage, since the purpose of this study is to understand the most relevant features to identify depression, a qualitative analysis of the most important features has been done.

From the models that best fit the distribution of the features extracted from each dataset, the names of the top 10 descriptors were printed out and analysed qualitatively, sorted according to their *feature importance score*.

Lastly, a qualitative understanding enabled us to arrive at some important conclusions for the task and to be able to comment on some evidence concerning how depression in healthy patients presents differently from depression in Parkinson's patients.

4.2 Hyper-parameter tuning

To find the most efficient set of hyperparameter values for a specific model, grid search[2] is a well-known hyperparameter optimisation technique. The process involves selecting the best set of hyperparameters before the training phase, which significantly impacts the model's performance.

Each combination of values in the grid is used to train and evaluate the model using a predefined evaluation metric described in Section 3.4. This systematic evaluation process helps to assess the model's performance across all possible hyperparameter combinations.

The ultimate objective of grid search is to find the most efficient set of hyperparameters that produces the best performance on the evaluation measure. This combination is then selected as the most suitable set to evaluate the algorithm on a test set taken out before the k-fold, according to the protocol splitting for each dataset.

4.3 Hypotesis and Results

As we have previously stated, selecting the right set of biomarkers to accurately indicate depression is crucial. Biomarkers that reflect changes in an individual's

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html`

Model	parameters	Grid search values
SVM	C γ Kernel	[0.1, 1, 10, 100] [0.001, 0.01, 0.1, 1] Linear, RBF, polynomial, sigmoid
RF	Number of estimators Maximum depth Minimum samples split Minimum samples leaf	[10, 20, 30, 40, 50, 70, 80, 100, 150, 200] [5, 7, 10, 20] [2, 3, 5, 7, 10, 15] [3, 4, 5]
GB	Number of estimators Maximum depth Minimum samples split Minimum samples leaf Learning rate	[10, 20, 30, 40, 50, 70, 80, 100, 150, 200] [5, 7, 10, 20] [2, 3, 5, 7, 10, 15] [3, 4, 5] [0.001, 0.01, 0.1]

Table 4.1: Grid search parameters and value for SVM, RF and GB

emotional and mental state are becoming increasingly important. Although not consistently reported in all studies, certain vocal source-related features such as shimmer and jitter of vocal-fold vibration, degree of aspiration, dynamics of the fundamental frequency, and frequency dependence of variability and velocity of energy have been shown to have a statistically significant association with the presence of depression [51] [52].

Considering these factors, the goal of ongoing research is to demonstrate through analytical findings that there is a substantial correlation between vocal source characteristics and depression. It is these characteristics that are most important in distinguishing between depressed patients and those who are not.

The following section describes the results of experiments on PD-D and DAIC datasets applying the traditional pipeline. For each dataset, first, the performance will be discussed, and the different classifiers for each feature set will be compared. Then, for the best-performing models, a qualitative analysis of the most representative features for depression detection will be presented.

4.3.1 PD: classification results

Table 4.2 presents the results of the classification of the three chosen rankings. The first column indicates the model used and the handcrafted features extracted from the PD-D dataset. The second column lists the optimal parameters from the optimisation phase that best fit the data distribution. Finally, the last three columns present the F1 score, precision and recall values grouped for each class: depressed (D) and non-depressed (ND). The unweighted average between the F1 score values for the two labels was the reference for deciding the best classification model. Finally, the value in bold indicates the best result discriminating between depressed and non-depressed in the Depression in Parkinson's disease dataset.

On comparing the models' performance, this study found that the SVM classifier is the model that distinguishes optimally between depressed and not-depressed in the Depression-PD database for both $eGeMAPS$ and $ComPARE$ feature spaces. For the $ComPARE$ features set, we obtained the best classification result for PD-D with an overall value of 0.64 for the F1 score. Also, SVM$_{ComPARE}$ presents the highest percentage of Recall for D class: over out of all the instances that truly belong to *Depressed*, the model correctly identifies 64% of them as Depressed class. Below 0.5 score overall (0.59), but still the best for $eGeMAPS$ representation, the SVM model ranks the depressed patients with an F1-score of 0.54 and a recall of 0.56. The two top models, however, had different hyperparameters: SVM$_{ComPARE}$ exploited a linear kernel, while SVM$_{eGeMAPS}$ relied on a polynomial kernel by transforming the input features into a higher dimensional space in which the features become linearly separable. The Random Forest classifier is the second-best with an overall F1 score of 0.56 and 0.45 for $eGeMAPS$ and $ComPARE$ respectively, and in this occurrence, the optimal parameters for depression classification are the same for both feature sets. Finally, the Gradient boosting system is the one with the lowest values.

4.3.2 DAIC: classification results

Table 4.3 summarizes the results obtained by applying the handcrafted features extraction method on the DAIC-WOZ dataset. The meaning of each column in the table is the same as described in the previous section 4.3.1; this time the labels refer to *depressed* (D) and *control* (C) individuals. The label O indicates the unweighted average between the F1-score values for the two classes.

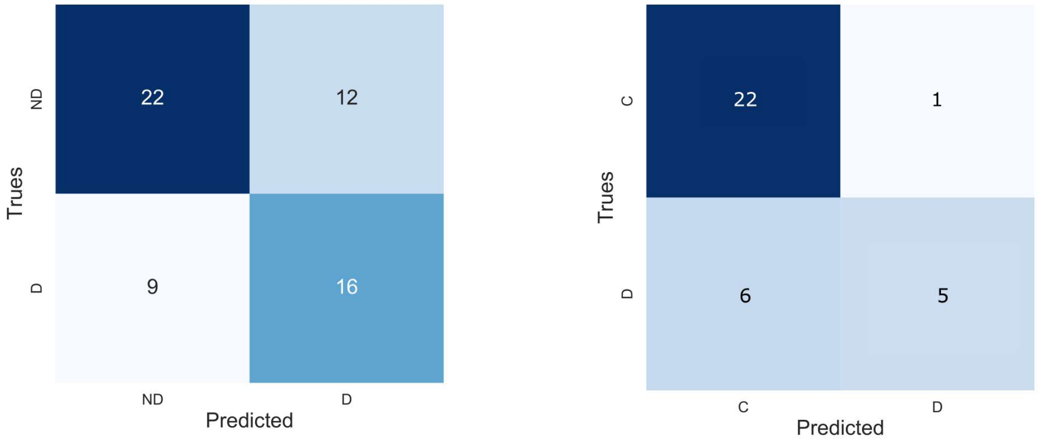
If for PD-D we had found SVM to be the best classification algorithm whatever the features selected, different results were obtained in the case of $eGeMAPS$ and $ComPARE$ in DAIC. It stands out that no model among SVM, RF and GB produces F1 scores below 0.5 (0.48, 0.39, 0.47 respectively) for $ComPARE$ features. This could indicate that the models are making many errors both in correctly classifying true positives (low recall) and in preventing false positives (low precision). The opposite is achieved by classifying the depression using $eGeMAPS$ features. The three models discriminate between *depressed* and *control* individuals with good results. The best F1 scores are obtained using algorithms based on decision trees, then RF and GB. The latter is the best-performing model with an overall F1 score value of 0.74, a precision of 0.9 and a recall of 0.47 for class D. This outcome is definitely the best classification result obtained in this study for the handcrafted approach.

Model	*Parameters*	*F1-score*			*Precision*		*Recall*	
		O	D	ND	D	ND	D	ND
		PD-D						
SVM$_{eGeMAPS}$	k: poly γ: 1 C: 10	0.59	0.54	0.64	0.52	0.66	0.56	0.62
RF$_{eGeMAPS}$	max_feat: log2 min_sample_leaf: 5 min_samples_split: 15 max_depth: 20 n_estimators: 10	0.56	0.41	0.71	0.57	0.62	0.32	0.82
GB$_{eGeMAPS}$	max_feat: log2 min_sample_leaf: 1 min_samples_split: 10 max_depth: 20 n_estimators: 100 loss: log_loss learning_rate : 0.1	0.54	0.4	0.69	0.53	0.61	0.32	0.79
SVM$_{ComPARE}$	k: linear degree: 1 C: 1	**0.64**	0.6	0.68	0.57	0.72	0.64	0.65
RF$_{ComPARE}$	max_feat: log2 min_sample_leaf: 5 min_samples_split: 15 max_depth: 20 n_estimators: 10	0.45	0.34	0.56	0.44	0.58	0.28	0.74
GB$_{ComPARE}$	max_feat: log2 min_sample_leaf: 1 min_samples_split: 10 max_depth: 20 n_estimators: 100 loss: log_loss learning_rate: 0.1	0.43	0.3	0.56	0.33	0.53	0.28	0.59

Table 4.2: Performance of the different classifiers on the Depression Parkinson's Diseases data. D points out *depressed*, ND indicates *not depressed*, and O denotes the *overall* score by un-weighted average over the two classes. Out of all the suggested techniques, the system with the highest total F1 score is indicated in bold type.

Model	*Parameters*	*F1-score*			*Precision*		*Recall*	
		O	D	C	D	C	D	C
DAIC-WOZ								
SVM$_{eGeMAPS}$	k: rbf γ: 1 C: 10	0.67	0.52	0.83	0.73	0.75	0.41	0.92
RF$_{eGeMAPS}$	max_feat: log2 min_sample_leaf: 5 min_samples_split: 15 max_depth: 20 n_estimators: 20	0.69	0.54	0.85	0.86	0.76	0.4	0.97
GB$_{eGeMAPS}$	max_feat: log2 min_sample_leaf: 1 min_samples_split: 10 max_depth: 20 n_estimators: 100 loss: log_loss learning_rate : 1	**0.74**	0.62	0.87	0.9	0.78	0.47	0.97
SVM$_{ComPARE}$	k: poly degree: 1 C: 10	0.48	0.3	0.66	0.41	0.58	0.18	0.8
RF$_{ComPARE}$	max_feat: log2 min_sample_leaf: 7 min_samples_split: 5 max_depth: 7 n_estimators: 20	0.39	0.07	0.72	0.35	0.58	0.04	0.95
GB$_{ComPARE}$	max_feat: log2 min_sample_leaf: 1 min_samples_split: 10 max_depth: 10 n_estimators: 60 loss: log_loss learning_rate: 0.1	0.47	0.24	0.7	0.45	0.59	0.16	0.86

Table 4.3: Performance of the different classifiers on the DAIC-WOZ dataset. D points out *depressed*, C indicates *control*, and O denotes the *overall* score by unweighted average over the two classes. Out of all the suggested techniques, the system with the highest total F1 score is indicated in bold type.



(a) Confusion Matrix on the PD-D set, extracting *ComPARE* feature set using SVM classifier

(b) Confusion Matrix on the DAIC-WOZ dev set, extracting *eGeMAPS* feature set using Gradient Boosting classifier

Figure 4.2: Confusion matrices for depression prediction on PD-D (a) and DAIC-WOZ (b) dataset using their respectively best classifiers

4.3.3 PD: feature importance analysis

Table 4.4 displays the ten most significant features of the *ComPARE* feature set in ascending order of feature importance score. The first column shows the original name of the descriptor. The second, third, and fourth columns refer to the origin of the descriptors, indicating whether they relate to the vocal source (source-related), the vocal tract system (system-related), or convey global and general information. The last column shows the relative feature importance score, which is calculated using the *coef_* attribute provided by the SVM library[3].

Before assigning a feature to either the system or source category, we conducted a qualitative investigation of the name's semantic significance. We examined and interpreted different parts of the label based on the string to assign the correct value. To help understand the feature names, we provide some examples by separating the first entry's name in the table. This will better prepare us for the data analysis.

`"pcm_fftMag_spectralVariance_sma_de_peakMeanRel"`

- `pcm_fftMag`: This refers to the magnitude spectrum obtained from the Fast

[3] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

Feature Name	Source-related	System-related	General	Importance score
PD-D, ComPARE, SVM				
pcm_fftMag_spectralVariance_sma_de_peakMeanRel			X	0.056435
pcm_fftMag_spectralCentroid_sma_de_peakMeanRel		X		0.049565
pcm_fftMag_psySharpness_sma_de_peakMeanRel			X	0.049564
pcm_RMSenergy_sma_minPos			X	0.041844
pcm_fftMag_spectralRollOff90.0_sma_de_peakMeanRel			X	0.041256
mfcc_sma[14]_minPos		X		0.040951
logHNR_sma_de_upleveltime25	X			0.039373
pcm_fftMag_spectralEntropy_sma_de_peakMeanRel			X	0.039115
pcm_fftMag_spectralEntropy_sma_de_minPos			X	0.036591
audspec_lengthL1norm_sma_de_upleveltime50		X		0.035357
DAIC-WOZ, eGeMAPS, GB				
F0semitoneFrom27.5Hz_sma3nz_percentile50.0	X			0.024646
loudness_sma3_amean	X			0.023184
F3frequency_sma3nz_amean		X		0.020240
F0semitoneFrom27.5Hz_sma3nz_percentile20.0	X			0.017258
loudness_sma3_percentile50.0	X			0.017234
equivalentSoundLevel_dB			X	0.017009
loudness_sma3_pctlrange0-2	X			0.016828
F0semitoneFrom27.5Hz_sma3nz_percentile80.0	X			0.016502
F0semitoneFrom27.5Hz_sma3nz_amean	X			0.016470
hammarbergIndexV_sma3nz_amean		X		0.016416

Table 4.4: The table shows the names and relative scores of the top 10 most significant features for depression classification. The upper part refers to the experiment conducted on the PD dataset, and the lower part refers to the results obtained using DAIC-woz samples. The middle columns specify when the descriptors carry source-related, vocal system-related or general information.

Fourier Transform (FFT) of the audio signal, which represents the distribution of signal energy across different frequency components.

- `spectralVariance`: identifies the variance or spread of power across these frequency components in the spectrum. It describes how the energy is distributed among various frequencies and indicates the variability of energy within different frequency bands.

- `sma_de`: This part implies the application of some form of differential analysis, such as differential computation or features computed using a Simple Moving Average.

- `peakMeanRel`: This indicates a calculation related to the relative mean of peaks in the spectrum. It likely refers to the relationship between the average intensity of the peaks found in the frequency domain.

`mfcc_sma[14]_minPos`

- `mfcc`: Stands for Mel-frequency cepstral coefficients, which are coefficients derived from the Fourier transform of a signal. They are used to represent the short-term power spectrum of an audio signal.

- *sma[14]*: Refers to the computation or extraction of this feature using a specific function or process. The '[14]' represents the 14th coefficient within the MFCCs;
- `minPos`: Indicates the position or time index where the minimum value of the 14th MFCC occurs within the analyzed segment of the speech signal.

Against our hypothesis formulated in Section 4.3, the dataset of depressed Parkinson's patients presents more features carrying information related to the vocal tract or general measure of the spectral characteristics. The spectral variance, spectral entropy, spectral Rolloff, Root Mean Square (RMS), and psychoacoustic sharpness all describe a signal's overall energy distribution. For example: the spectral variance measures the signal's spectral content variability over time [53]; The spectral entropy measures the randomness of a signal's spectral content [54]; the spectral roll-off is the frequency below which a specified percentage of the total spectral energy lies. This measure distinguishes voiced from unvoiced speech-unvoiced speech has a high proponion of energy contained in the high-frequency range of the spec", where most of the energy for unvoiced speech and music is contained in lower bands [55] [56]. These features are all influenced by both the vocal source (the vocal cords) and the vocal tract system (the airways, nasal cavity, and larynx) since they shape the overall spectral characteristics of the sound produced.

The spectral centroid represents the centre of gravity of the signal's spectral content, and it is more directly related to the overall characteristics shaped by the vocal tract [57] [58] [59].

With an importance score of roughly 0.040, the only feature related to the voice source that matters is `logHNR_sma_de_upleveltime25`. This feature logs the Harmonics-to-Noise Ratio (HNR) and highlights the clarity and periodicity of voiced sounds [60].

However, we based our idea on studies of depression in control patients without any specific pathology, rather than taking into account the influence of a disorder on our speech analysis. This oversight led us to make subsequent considerations and analyses of the impact of the disorder on depression detection. Therefore, more research on Parkinson's effects on speech is needed to explain the outcome (refer to Section 4.4 for more information).

4.3.4 DAIC: feature importance analysis

Similarly, the ten most important features of the *eGeMAPS* set are displayed at the bottom of Table 4.4, arranged in ascending order by feature importance score.

The descriptor's original name is displayed in the first column, where it is possible to distinguish between several components: the root refers to the LLD, and the desinences specify the functionals applied to the related descriptor. More details on LLDs and functionals can be found in Section 2.3.1. The GB model library[4] supplied the *feature_importances_* property, which was exploited to derive the relative feature importance score, which is shown in the last column.

Before analyzing the results, let us give examples of how to interpret the feature names by breaking down each part of the name of the first entry in Table 4.4 bottom part.

`"F0semitoneFrom27.5Hz_sma3nz_percentile50.0"`

- `F0semitoneFrom27.5Hz`: this component refers to the fundamental frequency (F0) measured in semitones from a reference frequency of 27.5Hz. F0 represents the perceived pitch of the speech signal.
- `sma3nz`: "sma" stands for "simple moving average," indicating that smoothing has been applied to the F0 values. The "3nz" further specifies the smoothing technique, which could involve a specific window size and other parameters. The smoothing process helps reduce noise and variations in the F0 values, making them more stable and suitable for analysis.
- `percentile50.0`: this represents the 50th percentile of the F0 values. The 50th percentile, also known as the median, is the value below which 50% of the F0 measurements fall and above which the other 50% fall. It provides a measure of the central tendency of the F0 values after applying the specified smoothing technique.

`"loudness_sma3_amean"`

- `loudness`: this LLD refers to the perceived intensity of the audio signal. It is a psychoacoustic measure that aims to quantify how the human ear perceives sound intensity.
- `sma3`: almost as before.
- `amean`: This component specifies that the feature calculates the arithmetic mean or average value of the loudness values processed using the "sma3" algorithm.

It is immediately noticeable that the majority of features are related to the vocal source, suggesting that it is this group that dominates and thus is representative

[4] `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html#sklearn.ensemble.GradientBoostingClassifier.feature_importances_`

of depression. Despite this, there are also two features related to the vocal tract among the top 10: `"F3frequency_sma3nz_amean"` with a score of 0.02024 in the top three highest, and `"hammarbergIndexV_sma3nz_amean"` in the last position with an importance score of 0.016416. Among those related to the source, several descriptors are associated with the fundamental frequency F0 (first, fourth, octave and ninth place) and the loudness (second and fifth place).

It is interesting to remark that for DAIC-WOZ, the three percentile levels are present (50^{th}, 80^{th}, 20^{th}). In the depression classification task, those features suggest that pitch-related characteristics at different percentile levels may be relevant in capturing differences between depressed and non-depressed individuals. This could indicate a complex relationship between pitch characteristics and depression, potentially highlighting various aspects of pitch distribution.

4.4 Mutual Information: Parkinson's *vs* Depression

Feature analysis presented in Section 4.3.3 from the classification of Depressive Parkinson's patients reveals vocal tract-related features as significant in depression, thus an involvement of disease effects on speech. Indeed, it has been shown that one of the main consequences of Parkinson's disease is its motor symptoms, such as tremors and rigidity, including articulation difficulties (dysarthria) [52]. This disturbance of motor control of speech can affect one or all of the respiratory, phonatory, resonatory or articulatory components of the speech production mechanism. Therefore, these abnormalities are reflected when extracting features contributing to classification. In light of these considerations, features related to the vocal tract become significant and may interfere in the classification of depression in people with Parkinson's disease.

To justify the presence of vocal tract and global related features in the depression classification of individuals with Parkinson's disease, we exploit the Mutual Information method.

Mutual Information (MI)[5] is a measure that quantifies the mutual dependence between two random variables. In other words, it measures the amount of knowledge that can be gleaned from one random variable through observation of another random variable (called reference target). High mutual information suggests a strong relationship or dependency between the variables, while low mutual

[5] `https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_classif.html#id3`

information indicates little to no relationship[6]. For our purposes, an additional dataset comprising people with Parkinson's disease and healthy people was used: *PC-GITA corpus*.

PC-GITA corpus [61] The database includes speech recordings of 50 people with PD and 50 healthy controls, 25 men and 25 women in each group. All the participants are Colombian Spanish native speakers. The age of the men with PD ranges from 33 to 77 years old (mean 62.2 ± 11.2), and the age of the women with PD ranges from 44 to 75 years old (mean 60.1 ± 7.8). For the case of healthy controls, the age of the men ranges from 31 to 86 (mean 61.2 ± 11.3), and the age of the women ranges from 43 to 76 years old (mean 60.7 ± 7.7).

We have calculated the mutual information score between *ComPARE* features and labels for each PD-D and PC-GITA dataset. This helps us to evaluate the amount of information that each feature carries for a specific task. In simpler terms, we can determine how much information a feature contains for Parkinson's disease when compared to PD-D data or PC-GITA audio data. Our goal is to determine whether the features listed in Table 4.4 are more indicative of depression or Parkinson's disease by comparing their MI scores. For simplicity's sake, here we have reported the MI value referring only to the features' LLD without considering their functionals and deltas.

The table 4.5 displays the top 10 LLDs already presented in section 4.3.3, categorized by type: source-related, vocal tract system-related, or global-related. The sixth column of the table represents the mutual dependencies value of these features with the PD-D dataset, indicated as "MI-Depression". The seventh column shows their MI score with the PC-GITA dataset, indicated as "MI-Parkinson".

The grey highlighted rows contain feature features with higher mutual information value in the last column, thus more indicative of Parkinson's.

Upon careful analysis of the table 4.5, it becomes clear that the characteristics that strongly suggest the presence of Depression are not particularly prominent. Elevated MI-Parkinson values are associated with the vocal tract system or offer a broad overview. Upon a closer examination of the values, we realize that the difference between them is only a delta of plus or minus 0.06. However, within the scope of this examination, the feature 'pcm_fftMag_psySharpness' is the most significant indicator of Parkinson's. Meanwhile, 'LogHNR' remains the feature with the highest delta and is connected to the source. The lack of a strongly significant value to indicate depression among Parkinson's patients is understandable, as the analysis involves two datasets that have a majority of samples with overlapping

[6] `https://en.wikipedia.org/wiki/Mutual_information`

FI rank position	LLD name	Source	System	General	MI-Depression	MI-Parkinson
1	pcm_fftMag_spectralVariance			X	0.1069	0.1663
2	pcm_fftMag_spectralCentroid		X		0.0866	0.1893
3	pcm_fftMag_psySharpness			X	0.0953	0.2146
4	pcm_RMSenergy			X	0.0709	0.1308
5	pcm_fftMag_spectralRollOff90.0			X	0.1321	0.176
6	mfcc		X		0.2189	0.1803
7	logHNR	X			0.116	0.0777
8,9	pcm_fftMag_spectralEntropy			X	0.1291	0.1859
10	audspec_lengthL1norm		X		0.1182	0.1142

Table 4.5: The table shows the names and relative scores of the top 10 most significant features for depression classification on the PD-D (see Tab 4.4. The second column indicates the LLD name of each feature; the middle columns specify when the descriptors carry source-related, vocal system-related or general information. Lastly, the MI scores referred to PD-D and PC-Gita

Parkinson's patients without depression. Those with Parkinson's and depression, as well as healthy individuals, are in the minority. Moreover, speech production is strongly influenced by Parkinson's pathology, likely due to phonological difficulties experienced by patients.

Going forward, it may be beneficial to proceed in smaller steps, integrating one more piece of information each time and looking for distinctive and characterizing insights into depression in individuals with Parkinson's disease.

4.5 Summary

In Chapter 4 of our study, we discuss the traditional method of extracting features from signals and classifying them. We have used two feature sets, eGeMAPS and ComPARE, that are provided by the OpenSMILE toolkit. We have used different methods to classify depressed patients into PD-D and discriminate between depressed and control for DAIC. Our results show that SVM is the best method to classify depressed patients into PD-D, while Gradient Boosting is the best method to discriminate between depressed and control for DAIC. We have found that the descriptors that carry information related to the voice source of the signal are more indicative of discriminating depression in non-pathological patients. However, the most characteristic features of Parkinson's patients describe general speech information, and it's more controversial. After comparing the mutual information scores, we have found that there are no significant and characterizing values for our task, suggesting a complex correlation between the effects of Parkinson's on speech.

Chapter 5

End-to-End CNN study

The second approach involves directly feeding the waveform data into a neural network, utilizing an end-to-end CNN architecture. In an end-to-end approach, the model is trained to complete a task directly using the raw input data taken with a fixed length.

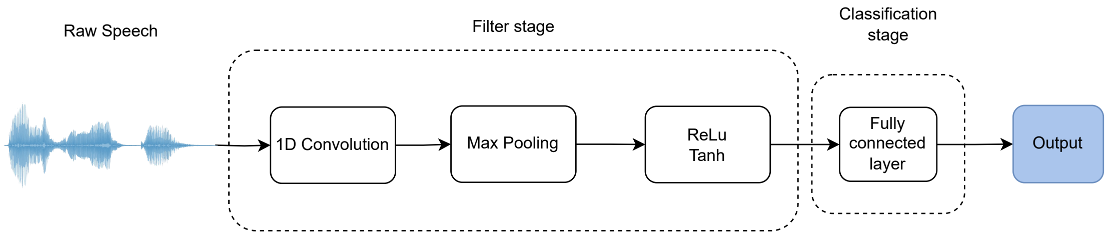


Figure 5.1: Overview of CNN architecture

As depicted in Fig 5.1, the model consists of a filter stage repeated several times, followed by a classification stage that comprises fully connected layers. Each convolution layer is composed of 3 operations: 1D convolution, max-pooling, and the activation function.

5.1 System architecture

We employed an architecture already adopted in several studies mentioned in section 2.4.4. Fig 5.2 shows how the row waveform is fed into the first convolutional layer. Let's consider a signal of length w_{seq}; the CNN takes as input a 250 ms (kW) fixed-length signal overlapped with a 10 ms (dW) window shift. n_f represents the number of filters in each layer.

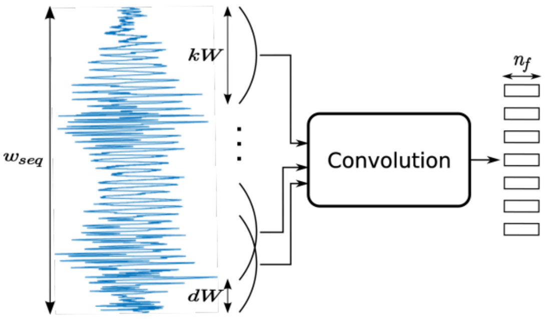


Figure 5.2: Illustration of the first convolution layer processing[1]

Thus, the first layer's output may be seen as a time-frequency representation like a spectrogram, except that the frequency axis has no particular order (unlike a standard spectrogram) and the channels can be connected based on the frequency responses of the filters. Figure 5.3 shows the structure of the network. Each layer's output is subjected to a nonlinearity, a rectified linear unit (ReLU), and, a max-pooling operation along the time axis. To obtain the probabilities of detecting depression, the output of the feature learner is fed to fully connected layers, with ReLU activations at the hidden layers and sigmoid activations at the output layer. The classifier component of the CNN is made up of a single hidden fully connected layer with 10 nodes. During training, the parameters are updated by backpropagating a cross-entropy loss calculated between the predictions and the targets. All the frames of the depressed group were labelled 1, and the rest 0. In this study, networks were trained using Tensorflow [62] [63].

Two CNNs can be distinguished based on the length of the kernel in the first convolution layer:

- *Subsegmental modelling* (*Subseg*): the kernel length is about 1.5 ms, considered less than a pitch period. This technique offers a reliable time resolution.
- *Segmental modelling* (*Seg*): the filter width is about 15 ms, values that allow to catch 1-5 pitch periods. It provides a better frequency resolution.

We examined both the subsegmental approach, which is useful for capturing locally present information related to glottal pulses, often requiring high time resolution, and the segmental approach to enhance the modelling of source-related information [44].

[1] https://infoscience.epfl.ch/record/270134?ln=fr

Table 5.1 summarizes and shows the architecture of the CNNs.

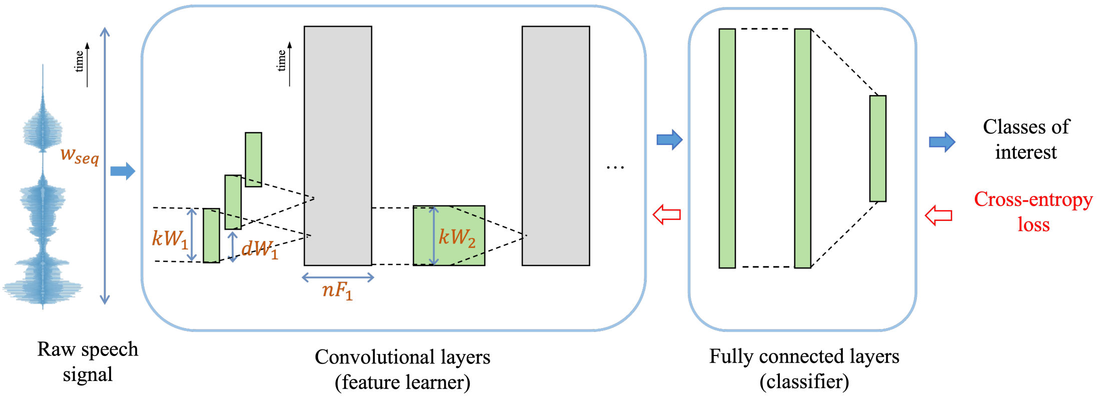


Figure 5.3: Automatic depression detection using raw speech CNNs

Model	Layer	$\mathrm{N_f}$	Conv kW	dW	MP
subseg	1	128	30	10	2
	2	256	10	5	3
	3	512	4	2	-
	4	512	3	1	-
seg	1	128	300	100	2
	2	256	5	2	-
	3,4	same as subseg			

Table 5.1: CNN architectures. N_f refers to the number of filters; kW indicates the kernel width; dW denotes the kernel shift; MP is for max-pooling

5.2 Modelling source and system based signals

In order to reproduce the feature classification analysis into source-related or vocal tract-related and to interpret the results, we utilized the method of filtering signals to enhance both the source-specific information and system information. We conducted three types of experiments to detect depression detection by feeding the network with three kinds of signals, which are as follows:

1. *Original raw signals*, Method 1 in Figure 5.4, contain both vocal source and vocal tract system information;

2. *Zero Frequency filtered* (ZFF) signals, Method 2 in Figure 5.4, carry source-related information;

3. *Composite signals* (CS), Method 3 in Figure 5.4, include system-related information.

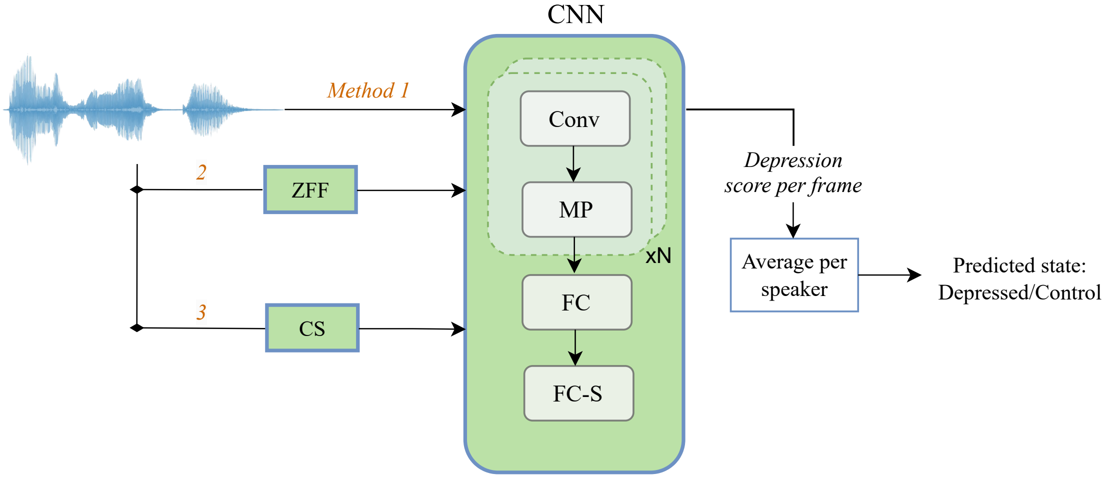

CNN
Method 1
2
ZFF
3
CS
Conv
MP
xN
FC
FC-S
Depression score per frame
Average per speaker
Predicted state: Depressed/Control

Figure 5.4: The proposed method

5.2.1 Original raw signals

In the first step, the original signals were used without applying any filter. Each one carries all the information relating to both the vocal source and the system. We used these raw signals to get reference results and then tried to interpret reasonably how the neural network learns according to different input types.

5.2.2 Zero Frequency Filtering

Zero frequency filtering (ZFF) is a technique that characterizes glottal source activity [64] [65]. It takes advantage of the characteristic of an impulse-like excitement at the glottal closure instance to detect glottal closure instants (GCIs). To obtain Zero Frequency Filtering (ZFF) signals, pre-emphasized speech signals are passed through a sequence of two ideal digital resonators positioned at 0Hz. Subsequently, any underlying trends in the resulting signals are removed by subtracting the average value within a window of a size ranging from 1 to 2 pitch periods. In addition to detecting GCIs, ZFF signals allow for the estimation of the strengths of glottal excitations, the fundamental frequency, and the instants of glottal opening.

Recent research has also demonstrated that Convolutional Neural Networks can be employed to model ZFF signals for paralinguistic applications, such as predicting factors like sleepiness [66] and dementia [67].

5.2.3 Composite signals

To collect information related to the vocal tract system we created a signal composed of different zero-frequency filtered signals by combining the filter outputs together to compose a signal carrying F1 and F2 related information. Then, a dynamic threshold was applied based on spectral entropy-based weighting [68].

5.2.4 Comparison between signals

For a more intuitive understanding of filtering on datasets, let's compare the spectrograms of four typical situations that can occur in our datasets:

1. *Healthy without Depression*: Subfigure (a) in figure 5.5 refers to 5s chunk "475_P_11" of DAIC-WOZ dataset. The selected patient is a healthy male.

2. *Healthy with Depression*: Subfigure (b) in figure 5.5 refers to 7s chunk "339_P_10" of DAIC-WOZ dataset. The selected patient is a healthy male with depression.

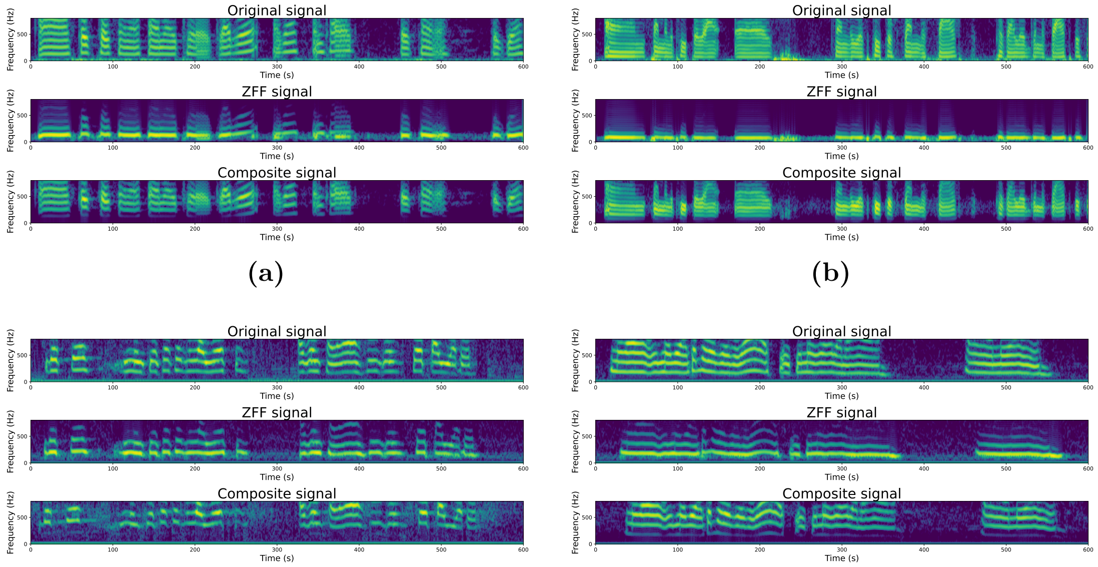
3. *Parkinson's disease without Depression*: Subfigure (b) in figure 5.5 refers to 7s of audio, clipped from the signal "005PD" in PD-D. It is a female patient with Parkinson's disease and depression.

4. *Parkinson's disease with depression*: Subfigure (b) in figure 5.5 refers to the first 6s of audio, clipped from the signal "006PD" in PD-D. It is a female patient with Parkinson's disease and depression.

Each subfigure in Figure 5.5 displays, from top to bottom, the original raw speech, the ZF-filtered signal, and the composite signal of the narrowband spectrogram for each situation described above. We use a narrowband representation of the spectrogram because it provides better frequency resolution and lets us visualize the fundamental frequency F0.

As mentioned in Section 5.2.2, using the Zero Frequency Filter on a vocal audio signal isolates and emphasizes fundamental information associated with the vocal source, highlighting the glottal closure instances and vocal pulse characteristics. The spectrograms at the centre of the subfigures in Figure 5.5 highlight these differences. The original signal's spectrogram (at the top of each subfigure) depicts a broad range of frequencies, whereas the ZFF signal's spectrogram is centred around vocal-relevant features such as the fundamental frequency (F0) and eliminates non-relevant information for source analysis.

In contrast, the composite signal, derived from the combination of zero-frequency filtered signals of the first two formats (see Section 5.2.3) and depicted in the spectrogram at the bottom of each subfigure, exhibits more pronounced resonances around specific frequencies. This reflects a notable concentration of information related to the formants and a decrease in the intensity of the F0, indicating an attenuation of information not directly linked to vocal tract characteristics. These differences indicate an accentuation and a more focused analysis of specific acoustic features of the vocal tract in the composite signal.



(a)

(b)

(c)

(d)

Figure 5.5: Comparison between narrow spectrograms between (a) Healthy without depression male patient, (b) Healthy with depression male patient; (c) female patient with Parkinson's disease without depression, (d) female patient with Parkinson's disease with depression.

5.3 Results

The following section describes the results of experiments on PD-D and DAIC datasets applying the End-to-End CNN method. The experiments with the three signals, Raw speech, ZFF signal and Composite signal will be compared in terms of network performance, and an interpretation of the results will be provided.

5.3.1 PD results

The table in Table 5.2 shows the CNN approach results for the dataset on Depression in Parkinson's disease. The first column indicates the experiment type, including the signal type (Raw speech, Zero frequency filtered, or Composite) and architecture type (segmental or subsegmental). The F1 score, precision, and recall values are grouped for each class - depressed (D) and non-depressed (ND) - in the final three columns. The optimal classification model was selected based on the unweighted average of the F1 score values for the two labels. The value in bold denotes the best outcome found in the PD-D database for differentiating between depressed and non-depressed individuals.

The model's predictions on the test samples could have been more satisfactory. However, due to the insufficient size of the dataset, it was challenging to use neural network models that needed many samples in the training phase. This led to the model's almost random behaviour in predicting the test samples.

However, having already carried out experiments using the first method (see Chapter 4) allowed us to anticipate the possible behaviour of the CNN when given the Composite and the ZFF signals. Even though the highest F1 score is obtained by feeding the CNN with the raw speech signal, we can see that CS in both structures (seg and subseg) behaves better than the zero frequency filtered audio.

It is important to recall that the composite signal carries information related to the first two formants, which are characteristics of the vocal tract. Thus, our E2E experiments confirmed the hypothesis that Parkinson's disease affects the vocal tract's characteristics, such as tremors and muscle rigidity, affecting vocal production.

5.3.2 DAIC results

Table 5.3 shows the results obtained by using CNN architectures on the DAIC-WOZ dataset. The table presents an overview of the outcome for each column, which represents *control* (C) and *depressed* (D) individuals, respectively. The label *O* represents the unweighted average of the F1 score values for the two classes. Each value displayed is the average performance obtained by training the proposed CNN five times. This was done to ensure that the proposed methods are not sensitive to CNN initialization and that the outcomes are repeatable. The best F1 score was achieved using the ZFF signal with the *subseg* architecture model, producing an overall value of 0.57 and 0.54 for depressed individuals. The confusion matrix in Figure 5.6 subfigure (b) represents the true positives, false positives, false negatives, and true negatives produced by the best performing seed in the ZFF signal experiment. Overall, the F1 score values are around 0.5, with slight variations in the second decimal place. While no particularly distinctive or enlightening

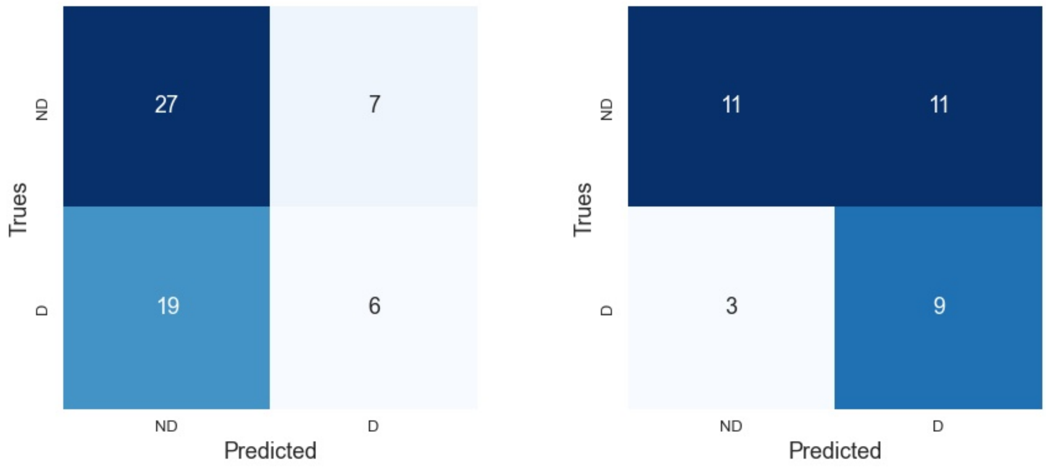
Experiment	*F1-score*			*Precision*		*Recall*	
	O	D	ND	D	ND	D	ND
PD-D							
Raw speech - *subseg*	**0.5**	0.32	0.68	0.46	0.59	0.24	0.79
Raw speech - *seg*	0.4	0.12	0.67	0.29	0.56	0.08	0.65
ZFF signal - *subseg*	-	-	0.64	-	0.53	-	0.83
ZFF signal - *seg*	0.35	0.06	0.64	0.12	0.53	0.04	0.79
CS signal - *subseg*	0.44	0.22	0.66	0.4	0.55	0.15	0.82
CS signal - *seg*	0.41	0.17	0.65	0.33	0.53	0.12	0.82

Table 5.2: The table shows the performances of different methods and architectures used on PD-D dataset. D indicates *depressed*, ND indicates *not-depressed* and O denotes the *overall* score by unweighted average over the two classes. The bold type value indicates the best system in terms of F1 score.

results were found, this study provides an opportunity for further improvement in future studies through protocol variations, diverse data augmentation techniques, or exploiting self-supervised learning methods (SSL). Finally, we confirmed our hypothesis that for healthy patients, feeding a CNN with signals containing source-related information produces the best results in classifying depression, even if only by a small margin.

Experiment	*F1-score*			*Precision*		*Recall*	
	O	D	C	D	C	D	C
DAIC-WOZ							
Raw speech - *subseg*	0.52	0.48	0.56	0.39	0.69	0.56	0.52
Raw speech - *seg*	0.55	0.52	0.59	0.424	0.73	0.67	0.50
ZFF signal - *subseg*	**0.57**	0.54	0.60	0.43	0.75	0.70	0.50
ZFF signal - *seg*	0.55	0.52	0.59	0.42	0.73	0.67	0.50
CS signal - *subseg*	0.53	0.50	0.56	0.40	0.70	0.67	0.46
CS signal - *seg*	0.55	0.52	0.58	0.42	0.73	0.58	0.49

Table 5.3: The table shows the performances of different methods and architectures used on DAIC-WOZ dev set. D indicates *depressed*, C indicates *control* and O denotes the *overall* score by unweighted average over the two classes. The bold type value indicates the best system in terms of F1 score.



(a) Comfusion matrix on the PD-D set using raw speech signals and the *subsegmental* model.

(b) Confusion Matrix of the best performing seed in the ZFF signal experiment for DAIC-WOZ database.

Figure 5.6: Confusion matrices for depression prediction on (a) PD-D and (b) DAIC-WOZ dataset using CNN approach

5.4 Summary

In this chapter, we discuss the results of our investigation into the use of convolutional neural networks (CNNs) that were fed with three different input signals to model the source, filter, and overall combined information. The three signals were the ZFF signal, the composite signal, and the original raw speech.

Our findings suggest that the information contained in the F0 signal is the most useful for the network to classify depressed patient in DAIC. However, in the case of Parkinson's patients, there is a drop in the value of the F1 score when the network is fed with ZFF signals or Composite signals when using the subseg architecture. This drop in score may be due to vocal tract articulation problems experienced by these patients.

Overall, our results suggest that all spectral content may be useful in the detection of depression in pathological patients.

Chapter 6

Conclusions

The present research aims to investigate the potential of utilizing speech analysis as a reliable and non-invasive tool for detecting depression. The primary focus of this study lies in examining the acoustic traits present in healthy speech associated with depression and exploring how these features can be distinguished from pathological speech patterns.

The thesis investigated two alternative methods for depression detection. The first approach involves a traditional process that extracts handcrafted features such as *eGeMAPS* and *ComPARE* from the audio signal, followed by a classification module utilizing diverse machine learning models. The second method involves utilizing convolutional neural networks (CNNs) that were fed with three different signals as input- the ZFF signal, the composite signal, and the original raw speech signal, to model source, filter, and overall combined information respectively.

To conduct our experiments, we used two datasets: Depression in Parkinson's disease (PD-D) and The Distress analysis interview corpus Wizard of Oz (DAIC-WOZ).

Throughout the experiments, the guiding hypothesis was to use the knowledge from pre-existing literature that vocal source features could be accurate for depression. The study of healthy patients with depression suggests this assumption. The top-10 most indicative features derived from the Gradient Boosting classifier, which was found to be the best in terms of performance, were mostly source-related. This result was confirmed by the application of the CNN approach, which showed that Zero Frequency filtered signals, carrying information related to the fundamental frequency of the spectrum, had the highest F1 score value of 0.57.

In contrast, the study of patients with Parkinson's disease and depression was more controversial. The best F1 score was achieved by providing the CNN with the original raw signal, including all spectral content, scoring 0.5. Interestingly, the performance trend of the *seg* architecture remains almost unchanged when working only with the F0 frequency or a composite signal. However, the *subseg* model performance worsens to 0.4 for composite signals. This outcome could suggest that retaining the information from the fundamental frequency F0 significantly improves results.

Further analysis revealed that the most relevant descriptors for discrimination of depression in pathological patients are features influenced by both source-related and vocal tract-related components. The development implies that the presence of motor impairments, such as tremors and hypokinetic dysarthria (articulation difficulties), may be the cause for the observed outcome. The comparison of mutual information values between features and datasets leaves no doubt that the results point out that depression cannot be treated in Parkinson's patients without taking into account the impact of the illness on their speaking ability.

The thesis attempted to interpret the outlines obtained through the utilization of two methods and to extract progress knowledge from the current literature. However, recent advancements in self-supervised learning techniques and State-of-the-Art models have achieved better results. Hence, for future works, it would be advantageous to emphasize improving the performance of the model while also leveraging the insights from this thesis work to make better-informed decisions.

Appendix A

openSMILE features

Feature	Description
Waveform	Zero-Crossings, Extremes, DC
Loudness	Energy, intensity, auditory model loudness
FFT spectrum	Phase, magnitude (lin., dB, dBA)
ACF Cepstrum	Autocorrelation and Cepstrum
Mel/Bark spectr.	Bands 0-N_{mel}
Semitone spectr.	FFT based and filter based
Cepstral	Cepstral features, e.g. MFCC, PLP- CC
Pitch	F0 via Autocorrelation and sub-harmonic summation, smoothed by Viterbi algorithm
Voice Quality	HNR, Jitter, Shimmer, Voice Prob
LPC	LPC coeff., reflect. coeff., residual Line spectral pairs (LSP)
Auditory	Auditory spectra, psychoacoustic sharpness
Formants	Centre frequencies and bandwidths
Spectral	Energy in N user-defined bands, roll-off points, centroid, entropy, flux, and rel. pos. of max./min., har- monicity
Tonal	CHROMA, CENS, CHROMA-based features

Table A.1: openSMILE's low-level-descriptors

6 Frequency related parameter	**Meaning**
Pitch	logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz (semitone 0)
Jitter	deviations in individual consecutive F0 period lengths
Formant 1, 2, and 3	frequency, centre frequency of first, second, and third formant
Formant 1	bandwidth of first formant.
3 Energy/amplitude related parameter	**Meaning**
Shimmer	difference of the peak amplitudes of consecutive F0 periods
Loudness	estimate of perceived signal intensity from an auditory spectrum
Harmonics-to-Noise Ratio (HNR)	relation of energy in harmonic components to energy in noise-like components
9 Spectral (balance) parameters	**Meaning**
Alpha Ratio	Ratio of the summed energy from 50–1000 Hz and 1–5 kHz
Hammarberg Index	Ratio of the strongest energy peak in the 0–2 kHz region to the strongest peak in the 2–5 kHz region
Spectral Slope 0–500 Hz and 500–1500 Hz	Linear regression slope of the logarithmic power spectrum within the two given bands
Formant 1, 2, and 3 relative energy	Ratio of the energy of the spectral harmonic peak at the first, second, third formant's centre frequency to the energy of the spectral peak at F0
Harmonic difference H1–H2	Ratio of energy of the first F0 harmonic (H1) to the energy of the second F0 harmonic (H2)
Harmonic difference H1–A3	Ratio of energy of the first F0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3)

Table A.2: Minimalistic parameter set (GeMAPS)

LLD	Functionals
Group A: 59	
Loudness	Root-quadratic mean, flatness
Modulation loudness	Standard deviation, skewness, kurtosis
RMS energy, ZCR	Quartiles 1–3
RASTA auditory bands 1–26	Inter-quartile ranges 1–2, 2–3, 1-3
MFCC 1–14	99th and 1-st percentile, range of these
Energy 250–650 Hz	Relative position of max. and min. value
Energy 1–4 kHz	Range (maximum to minimum value)
Spectral RoP .25, .50, .75, .90	Linear Regression slope
Spectral flux, entropy, variance	Linear regression quadratic error
Spectral skewness and kurtosis	Quadratic regression coeff.
Spectral slope	Quadratic regression quadratic error
Spectral harmonicity	Temporal centroid
Spectral sharpness (auditory)	Peak mean value and dist. to arithm. mean
Spectral centroid (linear)	Mean and std. dev. of peak to peak distances. Peak and valley range (absolute and relative), Peak-valley-peak slopes mean and std. dev.
Group B: 6	
F0 via SHS, Prob. of voicing	Segment length mean, min., max., std. dev.
Jitter (local and delta)	Up-level time 25%, 50%, 75%, 90%
Shimmer	Rise time, left curvature time
logHNR(time domain)	Linear Prediction gain and coefficients 1–5

Table A.3: INTERSPEECH 2013 Computational Paralinguistics ChallengE (ComParE) set

Appendix B

Mutual information

Feature name	Source	System	Global	MI-Depression	MI-Parkinson
F0final	X			0.0568	0.1692
audSpec_Rfilt		X		0.1557	0.2115
audspecRasta_lengthL1norm		X		0.1162	0.1135
audspec_lengthL1norm		X		0.1182	0.1142
jitterDDP	X			0.1753	0.068
jitterLocal	X			0.0703	0.1178
logHNR	X			0.116	0.0777
mfcc		X		0.2189	0.1803
pcm_RMSenergy			X	0.1308	0.0709
pcm_fftMag_fband1000-4000			X	0.0502	0.0908
pcm_fftMag_fband250-650	X			0.0684	0.1181
pcm_fftMag_psySharpness			X	0.0953	0.2146
pcm_fftMag_spectralCentroid		X		0.0866	0.1893
pcm_fftMag_spectralEntropy			X	0.1291	0.1859
pcm_fftMag_spectralFlux			X	0.1238	0.1
pcm_fftMag_spectralHarmonicity		X		0.0651	0.0968
pcm_fftMag_spectralKurtosis			X	0.1277	0.0997
pcm_fftMag_spectralRollOff25.0			X	0.1027	0.1429
pcm_fftMag_spectralRollOff50.0			X	0.1195	0.1649
pcm_fftMag_spectralRollOff75.0			X	0.1526	0.1647
pcm_fftMag_spectralRollOff90.0		X		0.1321	0.176
pcm_fftMag_spectralSkewness			X	0.1537	0.1517
pcm_fftMag_spectralSlope			X	0.1055	0.0747
pcm_fftMag_spectralVariance			X	0.1069	0.1663
pcm_zcr	X			0.0835	0.1919
shimmerLocal	X			0.0582	0.1134
voicingFinalUnclipped	X			0.1364	0.1345

Table B.1: Entire list of Mutual information values computed in the study

Feature name	**Meaning**
F0final	The smoothed fundamental frequency contour
audSpec_Rfilt	Relative Spectral Transform (RASTA)-style filtered applied to Auditory Spectrum
audspecRasta_lengthL1norm	Relative Spectral Transform applied to Auditory Spectrum and engthL1norm is the magnitude of the L1 norm
audspec_lengthL1norm	Magnitude of L1 norm of Auditory Spectrum
jitterDDP	The differential frame-to-frame Jitter (the ‘Jitter of the Jitter’)
jitterLocal	The local (frame-to-frame) Jitter (pitch period length deviations)
logHNR	Log of the ratio of the energy of harmonic signal components to the energy of noise like signal components
mfcc	Mel-frequency cepstral coefficients 1–14
pcm_RMSenergy	Root-mean-square signal frame energy
pcm_fftMag_fband1000-4000	fft magnitude of frequency band between 1000Hz to 4000Hz
pcm_fftMag_fband250-650	fft magnitude of frequency band between 250Hz to 650Hz
pcm_fftMag_psySharpness	Psychoacoustic sharpness
pcm_fftMag_spectralCentroid	Spectral Features, represents the centre of gravity of the signal’s spectral content
pcm_fftMag_spectralEntropy	Measures the randomness of a signal’s spectral content
pcm_fftMag_spectralFlux	Evaluates the temporal variation of the logarithmically-scaled rate-map across adjacent frames
pcm_fftMag_spectralHarmonicity	Spectral Harmonicity

pcm_fftMag_spectralKurtosis	Indicates the presence of series of transients and their locations in the frequency domain.
pcm_fftMag_spectralRollOff25.0	Represents the frequency below which 25 percentage of the total spectral energy lies
pcm_fftMag_spectralRollOff50.0	Represents the frequency below which 50 percentage of the total spectral energy lies.
pcm_fftMag_spectralRollOff75.0	Represents the frequency below which 75 percentage of the total spectral energy lies
pcm_fftMag_spectralRollOff90.0	Represents the frequency below which 90 percentage of the total spectral energy lies
pcm_fftMag_spectralSkewness	Measures the symmetry of the spectrum around its arithmetic mean. The feature will be zero for silent segments and high for voiced speech where substantial energy is present around the fundamental frequency.
pcm_fftMag_spectralSlope	It is a high-frequency response of spectrum calculated using linear regression and the central wavelet of the signal for a window
pcm_fftMag_spectralVariance	Measures the signal's spectral content variability over time
pcm_zcr	Zero-crossing rate of time signal (frame-based)
shimmerLocal	The local (frame-to-frame) Shimmer (amplitude deviations between pitch periods)
voicingFinalUnclipped	The voicing probability of the final fundamental frequency candidate. Unclipped means,that it was not set to zero when is falls below the voicing threshold

Table B.2: ComPARE features explanation. Reference by [69] [70]

Bibliography

[1] *World Health Organization "Depression".* `https://www.who.int/news-room/fact-sheets/detail/depression`. Accessed: November 19, 2023. (cit. on p. 1).

[2] Florian Hönig, Anton Batliner, Elmar Nöth, Sebastian Schnieder, and Jarek Krajewski. «Automatic modelling of depressed speech: relevant features and relevance of gender». In: (2014) (cit. on p. 1).

[3] H WHO. «Depression: A global crisis». In: *World Ment Heal Day [Internet]* (2012) (cit. on p. 1).

[4] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. «Avec 2014: 3d dimensional affect and depression recognition challenge». In: *Proceedings of the 4th international workshop on audio/visual emotion challenge.* 2014, pp. 3–10 (cit. on p. 1).

[5] V Mitra and E Shriberg. «Effects of feature type, learning algorithm and speaking style for depression detection from speech. Acoustics, Speech and Signal Processing (ICASSP)». In: *2015 IEEE International Conference on IEEE.* 2015, pp. 19–24 (cit. on p. 1).

[6] Bernd Löwe, Jürgen Unützer, Christopher M Callahan, Anthony J Perkins, and Kurt Kroenke. «Monitoring depression treatment outcomes with the patient health questionnaire-9». In: *Medical care* (2004), pp. 1194–1201 (cit. on p. 1).

[7] Ran Ha Hong et al. «Implementing measurement-based care for depression: practical solutions for psychiatrists and primary care physicians». In: *Neuropsychiatric Disease and Treatment* (2021), pp. 79–90 (cit. on p. 1).

[8] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. «A review of depression and suicide risk assessment using speech analysis». In: *Speech communication* 71 (2015), pp. 10–49 (cit. on p. 1).

[9] Haihua Jiang, Bin Hu, Zhenyu Liu, Lihua Yan, Tianyang Wang, Fei Liu, Huanyu Kang, and Xiaoyu Li. «Investigation of different speech types and emotions for detecting depression using different classifiers». In: *Speech Communication* 90 (2017), pp. 39–46 (cit. on p. 1).

[10] Rachelle Horwitz, Thomas F Quatieri, Brian S Helfer, Bea Yu, James R Williamson, and James Mundt. «On the relative importance of vocal source, system, and prosody in human depression». In: *2013 IEEE international conference on body sensor networks*. IEEE. 2013, pp. 1–6 (cit. on p. 2).

[11] Antonina Kouli, Kelli M Torsney, and Wei-Li Kuan. «Parkinson's disease: etiology, neuropathology, and pathogenesis». In: *Exon Publications* (2018), pp. 3–26 (cit. on p. 2).

[12] Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass. «Detecting Depression with Audio/Text Sequence Modeling of Interviews.» In: *Interspeech*. 2018, pp. 1716–1720 (cit. on p. 2).

[13] Hamdi Dibeklioğlu, Zakia Hammal, and Jeffrey F Cohn. «Dynamic multimodal measurement of depression severity using deep autoencoding». In: *IEEE journal of biomedical and health informatics* 22.2 (2017), pp. 525–536 (cit. on p. 2).

[14] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas Quatieri. «A review of depression and suicide risk assessment using speech analysis». In: *Speech Communication* 71 (Apr. 2015) (cit. on p. 2).

[15] Stefan Scherer, Gale Lucas, Jonathan Gratch, Albert Rizzo, and Louis-Philippe Morency. «Self-Reported Symptoms of Depression and PTSD Are Associated with Reduced Vowel Space in Screening Interviews». In: *IEEE Transactions on Affective Computing* 7 (Jan. 2015), pp. 1–1. DOI: `10.1109/TAFFC.2015.2440264` (cit. on p. 2).

[16] Amber Afshan, Jinxi Guo, Soo Jin Park, Vijay Ravi, Jonathan Flint, and Abeer Alwan. «Effectiveness of Voice Quality Features in Detecting Depression». In: *Proc. Interspeech 2018*. 2018, pp. 1676–1680. DOI: `10.21437/Interspeech.2018-1399` (cit. on p. 3).

[17] Saurabh Sahu and Carol Espy-Wilson. «Speech Features for Depression Detection». In: Sept. 2016, pp. 1928–1932. DOI: `10.21437/Interspeech.2016-1566` (cit. on p. 3).

[18] Amber Afshan, Jinxi Guo, Soo Jin Park, Vijay Ravi, Jonathan Flint, and Abeer Alwan. «Effectiveness of Voice Quality Features in Detecting Depression». In: *Interspeech 2018* () (cit. on p. 3).

[19] Ray D Kent and Y-J Kim. «Toward an acoustic typology of motor speech disorders». In: *Clinical linguistics & phonetics* 17.6 (2003), pp. 427–445 (cit. on p. 3).

[20] Brian Stasak, Julien Epps, Nicholas Cummins, and Roland Goecke. «An Investigation of Emotional Speech in Depression Classification». In: *Proc. Interspeech 2016*. 2016, pp. 485–489. DOI: 10.21437/Interspeech.2016-867 (cit. on p. 3).

[21] Lang He and Cui Cao. «Automated depression analysis using convolutional neural networks from speech». In: *Journal of biomedical informatics* 83 (2018), pp. 103–111 (cit. on p. 3).

[22] Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. «Depaudionet: An efficient deep model for audio based depression classification». In: *Proceedings of the 6th international workshop on audio/visual emotion challenge*. 2016, pp. 35–42 (cit. on p. 3).

[23] L. R. Rabiner, B. Gold, and C. K. Yuen. «Theory and Application of Digital Signal Processing». In: *IEEE Transactions on Systems, Man, and Cybernetics* 8.2 (1978), pp. 146–146. DOI: 10.1109/TSMC.1978.4309918 (cit. on p. 5).

[24] Shaykhah Almaghrabi, Scott Clark, and Mathias Baumert. «Bio-acoustic features of depression: A review». In: *Biomedical Signal Processing and Control* 85 (May 2023). DOI: 10.1016/j.bspc.2023.105020 (cit. on p. 6).

[25] Alistair J Flint, Sandra E Black, Irene Campbell-Taylor, Gillian F Gailey, and Carey Levinton. «Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression». In: *Journal of psychiatric research* 27.3 (1993), pp. 309–319 (cit. on pp. 7, 8).

[26] F Tolkmitt, H Helfrich, Rt Standke, and Klaus R Scherer. «Vocal indicators of psychiatric treatment effects in depressives and schizophrenics». In: *Journal of communication disorders* 15.3 (1982), pp. 209–222 (cit. on p. 8).

[27] Harry Hollien. «Vocal indicators of psychological stress». In: *Annals of the New York Academy of Sciences* 347.1 (1980), pp. 47–72 (cit. on p. 8).

[28] Lu-Shih Alex Low, Namunu C Maddage, Margaret Lech, Lisa B Sheeber, and Nicholas B Allen. «Detection of clinical depression in adolescents' speech during family interactions». In: *IEEE Transactions on Biomedical Engineering* 58.3 (2010), pp. 574–586 (cit. on p. 8).

[29] Brian S Helfer, Thomas F Quatieri, James R Williamson, Daryush D Mehta, Rachelle Horwitz, and Bea Yu. «Classification of depression state based on articulatory precision.» In: *Interspeech*. 2013, pp. 2172–2176 (cit. on p. 8).

[30] Marc D Pell, Henry S Cheang, and Carol L Leonard. «The impact of Parkinson's disease on vocal-prosodic communication from the perspective of listeners». In: *Brain and language* 97.2 (2006), pp. 123–134 (cit. on p. 9).

[31] Florian Eyben, Martin Wöllmer, and Björn Schuller. «Opensmile: the munich versatile and fast open-source audio feature extractor». In: *Proceedings of the 18th ACM international conference on Multimedia.* 2010, pp. 1459–1462 (cit. on p. 10).

[32] Florian Eyben et al. «The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing». In: *IEEE Transactions on Affective Computing* 7.2 (Apr. 2016), pp. 190–202. DOI: `10.1109/TAFFC.2015.2457417` (cit. on p. 10).

[33] Ziqiang Bao, Shuai Zhao, Shuang Li, Guisong Jiang, Huazhi Sun, and Long Zhang. «Multi-dimensional Convolutional Neural Network for Speech Emotion Recognition». In: *International conference on Smart Technologies and Systems for Internet of Things.* Springer. 2021, pp. 296–303 (cit. on p. 10).

[34] Florian Eyben and Florian Eyben. «Acoustic features and modelling». In: *Real-time Speech and Music Classification by Large Audio Feature Space Extraction* (2016), pp. 9–122 (cit. on p. 10).

[35] Björn Schuller et al. «The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism». In: Aug. 2013, pp. 148–152. DOI: `10.21437/Interspeech.2013-56` (cit. on p. 11).

[36] Björn Schuller et al. «The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language». In: *17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), Vols 1-5.* Vol. 8. ISCA. 2016, pp. 2001–2005 (cit. on p. 11).

[37] Maximilian Schmitt, Fabien Ringeval, and Björn Schuller. «At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech». In: Sept. 2016, pp. 495–499. DOI: `10.21437/Interspeech.2016-1124` (cit. on p. 12).

[38] Corinna Cortes and Vladimir Vapnik. «Support-vector networks». In: *Machine learning* 20 (1995), pp. 273–297 (cit. on p. 12).

[39] Steve R Gunn et al. «Support vector machines for classification and regression». In: *ISIS technical report* 14.1 (1998), pp. 5–16 (cit. on p. 12).

[40] Dimitri Palaz, Ronan Collobert, and Mathew Magimai Doss. «Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks». In: *arXiv preprint arXiv:1304.1018* (2013) (cit. on p. 15).

[41] Dimitri Palaz, Mathew Magimai-Doss, and Ronan Collobert. «End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition». In: *Speech Communication* 108 (2019), pp. 15–32 (cit. on p. 15).

[42] Steve Renals and Pawel Swietojanski. «Neural networks for distant speech recognition». In: *2014 4th joint workshop on hands-free speech communication and microphone arrays (HSCMA)*. IEEE. 2014, pp. 172–176 (cit. on p. 15).

[43] Hannah Muckenhirn, Mathew Magimai Doss, and Sébastien Marcell. «On Learning Vocal Tract System Related Speaker Discriminative Information from Raw Signal Using CNNs.» In: *Interspeech*. 2018, pp. 1116–1120 (cit. on p. 15).

[44] Hannah Muckenhirn, Mathew Magimai Doss, and Sébastien Marcell. «Towards directly modeling raw speech signal for speaker verification using CNNs». In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 4884–4888 (cit. on pp. 15, 40).

[45] Selen Hande Kabil, Hannah Muckenhirn, and Mathew Magimai-Doss. «On Learning to Identify Genders from Raw Speech Signal Using CNNs.» In: *Interspeech*. 2018, pp. 287–291 (cit. on p. 15).

[46] Tilak Purohit, Sarthak Yadav, Bogdan Vlasenko, S Pavankumar Dubagunta, and Mathew Magimai- Doss. «Towards Learning Emotion Information from Short Segments of Speech». In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5 (cit. on p. 15).

[47] Paula Perez-Toro, Juan Vasquez, Tobias Bocklet, Elmar Noeth, and Juan Rafael Orozco. «User State Modeling Based on the Arousal-Valence Plane: Applications in Customer Satisfaction and Health-Care». In: *IEEE Transactions on Affective Computing* PP (Sept. 2021), pp. 1–1. DOI: `10.1109/TAFFC.2021.3112543` (cit. on p. 17).

[48] Jonathan Gratch et al. «The distress analysis interview corpus of human and computer interviews.» In: *LREC*. Reykjavik. 2014, pp. 3123–3128 (cit. on p. 18).

[49] Amy Fan, Tara Strine, Youjie Huang, Melissa Jordan, Senyoni Musingo, Ruth Jiles, and Ali Mokdad. «Self-Rated Depression and Physician-Diagnosed Depression and Anxiety in Florida Adults: Behavioral Risk Factor Surveillance System, 2006». In: *Preventing chronic disease* 6 (Feb. 2009), A10 (cit. on p. 18).

[50] Michel Valstar et al. «Avec 2016: Depression, mood, and emotion recognition workshop and challenge». In: *Proceedings of the 6th international workshop on audio/visual emotion challenge*. 2016, pp. 3–10 (cit. on p. 21).

[51] Thomas F Quatieri and Nicolas Malyska. «Vocal-source biomarkers for depression: A link to psychomotor activity». In: *Thirteenth annual conference of the international speech communication association*. 2012 (cit. on p. 27).

[52] Kris Tjaden. «Speech and swallowing in Parkinson's disease». In: *Topics in geriatric rehabilitation* 24.2 (2008), p. 115 (cit. on pp. 27, 35).

[53] Peter Pabon and Sten Ternström. «Feature Maps of the Acoustic Spectrum of the Voice». In: *Journal of Voice* 34.1 (2020), 161.e1–161.e26 (cit. on p. 33).

[54] Aik Ming Toh, Roberto Togneri, and Sven Nordholm. «Spectral entropy as speech features for speech recognition». In: *Proceedings of PEECS* 1 (2005), p. 92 (cit. on p. 33).

[55] Md Gulzar Hussain, Mahmuda Rahman, Babe Sultana, Ayesha Khatun, and Sakib Al Hasan. «Classification of Bangla Alphabets Phoneme based on Audio Features using MLPC & SVM». In: *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)*. IEEE. 2021, pp. 1–5 (cit. on p. 33).

[56] Marko Kos, Zdravko Kačič, and Damjan Vlaj. «Acoustic classification and segmentation using modified spectral roll-off and variance-based features». In: *Digital Signal Processing* 23.2 (2013), pp. 659–674 (cit. on p. 33).

[57] openSMILE. *"cSpectral"*. URL: %5Curl%7Bhttps://audeering.github.io/opensmile/_components/cSpectral.html#cspectral%7D (cit. on p. 33).

[58] *ScienceDirect "Spectral Centroids"*. https://www.sciencedirect.com/topics/engineering/spectral-centroid (cit. on p. 33).

[59] Theodoros Giannakopoulos and Aggelos Pikrakis. *Introduction to audio analysis: a MATLAB® approach*. Academic Press, 2014, pp. 69–103 (cit. on p. 33).

[60] João Paulo Teixeira, Carla Oliveira, and Carla Lopes. «Vocal acoustic analysis–jitter, shimmer and hnr parameters». In: *Procedia Technology* 9 (2013), pp. 1112–1122 (cit. on p. 33).

[61] Juan Rafael Orozco-Arroyave, Julián David Arias-Londoño, Jesús Francisco Vargas-Bonilla, María Claudia Gonzalez-Rátiva, and Elmar Nöth. «New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease.» In: *LREC*. 2014, pp. 342–347 (cit. on p. 36).

[62] M Abadi, A Agarwal, P Barham, et al. *TensorFlow: Large-scale Machine Learning on Heterogeneous Systems, v1. 14*. 2015 (cit. on p. 40).

[63] François Chollet. «Keras». In: 2015 (cit. on p. 40).

[64] K Sri Rama Murty and Bayya Yegnanarayana. «Epoch extraction from speech signals». In: *IEEE Transactions on Audio, Speech, and Language Processing* 16.8 (2008), pp. 1602–1613 (cit. on p. 42).

[65] B Yegnanarayana and Suryakanth V Gangashetty. «Epoch-based analysis of speech signals». In: *Sadhana* 36 (2011), pp. 651–697 (cit. on p. 42).

[66] Julian Fritsch, S Pavankumar Dubagunta, and Mathew Magimai- Doss. «Estimating the degree of sleepiness by integrating articulatory feature knowledge in raw waveform Based CNNS». In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 6534–6538 (cit. on p. 43).

[67] Nicholas Cummins et al. «A comparison of acoustic and linguistics methodologies for Alzheimer's dementia recognition». In: *Interspeech 2020*. ISCA-International Speech Communication Association. 2020, pp. 2182–2186 (cit. on p. 43).

[68] Eklavya Sarkar, RaviShankar Prasad, and Mathew Magimai Doss. «Unsupervised Voice Activity Detection by Modeling Source and System Information using Zero Frequency Filtering». In: *arXiv preprint arXiv:2206.13420* (2022) (cit. on p. 43).

[69] Alexander Lerch. *An introduction to audio content analysis: Applications in signal processing and music informatics*. Wiley-IEEE Press, 2012 (cit. on p. 57).

[70] Eric Scheirer and Malcolm Slaney. «Construction and evaluation of a robust multifeature speech/music discriminator». In: *1997 IEEE international conference on acoustics, speech, and signal processing*. Vol. 2. IEEE. 1997, pp. 1331–1334 (cit. on p. 57).