Transfer learning strategies to model speech impairment in patients with neurodegenerative diseases

Cristian David Ríos Urrego

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

Director
Prof. Dr.-Ing. Juan Rafael Orozco Arroyave

Asesor
MSc. Juan Camilo Vásquez-Correa

Universidad de Antioquia
Facultad de Ingeniería
Maestría en Ingeniería de Telecomunicaciones
Medellín, Antioquia, Colombia
2022

Cita	Rios Urrego [1]
Referencia	[1] C. D. Rios Urrego, "Transfer learning strategies to model speech impairment in patients with neurodegenerative diseases", Tesis de maestría, Maestría en Ingeniería de Telecomunicaciones, Universidad de Antioquia, Medellín, Antioquia, Colombia, 2022.
Estilo IEEE (2020)	

Maestría en Ingeniería de Telecomunicaciones, Cohorte XV
Grupo de Investigación Telecomunicaciones Aplicadas (GITA)

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia
Sistema de Bibliotecas
Biblioteca Carlos Gaviria Díaz

Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes
Decano/Director Jesús Francisco Vargas Bonilla
Jefe departamento: Augusto Enrique Salazar Jiménez

Transfer learning strategies to model speech impairment in patients with neurodegenerative diseases

UNIVERSIDAD DE ANTIOQUIA

1 8 0 3

Master's Thesis in Telecommunication Engineering

Cristian David Ríos-Urrego

Director: Prof. Dr.-Ing. Juan Rafael Orozco Arroyave
Advisor: MSc. Juan Camilo Vásquez-Correa

Faculty of Engineering

Department of Electronics and Telecommunications

University of Antioquia.

Acknowledgments

First of all thanks to my mother Beatriz Urrego, my father Hernán Ríos and my brother Camilo Ríos, they made all this possible with their invaluable support, unconditional love, and understanding. I would also thank my family, who have trusted me for the culmination of my research work, everybody in my family has shown to me the value and power of patience and perseverance.

Now, I want to thank my colleagues Felipe López and Daniel Escobar, for supporting me in this process, giving me their friendship, and their academic. I also want to thank the pattern recognition team of GITA Lab in the University of Antioquia: Felipe Gómez, Paula Pérez, Felipe Parra, Tomás Arias, Surley Berrio, and Nicanor García, for helping me and supporting me in this academic process. Likewise, I want to thank the other members of the GITA Lab because they made all the working hours more enjoyable.

I want to express also my gratitude to my director Prof. Dr.-Ing. Juan Rafael Orozco, and my advisor MSc. Juan Camilo Vásquez, for the knowledge transferred that I was able to obtain through their experiences, for their support, for always being ready to answer any questions, and for their patience during this long research process.

Finally, I want to thank the University of Antioquia for its financial support during the development of this master's thesis through the master's scholarship and the CODI project by grant Number 2017–15530.

Abstract

Nowadays, the interest in the automatic analysis of speech in different scenarios has increased. This biomarker has been explored for diagnostic support and monitoring of different neurodegenerative diseases such as Parkinson's disease (PD) and Huntington's disease (HD). Speech has the main benefit of being a non-invasive method, which can be captured remotely, at a low cost, and includes detailed information about each participant. Deep learning (DL) has facilitated the development of different robust computational models, their main advantages are that these systems receive the raw signal at the input and can process a large amount of information in parallel. However, obtaining large amounts of data for pathological speech, particularly in neurodegenerative diseases, is very difficult and expensive. Therefore, it is necessary to implement DL techniques to address this issue. Transfer Learning (TL) takes advantage of the experience obtained in a previously trained model to improve the learning of new target phenomena. The main aim of this work is to evaluate the suitability of using DL for pathological speech processing applications, particularly using TL methods. Different approaches are implemented including transfer knowledge in classical methods by crossing pathologies and languages for the classification of PD and HD in the Czech language. In addition, different Convolutional Neural Networks (CNNs) are trained to perform a fine-tuning strategy from cross-language and cross-pathology for the discrimination of patients with PD and HD with respect to Healthy Control (HC) subjects and also the estimation of their depression level. Subsequently, a freezing of layers strategy was performed for the classification of PD patients vs. HC subjects in different languages. Finally, a DL strategy based on embeddings is proposed to know which additional demographic information a CNN learns from pathological speech data.

Contents

Chapter 1

Introduction

1.1 Motivation

In recent years, the research about biomarkers has advanced to support the field of medicine. Recent studies have shown that digital measures will provide great benefits to research and clinical practice [1], [2]. In addition, it will offer a promising solution to issues such as repetitive clinical assessment, variability between clinicians, time, and cost [3]. Specifically, speech is a biomarker that the scientific community has explored for diagnostics support and monitoring of different neurodegenerative diseases such as Parkinson's disease (PD) and Huntington's disease (HD) [4].

On the one hand, PD has been a focus of research because it is the second most prevalent degenerative disorder in the world after Alzheimer's disease [5]. PD is characterized by symptoms such as resting tremor, bradykinesia, rigidity and freezing of gait [6]. Most of PD patients develop several speech deficits which are grouped and called hypokinetic dysarthria. Dysarthric speech appears as the result of losing movement control of the muscles and limbs involved in the speech production process. Typical characteristics of hypokinetic dysarthria include monoloudness, reduced voice quality, monotonicity, imprecise pronunciation of consonants and vowels, lack of fluency, voice tremor, and other characteristics [7]. All these factors affect also the intelligibility of patients and limit their quality of life and their effective communication [8]. On the other hand, HD is not a very common disease since it only affects between four and seven people per 100,000 inhabitants [9]. However, different studies have been developed due to the fact that it is a very aggressive pathology. HD manifests mainly by involuntary

movements or chorea [10], in addition to cognitive deficits [11], dystonia, and rigidity that may appear in patients in the early stages of the disease [12]. HD patients develop hyperkinetic dysarthria, which appears primarily as a consequence of chorea. The most relevant deficits in speech include phonatory dysfunction, unpredictable interruptions of articulation, and abnormal prosody [13]. The reduced quality of the speech production in HD appears due to involuntary contractions of the muscles, especially when stable functioning is required, i.e., to produce sustained vowel phonations [14].

Although PD and HD patients exhibit different motor symptoms, they also share characteristics like reduced mood and metabolic dysfunction [15]. Additionally, both are diagnosed based on physical, neurological and psychiatric exams, family history, and others. Due to the way they are diagnosed and certain similarities between the two diseases, PD and HD can sometimes be confused in their diagnosis, which is a great problem because HD can be more aggressive than PD and the required therapy is different [16]–[18]. For these reasons, and due to the number of common symptoms and deficiencies in both diseases, different Machine learning (ML) techniques have been explored to characterize each of the pathologies in order to support the diagnosis and perform automatic monitoring of the patients [19], [20]. Even different computational tools have been developed to support diagnosis and monitoring of PD [21], [22]. One of the main problems related to speech signals in ML is that the inputs and outputs of the system have to be insensitive to external factors or co-variables such as accent, age, and gender. In addition to different transmission channels and microphones that are used [23]. The traditional strategy to reduce these problems is to work with a human expert who designs feature extraction strategies to model the information encoded in the collected data. However, this approach is expensive and time-consuming [24], [25].

Therefore, different Deep Learning (DL) techniques have been developed to mitigate the aforementioned problem of classical ML algorithms. These methods receive the raw signal in their inputs, e.g., an image or a speech signal, to extract information from them [26]. Therefore, DL reduces the necessity of an expert person for the feature extraction procedure. The basis of the DL is the Deep Neural Network (DNN), which is composed of multiple layers to process the input information and produce an output with the required predictions. DNNs perform both linear and non-linear operations to obtain patterns of a phenomenon and solve a specific problem [24].

DL has facilitated the development of different computational models to evaluate and monitor the neurological state of patients with different neurodegenerative diseases based on speech signals. This has the main benefit of being a non-invasive method, which can be captured remotely, at low cost, and including detailed information about each participant. However, obtaining large amounts of data is very difficult and expensive. Therefore, it is necessary to implement ML and DL techniques to address this issue. In this work, we aim to explore the usability of the DL for pathological speech processing applications, particularly using Transfer Learning (TL) methods where the main idea is to use the experience obtained in certain models to improve the learning of new target phenomena. The knowledge (features, weights, and biases) of previously trained models (commonly with large databases) are used to support the performance of the new ones, where typically not enough data is available [27], [28]. Different contributions in the literature have shown the importance of including different TL strategies for pathological speech classification. Several works have succeeded in training models with a small amount of data by performing database TL with a large number of iterations [29]–[31]. However, there are several questions regarding TL in the classification of pathological speech. The first question is associated with how closely the base model should be related to the target model. Some work has shown that CNNs trained with images can complement spectrogram classification in pathological speech [30], but what restrictions are involved in performing this adaptation. Another important aspect is to analyze how best to transfer the parameters of the base model to the target model: is it necessary to fine-tune all its parameters or to freeze all the characterization layers, even if a better performance can be obtained by partially freezing the transferred model.

1.2 Research problem

TL has enabled the development of robust models re-training of small databases. The main aim of this work is to evaluate to which extent TL methods can take information from an external model and contribute to improve the accuracy of another one by using speech utterances. During the development of this work, we will evaluate which percentage of information extracted from an original base model is required to perform the automatic classification of Huntington's and Parkinson's disease patients. We hypoth-

esize that the models trained with other pathologies with common symptoms can successfully complement the discrimination of patients by using TL techniques [29]. In addition, we will evaluate different TL methods in a cross-language setting, the main idea being to assess whether there are language limitations in transferring knowledge from a base model to a target model with the same or similar objective (classification of pathological speech). Finally, to generate discussion on some of the problems mentioned in the previous section, we will evaluate what other traits or aspects a deep neural network considers for the classification of pathological speech, in order to reduce the gap of *black-box* in deep neural networks, including to know and implement an appropriate TL strategy.

1.3 State of the art

1.3.1 Assessment of neurodegenerative diseases with machine learning methods

There are several studies focused on the automatic evaluation of HD using speech signals from classical ML techniques. For example in [32], the authors proposed to objectively identify phonatory changes that may occur in pre-motor states of HD by assessing a wide combination of measures of dysphonia including: airflow insufficiency, aperiodicity, irregular vibration of vocal folds, signal perturbations, increased noise, vocal tremor, and articulation deficiency. A total of 28 Czech native speakers with the gene positive premanifest HD, namely PreHD, and 28 Healthy Control (HC) subjects were included, each participant performed two repetitions of the vowel /ah/. The authors discriminated between the PreHD and HC subjects, and reported a sensitivity of 91.5% and a specificity of 79.2%. In [33], an analysis of the phonatory impairments in HD was performed considering 34 patients and 34 HC subjects. The subjects pronounced sustained phonations of the vowels /ah/ and /i/. The authors computed features such as maximum phonation time, number of voice breaks, fundamental frequency variations, jitter, shimmer, harmonics-to-noise ratio, detrended fluctuation analysis, the maximum phonation time until the first voice break and Mel Frequency Cepstral Coefficients (MFCCs). A group of features accurately discriminated between HD patients and HC subjects (fundamental frequency variations, maximum phonation time until the first voice break, detrended fluctuation analysis

and MFCCs). The combination of these features produced an accuracy up to 94.1%.

Regarding the evaluation and monitoring of PD, much more research has been developed in recent years. For example, in [34], the authors proposed a model to classify PD patients vs. HC subjects based on acoustic features computed from different allophones. Speech recognition technologies such as forced alignment were used to segment the speech signal. Then, the allophones were used to train individual Gaussian Mixture Models (GMMs). Finally, the score for each class were calculated by means of the log-likelihood. The proposed model was evaluated in three corpora: Neurovoz [35], PC-GITA [36], and CzechPD [37]. Each participant performed different speech tasks such as diadochokinetic (DDK) repetition, a monologue, read sentences, and picture description. Finally, the results of a k-fold cross-validation provided accuracies between 81% to 94% depending on the corpus, while cross-corpora trials yielded accuracies between 66% to 76%. In [38], the authors performed an analysis of the relevance of the glottal source parametrization method for the evaluation of the presence of dysphonia in newly diagnosed (non-medicated) PD patients. For this study, the sharp vowel /ah/ of 40 PD patients and 40 HC subjects (all men) was considered. The speech signals were characterized with glottal, perturbation, and cepstral-based features. Each set was classified with a Support Vector Machine (SVM) classifier, and the authors reported an Area Under the ROC Curve (AUC) of 0.78 when the three sets were combined using an early fusion strategy. In [39], four speech dimensions (articulation, phonation, prosody, and intelligibility) were evaluated to predict dysarthria level using regression models. The authors consider a dataset with PD patients and HC subjects, and reported a Spearman's correlation of up to 0.64 between the predicted and real dysarthria severities. In [40], the articulation and phonation dimensions in vowels of 50 PD patients were evaluated with respect to two groups of controls: elderly healthy speakers and young healthy speakers. The authors performed a multi-class classification using an SVM and a Neural Network. The results indicated that the phonation and articulation capabilities of young speakers clearly differ from those exhibited by the elderly speakers (with and without PD). In [41], the authors proposed the classification of 188 patients with PD vs. 64 HC subjects. Different features were extracted, including: MFCCs, phonation features, wavelet features, and tunable Q-factor wavelet transform. Four classifiers were implemented and the authors reported accuracies

of up to 94.7% after selecting the 20 most discriminating using a wrapper algorithm.

Finally, few papers have performed and compared PD and HD perhaps due to the difficulty of collecting both databases. One of the first studies of this nature was proposed in [42], where the authors assessed syntactic changes in 9 PD and 10 HD patients taking as reference a set of 17 HC subjects. Each participant was asked to talk about themselves. Different linguistic features were calculated related to syntactic changes and cognition in each participant. The results showed that HD patients produced shorter expressions and a lower proportion of grammatical expressions than HC subjects. While PD patients did not differ significantly from HC subjects. In [43], the authors analyzed dysfunctions in the basal ganglia of PD and HD patients using data from 37 PD patients, 37 HD patients, and 37 HC subjects. The participants pronounced sustained phonations of the vowel /i/ and a read text in Czech language. The authors performed an acoustic analysis based on changes in a 1/3 octave band centered in 1kHz to reflect the nasality level, and pronunciation errors. 78% of HD patients showed a high incidence of intermittent hypernasality, while for PD patients there was an appearance of moderate mild hypernasality of 65%. Finally in [44], an automated method for the analysis of vocal tremor in multiple neurological diseases was designed. The authors included 240 participants divided in 9 groups of pathologies among which there were 40 PD patients and 20 HD patients. Each participant performed 2 repetitions of the vowels /ah/ and /i/. The results showed that the disease with the most abnormal tremor was HD (65%), followed by essential tremor (50%). The incidence for vocal tremor in PD patients was 20%.

1.3.2 Assessment of neurodegenerative diseases with deep learning methods

In recent years, different DL methods for speech analysis have been successfully implemented, the scientific community has worked on pathological speech analysis, for diagnosis and monitoring of patients. For example in [45], the authors modeled the articulatory deficits in PD patients with Convolutional Neural Networks (CNNs). Initially, the transitions between voiced and unvoiced segments were detected to model difficulties of patients to start/stop the vibration of the vocal folds. Then, a time-frequency representation for each transition was computed to train the CNNs. The authors

considered speech recordings to classify PD patients and HC subjects in 3 different languages (Spanish, German, and Czech), they obtained accuracies ranging from 70% to 89%, depending on the language. In [46], the authors proposed to evaluate the effectiveness of CNNs to discriminate between PD and HC using spectral features extracted from sustained vowels. A total of 81 subjects participated in the study (41 PD and 40 HC). The subjects were asked to produce the sustained vowel /ah/ for 5 seconds. Then, a transformation of each audio into a spectrogram was performed, using 25ms hamming windows with a time-shift of 15ms. Besides, the feature extraction procedure, data augmentation technique was applied upon frames, it mainly consisted in flipping (vertical and horizontal) and rotation of frames. Finally, the authors reported an accuracy of 75.7%. An additional DL model was proposed in [47] to classify patients with dysarthria and HC speakers, the architecture consisted in a combination of convolutional and recurrent layers trained with a multitask learning strategy to address two tasks: (1) to detect the presence of dysarthria, and (2) to reconstruct the Mel spectrogram from the input. The authors reported a recall of up to 0.93 for the dysarthria detection task. In [48], the authors proposed a classification of PD patients vs. HC subjects from DNNs and Long Short-Term Memory (LSTM). They used the Parkinson Speech dataset (PSD) containing recordings from 20 PD patients and 20 HC subjects [49]. The authors characterized each voice sample from 26 descriptors based on frequency, amplitude, harmonicity, pitch, pulse and voicing. Finally, they trained and evaluated different DNNs and LSTMs separately, the results obtained show an overall accuracy of 94% using DNNs and 98% using LSTMs. It should be clarified that the results may be optimistic and biased because the authors did not use a validation strategy in the experimentation. A work was introduced in [50] with a novel strategy based on unsupervised representation learning for automatic detection of pathological speech, specifically, for the classification of PD and cleft lip and palate. The proposed strategy is based on the use of recurrent and convolutional autoencoders trained to extract informative features to characterize the presence of pathologies in speech. In addition, a novel feature set based on the reconstruction error of the autoencoders is also proposed. In this study, the authors were able to classify the PC-GITA database with an accuracy of up to 84%. In [51], the authors proposed to apply a bidirectional-LSTM model to capture time-series dynamic features of speech signals for detecting PD, the dynamic features are measured based on computing the energy content

in the transition between voiced and unvoiced segments. A database with 30 PD patients and 15 HC subjects was considered, for each subject the sustained vowel /ah/ and a short sentence were recorded. The experimental results show that the proposed method remarkably improves the accuracy of PD detection over traditional ML models using static features, with an accuracy of up to 84%. In [52], a DNN model was proposed using the reduced input feature space of Parkinson's telemonitoring dataset to predict PD progression. The proposed DNN model analyzes the prediction performance of it in PD progression by predicting Motor and Total-Unified Parkinson's Disease Rating Scale (UPDRS) scores [53]. The database under consideration contains utterances of 42 PD patients recorded in different periods during 6 months. Each utterance contains age, gender, test time, Motor-UPDRS, Total-UPDRS, and 16 biomedical voice measures. The results of the proposed methodology show a root mean squared error of 2.2 and a coefficient of determination of 0.956. Finally, in [54] a pilot experiment to detect the presence of dysarthria in speech and detect the level of severity based on a DL approach for PD patients was implemented. Automated feature extraction and classification using CNNs showed 77.48% accuracy on test samples of the TORGO database [55] using spectrograms as input.

Due to the difficulty in finding a large database of speech disorders to train DL methods, the community has worked on TL techniques for the implementation of robust systems in one scenario and then to transfer knowledge to a specific application. For example, in [30], different experiments were performed for the detection of PD using the sustained vowel /ah/ and an architecture of CNNs called ResNet [56]. To train the network, spectrograms of the speech recordings were calculated and used as an input image to a pre-trained network with the Imagenet corpus [57] and with the Saarbrücken Voice Database (SVD) [58]. The PC-GITA database was used to evaluate the system. The accuracy obtained in the validation set was of 91%. In other experiment [29], the authors proposed to use a TL strategy among languages to discriminate between PD patients and HC subjects. The authors considered recordings in 3 different languages (Spanish, German, and Czech). Base models were created for each language and afterwards the parameters were transferred to the other languages. According to their results, the TL strategy improves the accuracy by up to 8%. In [31], the authors performed the classification of Amyotrophic Lateral Sclerosis (ALS), PD and HC using a CNN-LSTM. They used the National Institute of Mental Health

and Neurosciences (NIMHANS) database[1] that containing recordings of 60 ALS patients, 60 PD patients and 60 HC subjects. Each participant performed different tasks including spontaneous speech, DDKs, and sustained phoneme. Different experiments were proposed: (1) a bi-class classification of each disease, (2) a tri-class classification including HC subjects, and (3) a fine-tuning from the tri-class model to each bi-class model. The authors reported an accuracy of 96.9% for ALS vs. HC, 88.5% for PD vs. HC, and 85.2% for ALS vs. PD vs. HC. In addition, the fine-tuning experiments achieved an absolute improvement of 2.0% compared to the randomly initialized networks. In [59], the authors explored DNN and CNN models for the automatic prediction of the UPDRS-III scores. The architectures were previously trained using voice signals from patients of databases of organic and functional voice pathologies (SVD). Then, they adjusted each model by fine-tuning to predict the 3-stage split UPDRS-III score from the PC-GITA and Neurovoz databases. The authors were able to report accuracies of up to 70% in the three-class classification. A CNN for automated PD identification based on biomarkers-derived voice signals was developed in [60]. In this study, three pre-trained architectures (trained with ImageNet database) including ResNet101 [56], DenseNet161 [61], and SqueezeNet1_1 [62] were re-trained with the mPower database[2] for discriminate PD patients vs. HC subjects. They only considered a sustained voicing of "aaaah" and calculated the discrete cosine transformation to obtain a spectrum and feed the CNN. The performance results revealed that the proposed model could successfully identify the PD with an accuracy of 89.75% for DenseNet-161 architecture identified as the most suitable fine-tuning architecture. In [63], the authors proposed the classification and monitoring of PD patients based on sparse kernel TL combined with simultaneous sample and feature selection. The main idea of this methodology was to use sparse kernel TL to extract the effective structural information of PD speech features from the public TIMIT database. As target databases, the authors used the Sakar database that contains 20 PD patients and 20 HC subjects. The second database was collected by the same authors and it is used for the classification of patients before and after treatment (10 and 21 patients, respectively). The authors compared the proposed transfer learning method with different state-of-the-art methodologies and concluded that the proposed algorithm achieves improve-

[1] http://nimhans.ac.in/about-us
[2] https://www.synapse.org/mPower

ments in classification accuracy. Finally, in [64] a speech recognition model for PD patients was proposed to evaluate speech quality and help the patient with rehabilitation therapy. In this work, an LSTM was pre-trained with the LibriSpeech database, and then TL was performed on a Parkinson's database collected by the same authors extracted from Youtube videos (10 patients with PD). In this work, 3 TL strategies were evaluated: 1) fine-tuned all layers; 2) restoring and retraining certain layers of the pre-trained model; 3) discarding and reinitializing certain layers of the pre-trained model. Fine-tuning all layers method has shown to be the most effective method, where it has effectively reduced the WER by 12% on original speech dataset and 13% on denoised speech dataset.

1.4 Objectives

1.4.1 General objective

To implement and evaluate TL strategies for the detection of pathological speech in patients with Parkinson's and Huntington's disease, using base models of emotions and pathologies in other languages.

1.4.2 Specific objectives

1. To train and evaluate DL architectures able to detect and evaluate paralinguistic traits in speech using traditional learning techniques.

2. To implement and evaluate TL strategies able to transfer information from one speech pathology to a different one.

3. To evaluate the suitability of applying TL methods to transfer information from models to recognize emotions from speech into models to classify speech impairments.

4. To compare and validate the performance of the proposed transfer learning architectures with respect to traditional methods used to classify patients with PD and HD from HCs.

1.5 Contribution of this study

According to the state of the art, the scientific community has started to explore different TL methods for the assessment of pathological speech. Therefore, this research work is mainly focused on the use of different TL techniques for the assessment of PD and HD from different model bases. In order to contribute to this aim, the following are the main outcomes of this research work.

1. A transfer knowledge approach from classical ML methods for the classification of PD and HD patients. For this case, language and pathology crosses are performed using the maximum a posteriori adaptation technique in a Gaussian mixture model - universal background model [65].

2. A TL strategy for the classification of PD and HD patients including cross-pathology and cross-language, as well as trained models for emotion recognition.

3. A Novel TL strategy is proposed from sequentially frozen layers for the classification of PD patients vs. HC subjects in models trained for different languages.

4. A novel DL strategy based on embeddings is proposed to know which additional demographic information a CNN learns in pathological speech training from models trained with a language and several languages.

1.6 Outline

This research work is divided into five chapters. Chapter 1 contains the context, the research problem, the objectives and the contribution of the study. Chapter 2 includes the methods used in this work. Chapter 3 contains detailed information on the different databases. Chapter 4 presents details of the experiments and their results. Finally, Chapter 5 includes the main conclusions, discussion and future work.

Chapter 2

Theoretical background

This chapter shows the theoretical basis of each method and is divided into 4 main sections: (1) handcrafted feature extraction and time-frequency representation of speech signals, (2) pattern recognition methods, (3) deep learning methods, and (4) transfer learning techniques.

2.1 Speech dimensions

In this work, different features are computed to model 3 speech dimensions: articulation, phonation, and prosody. The features are extracted using Disvoice[1], a Python implemented toolbox to characterize pathological speech. The basis for most of the methods implemented in this toolbox are presented in [21]. Each dimension is explained in detail in the following subsections.

2.1.1 Articulation

This speech dimension is responsible for modeling aspects related to the modification of position, stress, and shape of various limbs and muscles involved in the speech production process. In this work, the articulatory capacity of each participant is measured from the transitions between unvoiced to voiced segments (onset) as it was originally introduced in [66]. A voiced segment is characterized as a quasi-periodic signal due to cyclic vibrations of the vocal folds. While in the production of an unvoiced segment there is no vibration of the vocal folds. Therefore, to detect each transition, the change between

[1] https://disvoice.readthedocs.io/en/latest/

a voiced and an unvoiced segment is detected based on the presence of Fundamental Frequency (F0). Once the borders are detected, we take 80ms of the signal to the left and to the right, forming segments with 160ms length.

The characterization of each transition is performed using 22 critical bands according to the Bark scale. The Bark scale is motivated by the concept of critical bands, which refers to frequency regions that reflect the excitation of the basilar membrane when it is stimulated by specific frequencies. In addition, 12 MFCCs and their first two derivatives are also calculated per segment to obtain a smooth representation of the voice spectrum in the transitions. MFCCs are coefficients for speech representation based on human auditory perception [39].

2.1.2 Phonation

Phonation is used to model abnormal patterns in vocal fold vibration, which are among the main problems that reduce intelligibility in patients with pathological speech. In this work, 7 phonation features were calculated. Each feature is extracted in voiced segments. Initially, the first and second derivatives of F0 are calculated, which are associated with the period of vibration of the vocal folds. Jitter and shimmer describe temporal perturbations in frequency and amplitude, respectively. Amplitude Perturbation Quotient (APQ) measures the long-term variability of the peak-to-peak amplitude of the speech signal, and Pitch Perturbation Quotient (PPQ) measures the long-term variability of the F0. Finally, the logarithm of the energy contour of each frame is also calculated [21], [40].

2.1.3 Prosody

Prosody analysis allows modeling changes in the intonation, timing, and loudness in speech. In PD patients, prosody helps in modeling speech deficits such as monotonicity and monoloudness, and non-motor symptoms such as depression, and anxiety. A large number of features are extracted to model prosody in pathological speech and can be divided into 3 parts: (1) features based on the F0, (2) features based on the short-term energy, and (3) features based on the duration. For this work, each voiced segment was modeled using coefficients of a 5-degree Lagrange polynomial for both the F0 and energy contours. In addition, the duration of each voiced segment was also included [21], [67].

2.2 Time-frequency representations

Short-Time Fourier Transform (STFT) is a Fourier-related transform used to determine the harmonic and phase frequency content in local sections of a signal. The STFT calculation consists of collecting a certain number of samples through a time window, then the frequency content (Ω) of the samples in the window is found, and the result is stored in a two-dimensional (time-frequency) matrix.

For audio signals, the information is divided into frames (which usually overlap each other to reduce border irregularities) and a Fourier transform from Equation 2.1 is performed on each frame.

$$X_m(\Omega) = \sum_{n=-\infty}^{\infty} x(n) f(n-m) e^{-j\Omega n} \tag{2.1}$$

Where $x(n)$ is the original signal and $f(n-m)$ is the window function centered at time m. In the STFT, the window establishes the degree of time and frequency resolution. If the length of the window is small, a short portion of the signal will be analyzed, which allows to have good time resolution but a bad frequency resolution. On the other hand, if the window length is large, you will have a high frequency resolution but low time resolution. The most commonly used windows to perform this type of analysis are the Blackman, Hamming, Hanning [68].

Finally, the magnitude response of the STFT is known as a spectrogram, which is a matrix that shows the variation of the spectrum and the signal energy in each frame. It is common to represent the signal energy in decibels (dB) as shown in Equation 2.2.

$$\mathbf{STFT}\{x(t)\}_{dB} = 10 \log_{10} |X_m(\Omega)| \tag{2.2}$$

Commonly, to analyze speech signals, a transformation to Mel scale frequencies is performed, this new representation aims to describe the human auditory system in a linear way. The filter bank used has a higher resolution at low frequencies and with a higher concentration around the area perceived by the human ear, while at high frequencies the information is not that relevant [69].

Figure 2.1 shows an onset transition from a 54 years old HC (left) and a 48 years old PD patient (right) with a MDS-UPDRS-III score of 9. The

red line indicates the contour of the F0. For the case of the HC subject, it is possible to identify the onset transition in time 80 milliseconds, because it is a minimum oscillation before the beginning of the voiced segment. For the case of the patients, it is not possible to see the beginning of the voiced segment due to large oscillations in the speech signal, this can be verified by the red line showing F0 contour in the unvoiced segment of the audio signal.

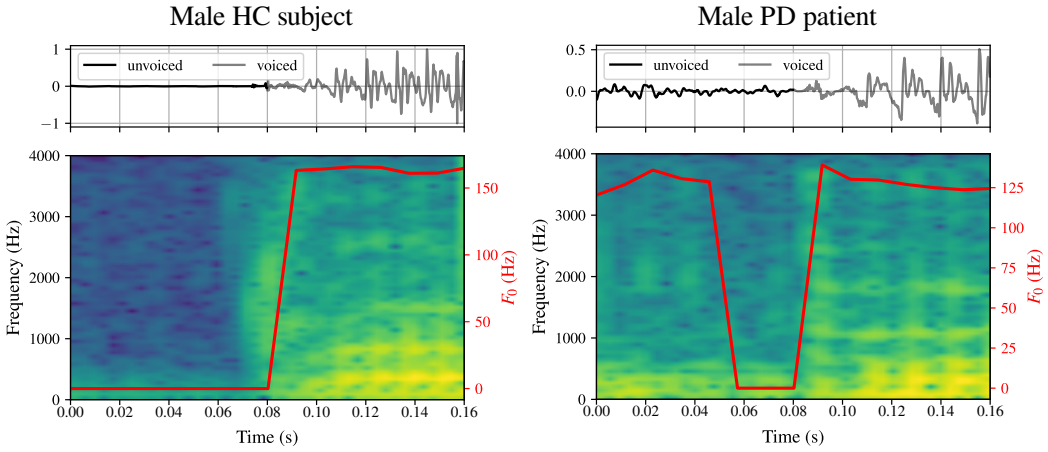

Figure 2.1. Spectrogram of an onset transition for a HC subject (left) and a PD patient (right).

2.3 Pattern recognition methods

2.3.1 Gaussian Mixture Models (GMM)

GMMs are probability models representing a population from a combination of Gaussian probability distributions. For a D-dimensional feature vector $\boldsymbol{x}$, the mixture density used for the likelihood function for N Gaussian is defined as:

$$p(\boldsymbol{x}|\lambda) = \sum_{i=1}^{N} w_i p_i(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{2.3}$$

Where $p_i(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ corresponds to a Gaussian density weighted by w_i that satisfies the constraint $\sum_{i=1}^{N} w_i = 1$. In addition, each p_i distribution is composed of a D $\times$ 1 mean vector $\boldsymbol{\mu}_i$, and a D $\times$ D covariance matrix $\boldsymbol{\Sigma}_i$ (see Equation 2.4). Collectively, the parameters of the density model are denoted as $\lambda = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ where $i = 1, ..., N$.

$$p_i(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{D/2}\,|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\intercal}(\boldsymbol{\Sigma}_i^{-1})(\boldsymbol{x}-\boldsymbol{\mu}_i)\right] \tag{2.4}$$

The parameters λ of the maximum likelihood function can be estimated using the Expectation Maximization (EM) algorithm [70]. This algorithm iteratively re-defines the λ parameters and increases the likelihood of the estimated model for the observed feature vectors; that is, for iterations k and $k+1$, $p(\boldsymbol{X}|\lambda^{(k+1)}) > p(\boldsymbol{X}|\lambda^{(k)})$, where $\boldsymbol{X}$ is an matrix containing the group of features $\boldsymbol{x}$ extracted from each participant in the database [65].

Gaussian mixture model - Universal Background Model (GMM-UBM)

A GMM trained with a large number of data is called a Universal Background Model (UBM). It is commonly defined as a reference and potentially represents the entire space of possible alternatives of the phenomenon of interest. The GMM-UBM paradigm was proposed in [71] and is a technique that has been explored mainly in speaker verification [65]. In a GMM-UBM system, the modeling of the parameters of each speaker is derived from the adaptation of a UBM using a process denoted Maximum A Posteriori (MAP) [72]. Unlike using only GMM and the EM algorithm, the main idea in MAP adaptation is to derive parameter updates from the UBM which is considered a robust and well-trained system. This provides a closer coupling between each model and the UBM. The process for the MAP adaptation is divided into 2 steps: (1) the probability that a feature vector belongs to each Gaussian of the UBM is estimated. (2) Refers to the adaptation of the new parameters taking into account both the probability of belonging obtained in step 1 and the parameters of previous iterations from an adaptation coefficient [65].

Given a UBM and a matrix $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_T\}$ that contains T feature vectors, we first determine the probability that the feature vector belongs to the i-th Gaussian component. This probability is shown in Equation 2.5.

$$Pr(i|\boldsymbol{x}_t) = \frac{w_i p_i(\boldsymbol{x}_t)}{\sum_{j=1}^{N} w_j p_j(\boldsymbol{x}_t)} \tag{2.5}$$

Then, $Pr(i|\boldsymbol{x}_t)$ and $\boldsymbol{x}_t$ are used to calculate the statistics denoted by n_i, $E_i(\boldsymbol{x})$, and $E_i(\boldsymbol{x}^2)$ that allow us to find the parameters λ.

$$n_i = \sum_{t=1}^{T} Pr(i|\boldsymbol{x}_t) \tag{2.6}$$

$$E_i(\boldsymbol{x}) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|\boldsymbol{x}_t)\boldsymbol{x}_t \tag{2.7}$$

$$E_i(\boldsymbol{x}^2) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|\boldsymbol{x}_t)\mathrm{diag}(\boldsymbol{x}_t\boldsymbol{x}_t^{\intercal}) \tag{2.8}$$

Finally, from the calculated statistics the parameters w_i', $\boldsymbol{\mu}_i'$, and $\boldsymbol{\Sigma}_i'$ are updated for a Gaussian mixture i using the following equations:

$$w_i' = [\alpha_i n_i/T + (1-\alpha_i)w_i]\,\gamma \tag{2.9}$$

$$\boldsymbol{\mu}_i' = \alpha_i E_i(\boldsymbol{x}) + (1-\alpha_i)\boldsymbol{\mu}_i \tag{2.10}$$

$$\boldsymbol{\Sigma}_i'^2 = \alpha_i E_i(\boldsymbol{x}^2) + (1-\alpha_i)(\boldsymbol{\Sigma}_i^2 + \boldsymbol{\mu}_i^2) - \boldsymbol{\mu}_i'^2 \tag{2.11}$$

Where γ is a scale factor that guarantees that $\sum_{i=1}^{N} w_i' = 1$. Also, α_i is known as the adaptive coefficient that controls the balance between the old and new parameters. The calculation of this coefficient is shown in Equation 2.12 where r is a relevance factor that has been defined in the literature as a standard term equivalent to 16 [65].

$$\alpha_i = \frac{n_i}{n_i + r} \tag{2.12}$$

GMM supervectors

The use of a supervector was proposed in [73] and it is a representation for each participant after the MAP adaptation of the GMM-UBM. Supervectors offer the solution to change from a frame wise feature matrix to a global static representation of each utterance. A GMM supervector is created by concatenating the vector of means $\boldsymbol{\mu}_i'$ and the diagonal of the covariance matrix $\boldsymbol{\Sigma}_i'$.

2.3.2 Support Vector Machine (SVM)

This is a learning algorithm that allows to discriminate M samples from a separation hyperplane that maximizes the margin between classes (see Figure 2.2). When the samples are linearly separable and an optimal hyperplane can be found, the classifier is known as Hard-Margin. In this case, all the points will satisfy the constraints.

$$\begin{aligned} \boldsymbol{w}^\intercal \boldsymbol{x}_i + w_0 \geq +1 \quad \text{for} \quad y_i = +1 \\ \boldsymbol{w}^\intercal \boldsymbol{x}_i + w_0 \leq -1 \quad \text{for} \quad y_i = -1 \end{aligned} \tag{2.13}$$

where $y_i \in \{-1, +1\}$ are the class labels, $\boldsymbol{x}_i$ is the feature vector, $\boldsymbol{w}$ is the vector normal to the hyperplane, w_0 is the bias parameter, and M is the number of samples.

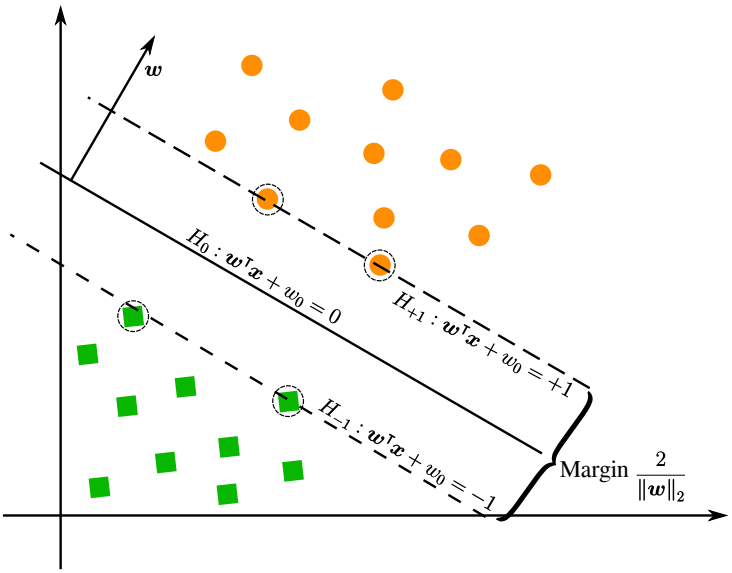

Figure 2.2. Linear separating hyperplane for separable data.

Note that the margin in Figure 2.2 is defined as $\frac{2}{\|\boldsymbol{w}\|_2}$, this is the perpendicular distance from the origin to the hyperplane of the points, for the case of $y_i = +1$ is $\frac{(1-w_0)}{\|\boldsymbol{w}\|_2}$ and to the hyperplane of the points with $y_i = -1$ is $\frac{(-1-w_0)}{\|\boldsymbol{w}\|_2}$.

Therefore, the SVM maximizing margin is the main idea of this method. Commonly, this maximization is reformulated as the following optimization problem.

$$\begin{aligned} &\underset{\boldsymbol{w},w_0}{\text{minimize}} && \tfrac{1}{2}\left\|\boldsymbol{w}\right\|_2^2 \\ &\text{subject to} && y_i(\boldsymbol{w}^\intercal \boldsymbol{x}_i + w_0) \geq 0 \quad i = 1, 2, \cdots M. \end{aligned} \tag{2.14}$$

The points that fulfill $y_i(\boldsymbol{w}^\intercal \boldsymbol{x}_i + w_0) = 0$ are in charge of constructing the decision margin and are known as support vectors. In the Figure 2.2 they are represented by a circle around each sample.

A standard approach to solve optimization problems with equality and inequality constraints is the Lagrange formalism [74] which leads to the primal form of the objective function, L_p.

$$L_p = \frac{1}{2}\boldsymbol{w}^\intercal \boldsymbol{w} - \sum_{i=1}^{M} \alpha_i (y_i(\boldsymbol{w}^\intercal \boldsymbol{x}_i + w_0) - 1) \tag{2.15}$$

where $\{\alpha_i, i = 1, 2, \cdots M; \alpha_i \geq 0\}$ are the Lagrange multipliers and the primal parameters are $\boldsymbol{w}$ and w_0.

The solution to the optimization problem shown in Equation 2.14 is equivalent to determining the saddle-point of the function L_p, at which L_p is minimized with respect to $\boldsymbol{w}$ and w_0 and maximized with respect to the α_i. Differentiating L_p with respect to $\boldsymbol{w}$ and w_0 and equating to zero yields.

$$\sum_{i=1}^{M} \alpha_i y_i = 0 \tag{2.16}$$

$$\boldsymbol{w} = \sum_{i=1}^{M} \alpha_i y_i \boldsymbol{x}_i \tag{2.17}$$

Substituting into (2.15) gives the dual form of the Lagrangian

$$L_D = \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^\intercal \boldsymbol{x}_j \tag{2.18}$$

Since this optimization problem is convex and the constrains are affine (linear), the Slater's condition holds, thus the duality gap is zero, that is, it is possible to formulate the Lagrangian dual problem L_D, and the optimal parameters in L_D will also be optimal for L_p. The Kuhn–Tucker conditions provide necessary and sufficient conditions to be satisfied when minimizing an objective function subject to inequality and equality constraints. In the primal form of the objective function, these are

$$\frac{\partial L_p}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{M} \alpha_i y_i \boldsymbol{x}_i = 0 \tag{2.19}$$

$$\frac{\partial L_p}{\partial w_0} = \sum_{i=1}^{M} \alpha_i y_i = 0 \tag{2.20}$$

$$y_i(\boldsymbol{x}_i^\intercal \boldsymbol{w} + w_0) - 1 \geq 0 \tag{2.21}$$

$$\alpha_i \geq 0 \tag{2.22}$$

$$\alpha_i(y_i(\boldsymbol{x}_i^\intercal \boldsymbol{w} + w_0) - 1) = 0 \tag{2.23}$$

In particular, the 2.23 condition known as the complimentary slackness condition is of great importance, since if $\alpha_i \geq 0$, then $y_i(\boldsymbol{x}_i^\intercal \boldsymbol{w} + w_0) - 1 = 0$; therefore, all the vectors x_i associated $\alpha_i \geq 0$ are elements of the boundary hyperplane and comprise the set of support vectors.

In many real-world practical problems there will be no linear boundary separating the classes and the problem of searching for an optimal separating hyperplane is meaningless. Therefore, it is necessary to include slack variables $\{\zeta_i, i = 1, 2, \cdots M\}$, into the constraints to give

$$\begin{array}{lll} \boldsymbol{w}^\intercal \boldsymbol{x}_i + w_0 \geq +1 - \zeta_i & \text{for} & y_i = +1 \\ \boldsymbol{w}^\intercal \boldsymbol{x}_i + w_0 \leq -1 + \zeta_i & \text{for} & y_i = -1 \\ \zeta_i \geq 0 & & i = 1, \cdots, M \end{array} \tag{2.24}$$

For a point to be misclassified by the separating hyperplane, we must have $\zeta_i > 1$(see Figure 2.3). Therefore, by introducing this variable in the objective function, the optimization problem is as follows

$$\underset{\boldsymbol{w}, w_0, \zeta_i}{\text{minimize}} \ \frac{1}{2} \|\boldsymbol{w}\|_2^2 + C \sum_{i=1}^{M} \zeta_i \tag{2.25}$$

where the term $C \sum_{i=1}^{M} \zeta_i$ can be thought of as measuring some amount of misclassification. Then, the primal form of the Lagrangian now becomes

$$L_p = \frac{1}{2} \boldsymbol{w}^\intercal \boldsymbol{w} + C \sum_{i=1}^{M} \zeta_i - \sum_{i=1}^{M} \alpha_i (y_i(\boldsymbol{w}^\intercal \boldsymbol{x}_i + w_0) - 1 + \zeta_i) - \sum_{i=1}^{M} r_i \zeta_i \tag{2.26}$$

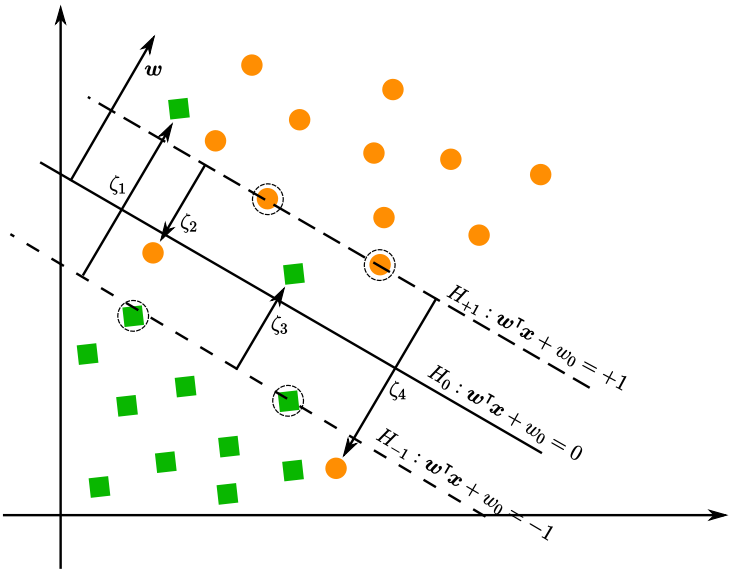

Figure 2.3. Linear separating hyperplane for non-separable data.

where $\alpha_i \geq 0$ and $r_i \geq 0$ are Lagrange multipliers, and r_i are introduced to ensure positivity of ζ_i. Now, Differentiating L_p with respect to $\boldsymbol{w}$ and w_0 and equating to zero yields the same equations for 2.16 and 2.17 were obtained; and differentiating with respect to ζ_i yields

$$C - \alpha_i - r_i = 0 \tag{2.27}$$

Substituting the above results. the dual form of the Lagrangian is obtained which is the same form as the maximal margin classifier (see Equation 2.18). This is maximized with respect to the α_i subject to

$$\sum_{i=1}^{M} \alpha_i y_i = 0 \qquad 0 \leq \alpha_i \leq C \tag{2.28}$$

The latter condition follows from (2.27) and $r_i \geq 0$. Thus, the only change to the maximization problem is the upper bound on α_i. Therefore, the Karush–Kuhn–Tucker conditions are

$$y_i(\boldsymbol{x}_i^\intercal \boldsymbol{w} + w_0) - 1 + \zeta_i \geq 0 \tag{2.29}$$

$$\zeta_i \geq 0 \tag{2.30}$$

$$\alpha_i(y_i(\boldsymbol{x}_i^\intercal \boldsymbol{w} + w_0) - 1 + \zeta_i) = 0 \tag{2.31}$$

$$r_i\zeta_i = (C - \alpha_i)\zeta_i = 0 \tag{2.32}$$

$$\alpha_i \geq 0; \qquad r_i \geq 0 \tag{2.33}$$

In this case, the border hyperplane is defined by the Equation 2.31 and is satisfied when $\alpha_i \geq 0$. The samples that satisfying $0 < \alpha_i < C$ must have $\zeta_i = 0$ and these support vectors are sometimes termed margin vectors.

2.4 Deep learning methods

Deep learning is a form of ML that enables computers to learn from experience and understand the world in terms of a hierarchy of concepts. This hierarchy of concepts allows the computer to learn complicated concepts by building them out of simpler ones. This is developed from a neural network which is the basis of deep learning. A neural network is a mathematical model created to simulate the activity of the human brain. It consists of a large number of units called artificial neurons, which are connected to transmit information from the input layer and flows through the entire neural network. The information flow within the network is subject to different nonlinear operations until the output values are obtained in the last layer [26].

Neural network architectures are typically formed by a combination of simple modules also called abstraction layers. Each layer is composed of neurons that receive a set of values at the input and compute the predicted y value. Vector $\boldsymbol{x}$ contains the values of the features from the training set. In addition, each unit has its own set of parameters denoted as $\boldsymbol{w}$ (vector of weights) and b (bias) which changes during the learning process. In each iteration, the neuron calculates a weighted average of the values of the vector $\boldsymbol{x}$, based on its current weight vector $\boldsymbol{w}$ and adds bias b. Finally, the result of this calculation is passed through a non-linear activation function denoted by ϕ. This process is summarized in Figure 2.4 and Equation 2.34.

$$y = \phi(w_1x_1 + w_2x_2 + w_3x_3 + ... + w_nx_n + b) = \phi(\boldsymbol{w}^\intercal\boldsymbol{x} + b) \tag{2.34}$$

In a neural network, each layer transforms the input and increases the level of abstraction of the model, which helps to reduce possible bias of the output due to external factors. For example, in a speaker verification

system that considers recordings of telephone channels, each layer can extract appropriate features to identify the speaker from speech, while eliminates non-relevant effects such as those introduced by the telephone channel and the microphone. These models are called feed-forward networks, and their objective is to approximate the function shown in Equation 2.34, to obtain the value of the parameters $\boldsymbol{w}$ and b such that provide the best approximation to the objective function [75].

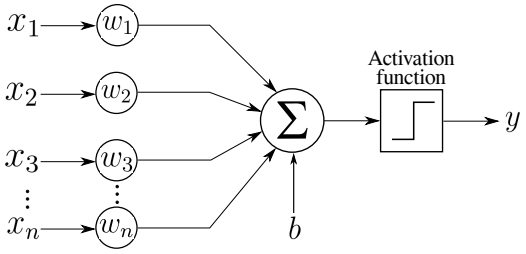

Figure 2.4. Single neuron.

2.4.1 Deep Neural Networks (DNNs)

DNNs are composed by 2 or more hidden layers that contain a set of linear and non-linear functions. The linear functions consist in the product of the weights with the input value plus the bias: $f(\boldsymbol{x}) = \boldsymbol{w}^\intercal \boldsymbol{x} + b$, and the non-linear functions $\phi(\boldsymbol{x})$ (also called activation functions) such as $sigmoid$, $tanh$, $softmax$, among others. The non-linear activations allows the system to "learn" more complex operations or functions [76]. The output of a DNN can be expressed according to Equation 2.35, where o represents the output layer and j is the number of layers and gives an idea of the depth of the model. Figure 2.5 shows the general scheme of a DNN with j layers of abstraction. Commonly, when a highly complex DNN is created, overfitting can occur, that is, the algorithm adapts perfectly to the training data and loses generalization capability to the actual problem. Overfitting is avoided with regularization strategies such as batch normalization [77] or dropout [78].

$$y = \phi_o(f_o(\phi_j(f_j \cdots (\phi_2(f_2(\phi_1(f_1(\boldsymbol{x}))))))))) \tag{2.35}$$

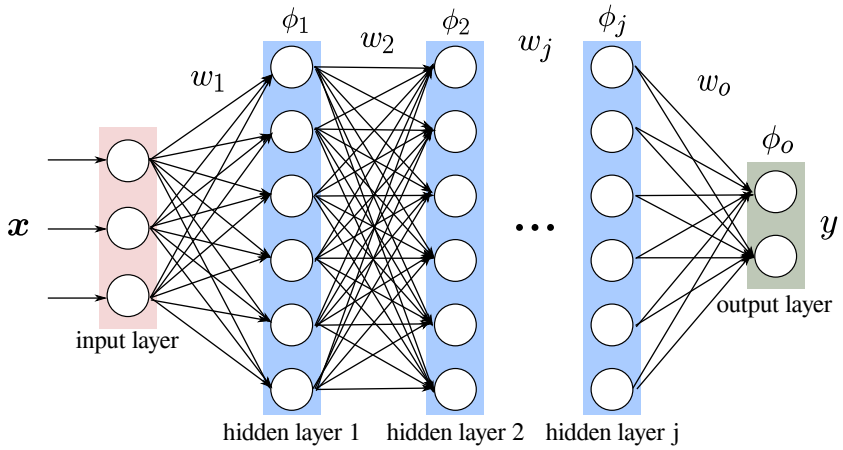

Figure 2.5. General scheme of a feed-forward DNN.

2.4.2 Convolutional Neural Networks (CNNs)

CNNs are another type of DL architecture. These are the state of the art in algorithms for object recognition in images. In addition, CNNs have been implemented to perform pathological speech classification [79], audio event detection [80] and speech recognition [81]. CNNs are designed to process data formed by multiple matrices, for example, a color image formed by three RGB (Red-Green-Blue) channels, or also two-dimensional matrices that correspond to time-frequency representations of audio signals. These networks contain a structure formed by convolutional filters and pooling layers, instead of the fully connected layers of a DNN (see Figure 2.6).

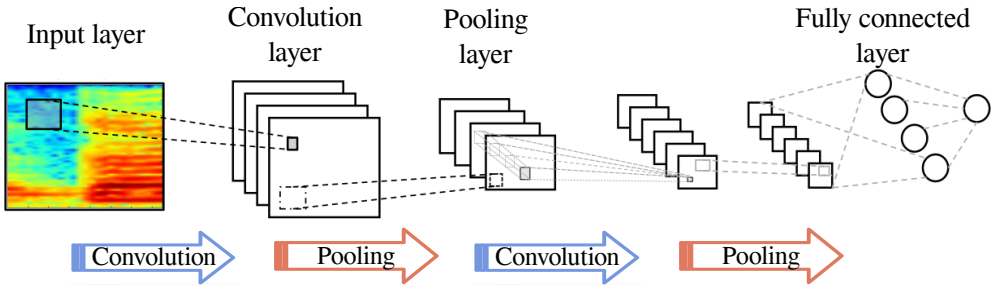

Figure 2.6. Typical structure of a CNN.

Convolutional layer

The input to a CNN can be a matrix or a tensor $\boldsymbol{X} \epsilon R^{\upsilon \times h \times c}$, where υ, h, c are, for instance the number of vertical and horizontal channels and the value of

an RGB image, or a time-frequency representation of a speech signal with c channels. A weight tensor $\boldsymbol{W} \epsilon R^{m \times m \times d}$ also called kernel, is convoluted with the input of each convolutional layer according to Equation 2.36, resulting in a hidden representation $\boldsymbol{H} \epsilon R^{v-m+1 \times h-m+1 \times d}$ of the extracted features, where m is the order of the convolutional filter and d is the number of hidden units in the layer, known as feature map.

$$\boldsymbol{H}(i,j,d) = \text{conv}(\boldsymbol{X}, \boldsymbol{W}_d)(i,j) = \sum_{l=1}^{m} \sum_{n=1}^{m} \boldsymbol{X}(i+l, j+n) \boldsymbol{W}_d(l,n) \quad (2.36)$$

In DL, discrete convolution can be viewed as matrix multiplication. The input matrix has several constraints because typically the kernel size is smaller than the input image (see Figure 2.7). This allows small pieces of information to be modeled consecutively and then combine the information in later layers of the network. One way to understand CNNs is that the first layer will try to detect borders and establish patterns of border detection. Then later layers will try to combine them into simpler shapes and finally into patterns of the different positions of objects, lighting, scales, among other things [26].

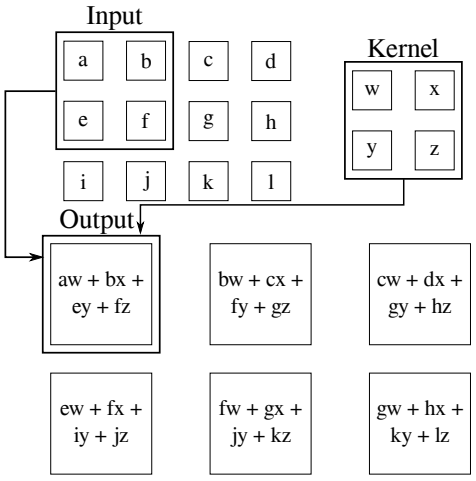

Figure 2.7. Example of a 2-D convolution. [a]

[a]Image taken from: I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT press, 2016.

Pooling layer

This layer is generally located after the convolutional one. Its main utility is reducing the spatial dimensions of the input layer from a statistical summary of the nearest outputs in the layer. The operation performed by this layer is also called under-sampling, because the size reduction results in loss of information. However, this type of loss can be beneficial to the network for two reasons: (1) the decrease in size leads to lower computational cost, (2) it reduces overfitting. One of the most common operations in this layer is called Max pooling (see Figure 2.8), a method that reports the maximum output value from its nearest neighbors in a rectangular array. There are other functions such as the average of a rectangular neighborhood, the L_2 norm, or a weighted average based on the distance from the central pixel [26], [82].

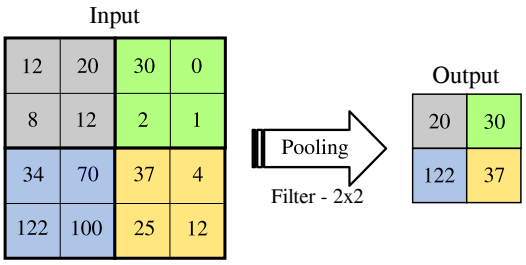

Figure 2.8. Max pooling method.

Fully connected layer

At the end of the convolution and pooling layers, networks generally use fully connected layers where each pixel is considered as a separate neuron as in a regular neural network. This last classification layer will have as many neurons as the number of classes to be predicted.

2.4.3 Residual networks

One of the topologies that have excelled in DL is the Residual Neural Network (ResNet). This topology aimed to solve the problem of vanishing gradient, where the error function is reduced exponentially through the propagation of the previous layers, making it difficult to train deeper networks [83]. ResNet allows continuous learning of the gradient by adding additional information

to the output of each block through the addition operation. The addition structure does not introduce additional parameters or computational complexity, allowing it to stand out from other architectures such as VGGNet and GoogLeNet, obtaining better performance at the same or lower computational cost [56].

A plain CNN topology has a nonlinear mapping function $y = F(x)$ within consecutive layers. The ResNet topology is composed of a series of layers and an identity mapping that adds a block input to the output, that is, instead of trying to learn from a direct mapping of x and with a function $H(x)$, ResNet defines a residual function $F(x) = H(x) - x$, which can be written $H(x) = F(x) + x$; where $F(x)$ represents the stacked layers of the neural network and x the identity function, which is carried through to the output of the block (see Figure 2.9).

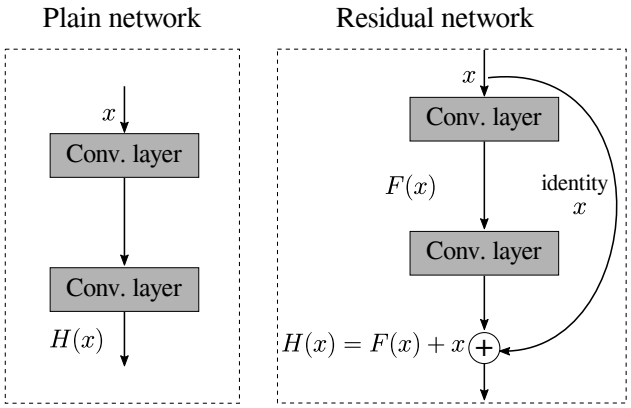

Figure 2.9. Identity mapping in residual blocks.

If the identity mapping is optimal, the architecture can easily take the residuals to zero ($F(x) = 0$) instead of adjusting an identity mapping (x, input = output) through the stack of network layers. In other words, it is easier to find a solution like $F(x) = 0$ instead of $F(x) = x$ using the stack of nonlinear layers of the CNN as function [56].

2.4.4 Training

Cost function

The cost function determines the error between the estimated value and the real labeled prediction. While minimizing the error, the corresponding parameters of the neural network will be the optimal ones. In many DNN

applications, a common cost function is the cross-entropy, where each probability estimated by the neural network is compared with the real label, then a score is calculated such that penalizes the probability as a function of distance from the expected value. Equation 2.37 shows this cost function, where H is the size of training data, $\boldsymbol{y_i}$ is the vector of real labels and $\boldsymbol{p_i}$ is the neural network predictions, mostly interpreted as probabilities. Cross-entropy is typically used for classification problems because the prediction converges quickly and more robustly, although the problem of the vanish gradient is common [84].

$$J(\boldsymbol{w}, b) = \frac{1}{H} \sum_{i=1}^{H} [\boldsymbol{y_i} \log(\boldsymbol{p_i}) + (1 - \boldsymbol{y_i}) \log(1 - \boldsymbol{p_i})] \tag{2.37}$$

Stochastic Gradient Descent (SGD)

The typical method to optimize the parameters of the DNN is the Stochastic Gradient Descent (SGD), which allows to minimize or maximize some function $f(x)$ when advancing iteratively on the functions, using its gradient with respect to x. Finally, the method stops at a local or global minimum.

The insight of SGD is that the gradient is an expectation. The expectation may be approximately estimated using a small set of samples. Specifically, on each iteration, a minibatch of examples $\mathbb{R} = \left\{x^{(1)}, x^{(2)}, ..., x^{(m')}\right\}$ sampled uniformly from the training set can be sampled. The minibatch size m' is typically chosen to be a relatively small number of examples, ranging from one to a few hundred [26]. The gradient estimation is expressed by Equation 2.38.

$$\boldsymbol{g} = \frac{1}{m'} \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{m'} J(\boldsymbol{\theta}) \tag{2.38}$$

where J can be any loss function (such as Cross-entropy) and $\boldsymbol{\theta}$ the network parameters. Finally, the SGD algorithm then follows the estimated gradient downhill:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \boldsymbol{g} \tag{2.39}$$

where η is the learning rate and indicates the size of the step in each iteration. In practice, it is necessary to gradually decrease the learning rate over time; therefore, we now denote the learning rate at iteration k as η_k.

This is because the SGD gradient estimator introduces a source of noise (the random sampling of m training examples) that does not vanish even when a minimum is arrived. In practice, it is common to decay the learning rate linearly until iteration τ:

$$\eta_k = (1 - \alpha)\eta_0 + \alpha\eta_\tau \tag{2.40}$$

with $\alpha = \frac{k}{\tau}$. After iteration τ, it is common to leave η constant. Considering that the parameters η_0, η_τ and τ may be chosen by trial and error [80].

Back-propagation algorithm

Back-propagation is the method by which a neural network adjusts its parameters to "learn" an internal representation of the information. In other words, it is the algorithm that calculates the direction of maximum variation in each layer and then updates the weights by means of the descending gradient.

We start from Equation 2.41, where there is a sum of the weights multiplied by the input plus a bias, all this is denoted as $\boldsymbol{z}$, which is then passed through an activation function ϕ, and finally this result is taken to the cost function J. In addition, L indicates the number of layers in the network.

$$J[\phi(\boldsymbol{z}_L)] = J[\phi(\boldsymbol{w}_L\boldsymbol{x} + b_L)] \tag{2.41}$$

From this composition of functions the derivative of the previous equation is made with respect to $\boldsymbol{w}_L$ through the chain rule, in order to find the direction of maximum variation for the layer L.

$$\frac{\partial J}{\partial \boldsymbol{w}_L} = \frac{\partial J}{\partial \phi_L} \cdot \frac{\partial \phi_L}{\partial \boldsymbol{z}_L} \cdot \frac{\partial \boldsymbol{z}_L}{\partial \boldsymbol{w}_L} \tag{2.42}$$

The first two terms of Equation 2.42 refer to the error in the cost function when there is a change in the sum of the neurons, this definition is known as the imputed error of the neurons and is denoted as $\boldsymbol{\delta}_L$. While the last term of Equation 2.42 is the representation of how $\boldsymbol{z}_L$ changes with respect to the weights and taking into account that the inputs to this layer are the outputs of the previous layer, we denote as ϕ_{L-1} the output of the layer $(L-1)$. Therefore, the change of the cost function with respect to the weights can be denoted as:

$$\frac{\partial J}{\partial \boldsymbol{w}_L} = \boldsymbol{\delta}_L \cdot \phi_{L-1} \tag{2.43}$$

Now continuing with the algorithm and performing the same analysis for the previous layers, we have to:

$$\frac{\partial J}{\partial \boldsymbol{w}_{L-1}} = \boldsymbol{\delta}_L \cdot \boldsymbol{w}_L \cdot \frac{\partial \phi_{L-1}}{\partial \boldsymbol{z}_{L-1}} \cdot \phi_{L-2} \tag{2.44}$$

In summary, the back-propagation algorithm works following 3 steps deduced from the previous explanation, starting from the last layer and performing the same process sequentially until the initial layer:

1. Calculate the error of the last layer (L):

$$\boldsymbol{\delta}_L = \frac{\partial J}{\partial \phi_L} \cdot \frac{\partial \phi_L}{\partial \boldsymbol{z}_L} \tag{2.45}$$

2. Propagate error to previous layer $(L-1)$:

$$\boldsymbol{\delta}_{L-1} = \boldsymbol{w}_L \cdot \boldsymbol{\delta}_L \cdot \frac{\partial \phi_{L-1}}{\partial \boldsymbol{z}_{L-1}} \tag{2.46}$$

3. Calculate the derivatives of the layer using the error:

$$\frac{\partial J}{\partial \boldsymbol{w}_{L-1}} = \boldsymbol{\delta}_{L-1} \cdot \phi_{L-2} \tag{2.47}$$

2.5 Transfer Learning (TL)

Conventional ML and DL algorithms, have been traditionally designed to work in isolation. These algorithms are trained to solve specific tasks. The models have to be rebuilt from scratch once the feature-space distribution changes. TL is the idea of overcoming the isolated learning paradigm and utilizing knowledge acquired for one task to solve related ones (see Figure 2.10). TL can use the knowledge (features, weights, and others) of previously created models to train new ones and address problems with small amounts of data [26].

Mathematically, a domain $\mathfrak{D} = \{\chi, P(X)\}$ consists of two components: a feature space χ and a marginal probability distribution $P(X)$, where $X =$

$\{x_1, x_2, \cdots, x_n\} \in \chi$. In addition, a task $\Gamma = \{\Upsilon, f(\cdot)\}$, where Υ is a label space and $f(\cdot)$ an objective predictive function. Then, given a source domain $\mathfrak{D}_S = \left\{(x_{S_1}, y_{S_1}), \cdots, (x_{S_{n_S}}, y_{S_{n_S}})\right\}$ where $x_{S_i} \in \chi_S$ is the data instance and $y_{S_i} \in \Upsilon_S$ is the corresponding class label. Similarly, the target domain data was denoted a $\mathfrak{D}_T = \left\{(x_{T_1}, y_{T_1}), \cdots, (x_{T_{n_T}}, y_{T_{n_T}})\right\}$, where the input $x_{T_i} \in \chi_T$ and $y_{T_i} \in \Upsilon_T$ is the corresponding output. In most cases, it is important to mention that $0 \leq n_T \ll n_S$. Therefore, TL aims to help improve the learning of the target predictive function $f_T(\cdot)$ in $\mathfrak{D}_T$ using the knowledge in $\mathfrak{D}_S$ and Υ_T, where $\mathfrak{D}_S \neq \mathfrak{D}_T$, or $\Upsilon_S \neq \Upsilon_T$ [27], [85]. Given the last 2 conditions, 4 situations may occur:

1. $\chi_S \neq \chi_T$. The feature spaces of the source and target domain are different. A example in the context of Natural Language Processing (NLP) is generally referred to as cross-lingual adaptation.

2. $P(X_S) \neq P(X_T)$. The marginal probability distributions of source and target domain are different. Following with the previous example, the languages are the same but the topics are different, this scenario is generally known as domain adaptation.

3. $\Upsilon_S \neq \Upsilon_T$. The label spaces between the two tasks are different. For example, the documents need to be assigned different labels in the target task.

4. $P(Y_S|X_S) \neq P(Y_T|X_T)$. The conditional probability distributions of the source and target tasks are different. It is common in unbalanced labels in the source and target domains.

Now, in order to provide a solution to each of the scenarios described above, it is necessary to divide TL into 3 different types:

1. Inductive transfer learning: In this scenario, the target task is different from the source task, regardless of whether the source and target domains are the same or not. The main objective is to achieve high performance in the target task by transferring knowledge from the source task. In addition, the data in the base domain must be labeled, while in the target domain some or all of the data should be labeled.

2. Transductive transfer learning: In this approach, the source and target tasks are the same, while the source and target domains are different.

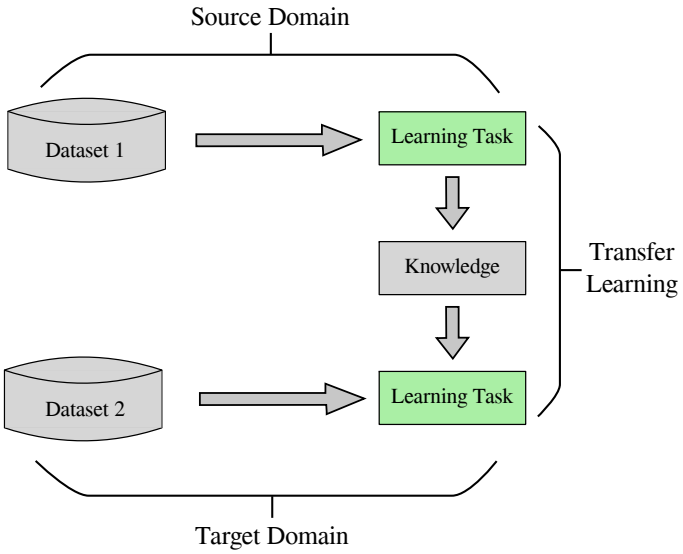

Figure 2.10. Learning process of TL.

The difference is that no labeled data is available in the target domain, while a lot of labeled data is available in the source domain.

3. Unsupervised transfer learning: This configuration is similar to inductive TL, but has the main difference that in this scenario, the labeled data is not available in any of the domains.

In this work, we focus on analyzing and using different strategies in inductive TL, since both in our base domain and target domain there are the labels, and our main focus is on helping to improve the discrimination of a target pathology from a different base pathology. Now, in order to perform the implementation of this type of TL there is a category denoted parameter-based approaches where TL tries to transfer knowledge through the shared parameters of the source and target domain learner model (see Figure 2.11). From this category, 3 strategies are derived that allow the development different experiments to provide a solution to the problem proposed in this work: pre-trained models as feature extractors, fine-tuning and pre-trained models.

Pre-trained models as feature extractors

DL systems and models are layered architectures that learn different features at different layers (hierarchical representations of layered features). These layers are typically connected at the end using a fully connected layer, in the

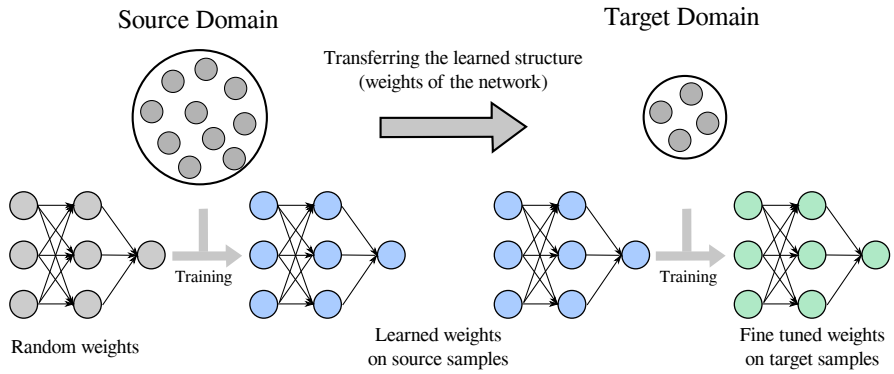

Figure 2.11. Parameter-based TL methods.

case of classification to get the final output. This type of architecture allows to use a pre-trained model without its final layer as a fixed feature extractor for other tasks. The main idea of this TL strategy is to take advantage of the characterization layers of a previously trained model, and then use these characteristics as an embedding to classify the new phenomenon.

Fine-tuning

Unlike the previous strategy, fine-tuning not only replaces the final layer (for classification/regression), but also selectively modifies some previous layers. It has been shown that in DNNs the initial layers capture generic characteristics, while the later layers are more focused on the specific task. Therefore, it is possible to freeze certain layers while they are being retrained, or tune the rest of them to fit our needs. In this case, the knowledge in terms of the overall architecture of the network is utilized and its states as the starting point for our retraining step are used. In addition, this helps in achieving better performance with less training time.

Pre-trained models

This strategy is supported by different DL architectures that are openly shared by the scientific community. These extend across different domains mainly to computer vision, emotion recognition, and NLP. The pre-trained models are usually shared in the form of the millions of parameters/weights that the model achieved while being trained to a stable state like VGG16 [86], ResNet-50 [56], and Word2vec [87]. Therefore, each person can make an adaptation or simply evaluate in each architecture its target database.

Chapter 3

Data

Seven corpora were considered for our experiments: (1-3) three databases with recording from PD patients and HC subjects in three different languages (Spanish, Czech, German), (4) a corpus from HD patients and HC subjects in Czech language. (5) A corpus of patients affected by different larynx pathologies (LP) in English. (6) An additional database to evaluate depression in PD in Spanish. Finally, (7) the IEMOCAP corpus for emotion recognition from speech [88]. Details about each corpus are found in the following sections. Each recording of the 7 databases used was down-sampled to the minimum sampling frequency of 16 kHz.

3.1 PD-Spanish

PC-GITA is a database containing speech recordings of 50 PD patients and 50 HC subjects sampled at 44.1 kHz [36]. All participants are native Colombians, balanced in age and gender. Table 3.1 shows additional information on the speakers in this database. Participants performed different speaking tasks, such as rapid repetition of syllables (DKKs) /pa-ta-ka/, /pa-ka-ta/, /pe-ta-ka/, /pa/, /ta/, /ka/, 10 sentences, a monologue, a read text, isolated words, sustained vowels, and modulated vowels. All patients were in ON state at the time of the recording, i.e., under the effect of their daily medication and each one was evaluated by an expert neurologist and labeled according to the MDS-UPDRS-III score.

Table 3.1. General information of the subjects in the PC-GITA database. [F/M]: Female/Male. Time since diagnosis and age are given in years. Values are reported in terms of mean ± standard deviation.

	PD patients	**HC subjects**	**Patients vs. controls**
Gender [F/M]	25/25	25/25	$^{*}p$=1.00
Age [F/M]	60.7±7/61.3±11	61.4±7/60.5±12	$^{**}p$=0.98
Range of age [F/M]	49-75/33-81	49-76/31-86	
Time since diagnosis [F/M]	12.6±12/8.7±6		
MDS-UPDRS-III [F/M]	37.6±14/37.8±22		
Speech item (MDS-UPDRS-III) [F/M]	1.3±0.8/1.4±0.9		

$^{*}p$–value calculated through Chi–square test.
$^{**}p$–value calculated through t-test.

3.2 PD-Czech

This corpus consisted of 50 PD patients and 49 HC captured at a sampling frequency of 48 kHz [89]. A total of 4 speech tasks are included: a monologue, a read text, a DDK (/pa-ta-ka/), and sustained vowels. All PD patients were recruited at their first visit to the clinic and were examined before symptomatic treatment to start. Disease duration was based upon patients' self-report [38]. Table 3.2 shows demographic and clinical information about the participants in this database.

Table 3.2. General information on speakers of the PD-Czech database. [F/M]: Female/Male. Time since diagnosis and age are given in years. Values are reported in terms of mean ± standard deviation.

	PD patients	**HC subjects**	**Patients vs. controls**
Gender [F/M]	20/30	19/30	$^{*}p$=0.94
Age [F/M]	60.1±9/65.3±10	63.5±11/60.3±12	$^{**}p$=0.37
Range of age [F/M]	41-72/43-82	40-79/41-77	
Time since diagnosis [F/M]	6.8±5/6.7±5		
MDS-UPDRS-III [F/M]	18.1±10/21.4±12		
Speech item (MDS-UPDRS-III) [F/M]	0.7±0.6/0.9±0.5		

$^{*}p$–value calculated through Chi–square test.
$^{**}p$–value calculated through t-test.

3.3 PD-German

This corpus consists of 176 native German speakers, divided into 88 PD patients and 88 HC subjects [90]. Details of the database can be found in Table 3.3. The German voice recordings were captured at a sample frequency

of 16 kHz. Each participant performed different speech tasks including: a monologue, a read text, reading of question-answer-pairs, 5 sentences, 7 isolated words, rapid repetition of 8 DDKs, and sustained vowels. The patients were recorded as well in ON state, i.e., under the effect of their daily medication.

Table 3.3. General information on speakers of the PD-German database. [F/M]: Female/Male. Time since diagnosis and age are given in years. Values are reported in terms of mean ± standard deviation.

	PD patients	**HC subjects**	**Patients vs. controls**
Gender [F/M]	41/47	44/44	$^{*}p$=0.76
Age [F/M]	66.2 ±9.7/66.7±8.7	62.6±15.2/63.8±12.7	$^{**}p$=0.07
Range of age [F/M]	42-84/44-82	28-85/26-83	
Time since diagnosis [F/M]	7.1±6/7.0±6		
MDS-UPDRS-III [F/M]	23.3±12/22.1±10		
Speech item (MDS-UPDRS-III) [F/M]	1.2±0.5/1.4±0.6		

$^{*}p$–value calculated through Chi–square test.
$^{**}p$–value calculated through t-test.

3.4 HD-Czech

This corpus is formed with 40 Czech native patients with genetically verified HD and 40 HC subjects balanced in age, with no history of neurological or speech disorders. 22 of the patients were treated with anti-psychotic medication [91]. Each participant was evaluated according to the Unified Movement Disorder Society, Hungtinton's disease rating scale (MDS-UHDRS) [92]. The speech tasks performed by the participants include the rapid repetition of the syllables /pa-ta-ka/, a read text, and a monologue, similar to the PD-Czech data. Table 3.4 shows demographic and clinical information about the participants in this database.

Table 3.4. General information on speakers of the HD-Czech database. [F/M]: Female/Male. Time since diagnosis and age are given in years. Values are reported in terms of mean ± standard deviation.

	HD patients	**HC subjects**	**Patients vs. controls**
Gender [F/M]	20/20	20/20	$^{*}p$=1.00
Age [F/M]	49.5±14.1/47.7±12.2	50.1±13.9/48.3±12.3	$^{**}p$=0.84
Range of age [F/M]	27-69/23-67	27-69/26-70	
MDS-UHDRS [F/M]	27.1±10.7/26.8±12.7		
Speech item (MDS-UHDRS) [F/M]	0.7±0.5/0.9±0.3		

$^{*}p$–value calculated through Chi–square test.
$^{**}p$–value calculated through t-test.

Figure 3.1 summarizes the distributions of the normalized severity scores (MDS-UPDRS-III and MDS-UHDRS) for the PD and HD corpora. The scores were standardized by the maximum value of the scale in order to have comparable scores between PD and HD. Note that subjects from the PD-Spanish (gray) and HD-Czech (purple) exhibit higher neurological impairments than patients from the PD-German (dark green) and PD-Czech (blue) corpora.

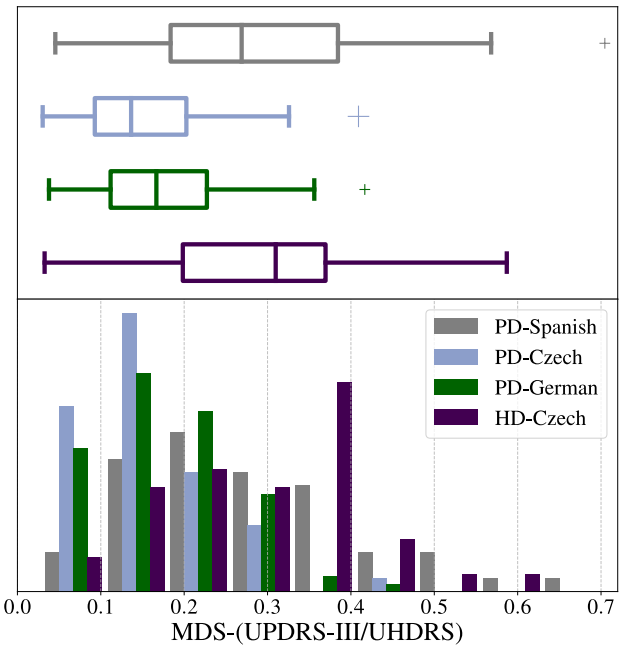

Figure 3.1. Distributions of the MDS-(UPDRS-III/UHDRS) score normalized for the PD and HD databases.

3.5 LP-English

We consider data from the Kay-elemetrics dataset [93], which contains utterances from 53 HC subjects and 175 pathological talkers who had a variety of organic, neurological, traumatic, and psychogenic voice disorders. The speech tasks performed by the participants include the vowel /ah/ and *the northwind and the sun* passage. Each of the patients was evaluated by a speech-language pathologist, the normal talkers exhibited no abnormal vocal characteristics and indicated no history of voice disorders.

3.6 Depression in PD

This corpus is composed of 60 native Colombian patients with PD. 25 patients were labeled as depressive and the remaining 35 as non-depressive, this label was set based on the depression item of the MDS-UPDRS-I, where a score equal to zero is equivalent to non-depressed patients and an item being higher than zero is equivalent to a depressive patient. Table 3.5 shows additional information on the speakers in this database. Each patient performed the same tasks described in Section 3.1 for the Spanish Parkinson's database.

Table 3.5. General information of the subjects in the Depression in Parkinson's Disease database. **D-PD patients**: Depressive Parkinson's disease patients. **ND-PD patients**: Non-Depressive Parkinson's disease patients. [F/M]: Female/Male. Time since diagnosis and age are given in years. Values are reported in terms of mean ± standard deviation.

	D-PD patients	**ND-PD patients**	**D-PD vs. ND-PD**
Gender [F/M]	16/9	17/18	$^{*}p$=0.36
Age [F/M]	66.2±10.3/69.0±10.2	60.3±13.6/66.3±10.6	$^{**}p$=0.21
Range of age [F/M]	55-87/48-86	29-80/53-90	
Time since diagnosis [F/M]	6.1±5.5/8.8±4.5	13.6±13.1/7.3±4.9	$^{**}p$=0.14
MDS-UPDRS-III [F/M]	32.6±19.1/36.9±17.8	26.5±10.6/36.3±17.1	$^{**}p$=0.56
Depression item (MDS-UPDRS-I) [F/M]	1.4±0.7/1.3±0.5		

$^{*}p$–value calculated through Chi–square test.
$^{**}p$–value calculated through t-test.

3.7 Interactive Emotional Dyadic Motion Capture (IEMOCAP)

IEMOCAP is an interactive emotional database collected by the University of Southern California [88]. This dataset was collected following theatrical theory in order to simulate natural dyadic interactions between actors. It contains approximately twelve hours of audio-visual data from five mixed gender pairs of actors. There are five sessions, each recorded session lasts approximately five minutes and consists of two actors interacting with each other. It is composed of an improvised dataset on topics that can elicit specific emotions and a scripted dataset. For each utterance in the dataset, each evaluator judged anger, sadness, happiness, disgust, fear, surprise, frustration, excitement, and contentment, and then labeled the utterance with the most selected emotion. A limitation of the speech dataset is that the distribution of these instances across each emotion classes is heavily unbalanced.

Therefore, we incorporated the excited turns into the happy class, resulting in four emotion categories for training and evaluation: anger, happiness, sadness, and neutral (see Figure 3.2).

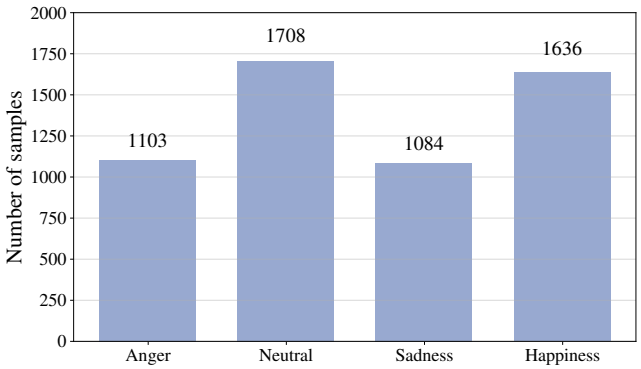

Figure 3.2. Class distribution of the speech samples in IEMOCAP dataset.

Chapter 4

Experiments and results

This chapter is distributed into five sections: (1) transfer knowledge using GMM-UBM for the classification of neurodegenerative diseases, (2) TL between languages and pathologies using CNNs, (3) TL in PD based on emotions, (4) freezing layers in CNNs to classify PD from speech, and (5) additional information acquired in a neural network trained for pathological speech classification.

4.1 Transfer knowledge using GMM-UBM for PD and HD classification of neurodegenerative diseases

Few works have explored the discrimination between pathologies using voice signals. This is due to the difficulty of collecting voice recordings from patients with different pathologies. In addition, due to differences in the acoustic conditions of each corpus, it is not possible to perform discrimination of these without having a bias of microphones and/or acoustic conditions. Therefore, the discrimination of two pathologies was proposed, in databases collected in the same acoustic conditions and in the same language. For this purpose, classification and comparison of hypokinetic dysarthria characteristic of PD and hyperkinetic dysarthria characteristic of HD was performed. Parkinson's and Huntington's databases in Czech language were classified using supervectors extracted from a GMM-UBM trained with Spanish and German databases collected for PD classification. In addition, different speech dimensions were used in order to model and compare the dysarthria characteristic of each disease and try to associate it with each speech dimension.

4.1.1 Methodology

The methodology addressed in this experiment consists of five main stages (see Figure 4.1): (A) the computation of articulation, phonation, and prosody features for each database. (B) The training of the UBM using the PD-Spanish and PD-German corpora. The UBMs are trained using articulation, phonation, and prosody-based features (described in section 2.1). (C) The adaptation of each speaker from the Czech corpora (PD and HD) using MAP to derive a specific GMM per subject. (D) The creation of GMM-supervectors for each subject in the target databases using the mean and covariances of the adapted GMM. (E) The supervectors are used to classify the subjects from the Czech corpora using an SVM classifier. Classification of PD and HD patients vs. their respective HC subjects is performed, the classification of both groups of patients, i.e., PD vs. HD, and the three-class problem discriminating PD vs. HD vs. HC. For each experiment, two speech tasks are considered, including the rapid repetition of the syllables /pa-ta-ka/ and a monologue.

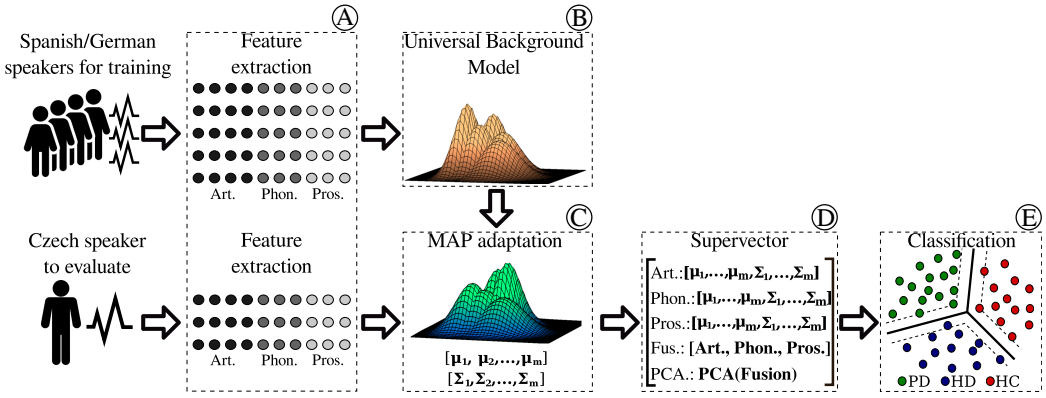

Figure 4.1. General methodology. **Art.**: Articulation, **Phon.**: Phonation. **Pros.**: Prosody. **Fus.**: Early fusion of supervectors from articulation, phonation, and prosody. **PCA.**: Principal component analysis.

4.1.2 Experiments and results

The number of Gaussian components M for the UBM is optimized up to multipliers of 2^n such that $M \in \{2, 4, 8, 16, 32, 64\}$ and the selection criterion is based on the accuracy in the training set. All experiments were validated using a stratified k-fold cross-validation strategy with 10 folds, this process was repeated 10 times for a better generalization of the results. In the classification stage, the parameters were optimized based on a grid-search where $C \in \{0.001, 0.005, 0.01, \cdots, 100, 500, 1000\}$ and $\gamma \in \{0.0001, 0.001, \cdots, 1000\}$. Finally, the optimal hyper-parameters obtained in the training process are computed from the mode over the repetitions of the process. The performance of each experiment was evaluated using different performance measures such as accuracy, sensitivity, specificity, AUC, F1 score, and Kappa coefficient [94].

For the speaker adaptation, UBMs with the data for Spanish, German, and their combination were created; besides, for each language (including their combination) the training of the UBM was performed using only HC subjects and the combination of PD patients and HC subjects. It is important to mention that training the UBM with only PD patients does not make sense, because it is the most difficult population to record, therefore it was not considered. Finally, for each UBM, the articulation, phonation, and prosody supervectors were obtained and evaluated. In addition, the early fusion of the 3 speech dimensions and the dimension reduction using Principal Component Analysis (PCA) of the fusion were considered.

Classification of PD vs. HC

Table 4.1 shows the accuracy obtained in the classification of the Parkinson's database in Czech (PD vs. HC) for each speech dimension, their early fusion, and their dimension reduction using PCA. It is possible to observe that in general, the best results range between 62-70%, obtaining comparable results with the state of the art for this same database [38] and considering that it is a database with patients in early stages of the disease (see Figure 3.1). In addition, it is important to mention that the dimension that allows better discrimination between PD patients and HC subjects is articulation with an accuracy of 66.3% in the case of monologue, and phonation with an accuracy of 66.4% for the /pa-ta-ka/ task.

Table 4.1. Accuracies obtained in the classification of PD patients vs. HC subjects for each speech dimension. **Ger.**: German, **Spa.**: Spanish, **Acc.**: Accuracy, **Art.**: Articulation, **Phon.**: Phonation, **Pros.**: Prosody, **Fus.**: Fusion. * UBM trained with HC subjects. ** UBM trained with HC subjects and PD patients. Values are reported in terms of mean ± standard deviation.

UBM	Monologue					/pa-ta-ka/				
	Acc. Art.	Acc. Phon.	Acc. Pros.	Acc. Fus.	Acc. PCA	Acc. Art.	Acc. Phon.	Acc. Pros.	Acc. Fus.	Acc. PCA
Ger.*	68.9±2.0	64.8±4.3	63.0±1.9	67.9±2.8	46.1±4.0	53.2±3.5	62.0±4.0	56.2±3.9	58.4±2.0	52.0±2.4
Spa.*	65.7±4.4	67.0±2.4	62.2±2.8	65.2±2.0	55.7±3.2	61.9±2.6	68.6±3.1	58.4±5.5	61.8±3.5	55.1±2.7
Ger.-Spa.*	65.7±2.6	63.3±2.7	62.9±2.1	61.6±2.8	44.6±3.1	54.4±4.3	68.2±2.7	59.4±3.5	58.9±4.1	48.2±2.3
Ger.**	**69.5±2.7**	65.6±1.8	62.3±2.6	67.5±3.3	47.0±5.2	52.8±3.3	63.8±2.4	55.5±4.0	57.9±2.7	55.8±3.0
Spa.**	65.7±2.3	66.8±2.7	62.0±2.4	59.4±3.4	43.4±4.3	60.1±3.9	65.7±3.1	58.0±2.4	57.2±4.7	66.9±2.8
Ger.-Spa**	66.1±3.4	62.6±2.0	62.8±1.8	65.1±3.4	42.4±3.6	56.5±2.7	**69.8±3.1**	60.3±4.2	57.8±3.6	57.8±2.2
Average	66.3±2.9	65.0±2.7	62.5±2.3	64.5±3.0	46.5±3.9	56.5±3.4	66.4±3.1	58.0±3.9	58.7±3.4	56.0±2.6

A more detailed result of the classification of PD patients vs. HC subjects is shown in Table 4.2, where it is possible to observe different performance measures such as accuracy, sensitivity, specificity, and AUC of the best results observed for each UBM in Table 4.1. The best result in this classification is obtained from the supervector built with the phonation features and adapted from the UBM trained with the combination of the Spanish and German databases (including controls and patients), this approach obtained an accuracy of 69.8%, an AUC of 0.73, and a good balance between sensitivity (70.4%) and specificity (69.2%).

Figure 4.2 shows the histogram and the probability density distribution of each sample evaluated at the separation hyperplane (located at zero). The figure on the left corresponds to the result obtained with the monologue task and using the articulation supervector (Acc.: 69.5%) and the figure on the right to the best result with the /pa-ta-ka/ task (Acc.: 69.8%). In general,

Table 4.2. Details of performance measures of the best results obtained in the classification of PD patients vs. HC subjects. **Ger.**: German, **Spa.**: Spanish, **Art.**: Articulation, **Phon.**: Phonation. * UBM trained with HC subjects. ** UBM trained with HC subject and PD patients. Values are reported in terms of mean ± standard deviation.

UBM	Monologue					/pa-ta-ka/				
	Feat.	Accuracy	Sensitivity	Specificity	AUC	Feat.	Accuracy	Sensitivity	Specificity	AUC
Ger.*	Art.	68.9±2.0	67.4±2.7	70.4±3.0	0.71±0.02	Phon.	62.0±4.0	59.8±6.8	64.2±7.3	0.67±0.04
Spa.*	Phon.	67.0±2.4	65.4±3.8	68.6±2.5	0.71±0.08	Phon.	68.6±3.1	66.2±4.9	71.0±4.2	0.73±0.03
Ger.-Spa.*	Art.	65.7±2.6	63.6±2.2	67.8±4.2	0.69±0.03	Phon.	68.2±2.7	67.2±4.7	69.2±2.3	0.74±0.02
Ger.**	Art.	**69.5±2.7**	**66.2±5.2**	**72.8±2.4**	**0.74±0.03**	Phon.	63.8±2.4	62.6±4.0	65.0±4.4	0.69±0.02
Spa.**	Phon.	66.8±2.7	62.2±3.7	71.4±2.5	0.73±0.03	PCA	66.9±2.8	68.6±4.0	65.1±4.4	0.70±0.03
Ger.-Spa.**	Art.	66.1±3.4	62.4±6.3	69.8±3.4	0.70±0.03	Phon.	**69.8±3.1**	**70.4±4.1**	**69.2±4.2**	**0.73±0.03**

the errors are balanced, with a better classification of the HC for the case of the monologue (Spec.: 72.8%) and better classification of the PD patients for the /pa-ta-ka/ task (Sens.: 70.4%).

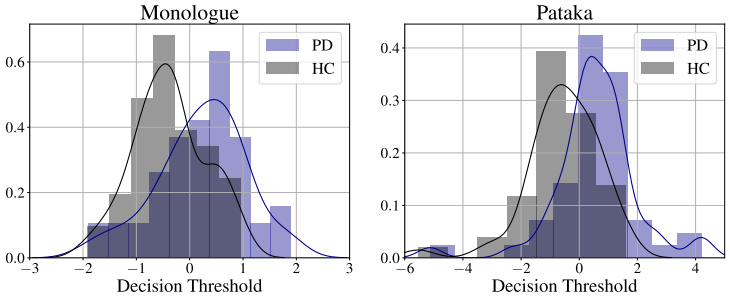

Figure 4.2. Histograms and the corresponding probability density distributions of the classification scores obtained from PD patients and HC subjects.

Classification of HD vs. HC

Table 4.3 shows the overall results for each speech dimension for the Huntington's database classification in Czech (HD vs. HC). In this case, the best results for both speech tasks are obtained by combining the 3 supervectors of the calculated dimensions. This suggests that these dimensions can be complementary and support the evidence in [39], [40] for the evaluation of PD patients, but never reporting on the HD. Considering each dimension individually, we can see that for the monologue task, phonation is the most discriminating dimension with an average accuracy of 81.5%. For the /pa-ta-ka/ task, prosody is the most accurate dimension (78.7%). These results are consistent with those found in [33], [95], where it is observed that the prosody and phonatory capacity of HD patients is considerably affected.

Table 4.3. Accuracies obtained in the classification of HD patients vs. HC subjects for each speech dimension. **Ger.**: German, **Spa.**: Spanish, **Acc.**: Accuracy, **Art.**: Articulation, **Phon.**: Phonation, **Pros.**: Prosody, **Fus.**: Fusion. * UBM trained with HC subjects. ** UBM trained with HC subjects and PD patients. Values are reported in terms of mean ± standard deviation.

UBM	Monologue					/pa-ta-ka/				
	Acc. Art.	Acc. Phon.	Acc. Pros.	Acc. Fus.	Acc. PCA	Acc. Art.	Acc. Phon.	Acc. Pros.	Acc. Fus.	Acc. PCA
Ger.*	76.3±3.5	81.9±2.0	81.0±1.2	85.8±2.0	64.9±3.6	69.1±5.2	77.4±2.1	77.9±2.9	81.6±2.6	64.5±2.9
Spa.*	74.6±2.7	82.6±2.8	77.9±1.9	**88.6±2.2**	49.8±3.6	70.4±3.6	77.9±2.3	79.1±2.5	82.1±2.8	65.4±1.5
Ger.-Spa.*	73.9±1.6	81.4±2.2	78.5±3.9	82.3±3.2	53.5±4.0	69.5±3.8	72.4±2.5	78.8±1.9	78.5±1.9	64.5±4.8
Ger.**	75.6±2.3	81.5±1.6	79.0±2.2	84.8±1.1	62.6±4.0	70.3±2.5	75.3±2.2	80.3±2.7	78.3±3.0	65.0±4.0
Spa.**	77.4±2.2	78.4±3.0	79.3±1.8	87.1±1.5	56.0±3.9	72.4±2.5	79.6±2.0	77.4±1.9	81.0±2.9	68.8±1.5
Ger.-Spa.**	78.6±2.5	83.4±2.5	80.9±1.9	84.6±2.1	63.9±3.9	71.4±3.5	78.4±1.8	78.6±3.0	**85.4±2.0**	68.9±3.8
Average	76.1±2.5	81.5±2.4	79.4±2.2	85.5±2.0	58.5±3.8	70.5±3.5	76.4±2.2	78.7±2.5	81.2±2.5	66.2±3.1

Table 4.4 shows additional performance measures of the best results obtained in the previous table. The best result has an accuracy of 88.6% and an AUC of 0.95. These results were obtained with the UBM built with the controls of the Spanish Parkinson's database and with the early fusion of the 3 extracted supervectors. It is important to highlight that in this experiment most of the results by the different UBMs were obtained by concatenating the 3 speech dimension supervectors. In addition, it is possible to observe that better discrimination is obtained in the continuous speech task than in the repetition task.

Table 4.4. Details of performance measures of the best results obtained in the classification of HD patients vs. HC subjects. **PD**: Parkinson's disease, **HC**: Healthy controls, **Fus.**: Fusion, **Pros.**: Prosody. * UBM trained with HC subjects. ** UBM trained with HC subjects and PD patients. Values are reported in terms of mean ± standard deviation.

UBM	Monologue					/pa-ta-ka/				
	Feat.	Accuracy	Sensitivity	Specificity	AUC	Feat.	Accuracy	Sensitivity	Specificity	AUC
Ger.*	Fus.	85.8±2.0	80.3±1.8	91.3±2.6	0.92±0.01	Fus.	81.6±2.6	80.8±5.3	82.5±3.4	0.90±0.01
Spa.*	Fus.	**88.6±2.2**	**85.3±2.1**	**92.0±2.7**	**0.95±0.01**	Fus.	82.1±2.8	85.0±3.5	79.3±5.1	0.90±0.02
Ger.-Spa.*	Fus.	82.3±3.2	83.0±1.9	81.5±5.1	0.93±0.02	Pros.	78.8±1.9	76.5±3.9	81.0±4.2	0.84±0.02
Ger.**	Fus.	84.8±1.1	83.4±2.0	86.0±1.2	0.93±0.02	Pros.	80.3±2.7	75.2±3.8	85.3±3.1	0.86±0.02
Spa.**	Fus.	87.1±1.5	85.4±1.0	88.7±2.6	0.95±0.01	Fus.	81.0±2.9	82.2±3.9	80.0±3.2	0.88±0.02
Ger.-Spa.**	Fus.	84.6±2.1	84.0±1.2	85.3±3.9	0.93±0.02	Fus.	**85.4±2.0**	**84.8±3.4**	**86.0±2.3**	**0.91±0.01**

Figure 4.3 shows the score distributions for the best results in the monologue (left) and /pa-ta-ka/ (right) tasks. In both figures, it is possible to observe that the distributions are separated. On the one hand, in the monologue figure, it is observed that the error for the classification of the HC is minimum (Spec.: 92%) while for the patients higher error is observed, classifying a small number of patients as controls with a high degree of certainty. On the other hand, in the /pa-ta-ka/ figure, it is possible to observe more

balanced distributions and with a similar error between patients and controls (Sens.: 84.8% and Spec.: 86.0%).

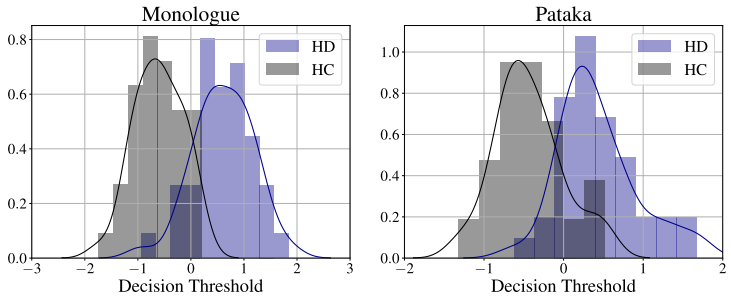

Figure 4.3. Histograms and the corresponding probability density distributions of the classification scores obtained from HD patients and HC subjects.

Classification of PD vs. HD

To guarantee that this experiment is not biased by differences in the acoustic conditions of the two databases in Czech language, a statistical test of Kruskal-Wallis with Bonferroni correction between HC was performed. For each dimension, we obtained results of $p \geq \alpha$, where α is the significance level after Bonferroni correction. This indicates that the difference between the population medians is not statistically significant; therefore, the databases can be merged. In addition, a classification of the HC speakers of both databases was performed and the results ranged from 50% to 55% accuracy.

Table 4.5 shows the overall results for the classification of PD vs. HD patients. In this case, performance ranges between 83-86% for the monologue task and between 75-81% in the /pa-ta-ka/ task, again continuous speech stands out for the discrimination of the patients. Moreover, in the monologue task, all results were obtained from phonation features with an average accuracy of 83.8%, supporting the result reported in [33] where the authors showed that phonatory features can detect different traits in the voice of HD patients. In addition, for the /pa-ta-ka/ task, the 3 dimensions were balanced and the best results were obtained with the early fusion of these dimensions (average accuracy of 77.7%). Although in the monologue the best dimension was phonation, the fusion of the features is only 1.7 percentage points below, which suggests that the fusion of the dimensions are complementary and fundamental to discriminate the different impairments caused by each of the diseases.

Table 4.5. Accuracies obtained in the classification of PD patients vs. HD patients for each speech dimension. **Ger.**: German, **Spa.**: Spanish, **Acc.**: Accuracy, **Art.**: Articulation, **Phon.**: Phonation, **Pros.**: Prosody, **Fus.**: Fusion. * UBM trained with HC subjects. ** UBM trained with HC subjects and PD patients. Values are reported in terms of mean ± standard deviation.

UBM	Monologue					/pa-ta-ka/				
	Acc. Art.	Acc. Phon.	Acc. Pros.	Acc. Fus.	Acc. PCA	Acc. Art.	Acc. Phon.	Acc. Pros.	Acc. Fus.	Acc. PCA
Ger.*	77.9±2.4	83.2±2.9	72.1±1.6	81.8±2.4	61.8±3.0	68.6±2.9	71.4±2.0	74.9±3.6	75.2±2.9	57.9±3.0
Spa.*	75.9±2.3	82.8±1.8	69.1±3.3	79.9±3.2	48.1±4.8	73.1±2.2	68.1±3.4	69.7±2.2	78.2±2.2	54.7±2.8
Ger.-Spa.*	75.7±2.5	83.1±1.6	71.1±2.5	81.2±3.0	56.3±3.4	68.8±4.0	67.1±2.8	68.4±1.8	77.7±2.8	52.0±3.3
Ger.**	74.8±3.0	**86.2±1.8**	73.9±2.0	84.1±1.7	65.2±1.8	66.6±2.3	69.3±4.3	68.1±3.7	75.2±1.5	47.2±3.9
Spa.**	75.2±1.9	83.7±2.1	69.1±2.0	83.2±2.2	67.9±2.9	70.6±2.7	73.2±2.0	67.0±3.4	**81.6±1.3**	66.7±2.2
Ger.-Spa.**	77.7±1.7	83.6±1.8	70.4±2.3	82.2±1.7	62.1±2.5	70.9±2.0	69.4±3.3	75.4±2.2	78.2±2.5	59.4±3.6
Average	76.2±2.3	83.8±2.0	71.0±2.3	82.1±2.4	60.2±3.1	69.8±2.7	69.8±3.0	70.6±2.8	77.7±2.2	56.3±3.1

A details of the best results can be seen in Table 4.6. The best result had an accuracy of 86.2% and its adaptation was obtained using a UBM trained with the full German database (including patients and controls). On the other hand, for the /pa-ta-ka/ task, the best result was obtained by concatenating the 3 supervectors and with the UBM trained with the entire Spanish database, with an accuracy of 81.6%.

Table 4.6. Details of performance measures of the best results obtained in the classification of PD patients vs. HD patients. **Ger.**: German, **Spa.**: Spanish, **Fus.**: Fusion, **Phon.**: Phonation. * UBM trained with HC subjects. ** UBM trained with HC subjects and PD patients. Values are reported in terms of mean ± standard deviation.

UBM	Monologue					/pa-ta-ka/				
	Feat.	Accuracy	Sensitivity	Specificity	AUC	Feat.	Accuracy	Sensitivity	Specificity	AUC
Ger.*	Phon.	83.2±2.9	86.6±2.4	80.0±4.3	0.90±0.02	Fus.	75.2±2.9	83.6±2.2	66.8±4.2	0.84±0.02
Spa.*	Phon.	82.8±1.8	84.0±2.4	81.5±2.5	0.89±0.01	Fus.	78.2±2.2	84.4±3.9	72.0±3.5	0.85±0.02
Ger.-Spa.*	Phon.	83.1±1.6	86.2±2.6	80.0±3.1	0.89±0.01	Fus.	77.7±2.8	82.4±2.3	73.0±4.4	0.85±0.02
Ger.**	Phon.	**86.2±1.8**	**88.6±1.8**	**83.7±2.8**	**0.93±0.01**	Fus.	75.2±1.5	84.4±2.3	66.0±4.1	0.84±0.01
Spa.**	Phon.	83.7±2.1	89.2±2.2	78.3±3.2	0.90±0.01	Fus.	**81.6±1.3**	**86.6±2.4**	**76.5±3.4**	**0.86±0.01**
Ger-Spa.**	Phon.	83.6±1.8	83.0±3.3	85.0±3.1	0.90±0.02	Fus.	78.2±2.5	70.5±3.8	86.2±3.1	0.86±0.01

Figure 4.4 shows the probability density distributions of the scores of the 2 results highlighted above. In the left figure it can be seen that the 2 distributions are clear and that the error for the two classes is similar, standing out the error of the HD patients classified as PD patients (Sens.: 88.6% and Spec.: 83.7%). The right figure shows a clear distribution for PD patients with a small error (Sens.: 81.6%), opposite case for HD patients since they tend to be confused with PD patients and have a higher error rate (Spec.: 76.5%).

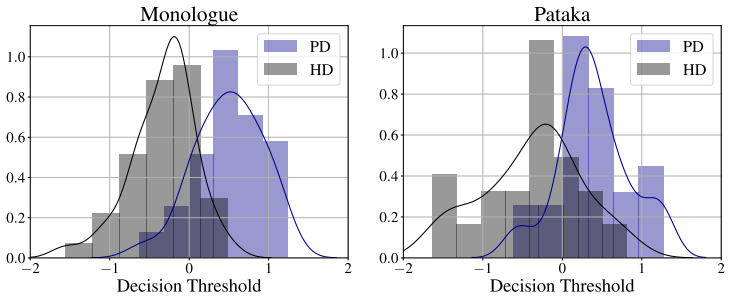

Figure 4.4. Histograms and the corresponding probability density distributions of the classification scores obtained from PD and HD patients.

Multi-class classification (PD vs. HD vs. HC)

A multi-class classification was performed and the results are presented in Table 4.7. For this experiment, HC subjects from the 2 Czech databases were combined and evaluated with respect to PD and HD patients using an SVM with a one-vs-rest decision. This table shows the balanced accuracy to compensate for the imbalance that exists between the classes (50PD, 40HD, and 90HC). In the results, it is possible to observe that again most of the best results were obtained using an early fusion of the 3 dimensions, 69% of average accuracy for the monologue task and 62.4% for /pa-ta-ka/ task. Therefore, it is possible to conclude again that these dimensions are complementary and the importance of including them to discriminate both pathologies in the healthy population.

Table 4.7. Accuracies obtained in the classification of PD patients vs. HD patients vs. HC subjects for each speech dimension. **Ger.**: German, **Spa.**: Spanish, **Acc.**: Accuracy, **Art.**: Articulation, **Phon.**: Phonation, **Pros.**: Prosody, **Fus.**: Fusion. * UBM trained with HC subjects. ** UBM trained with HC subjects and PD patients. Values are reported in terms of mean ± standard deviation.

UBM	Monologue					/pa-ta-ka/				
	Acc. Art.	Acc. Phon.	Acc. Pros.	Acc. Fus.	Acc. PCA	Acc. Art.	Acc. Phon.	Acc. Pros.	Acc. Fus.	Acc. PCA
Ger.*	67.1±1.5	63.7±2.6	60.9±1.8	**71.0±1.4**	48.2±2.3	52.5±2.2	61.9±2.2	54.3±1.8	61.0±2.1	41.5±1.4
Spa.*	65.0±2.4	62.8±2.2	61.2±3.2	67.9±1.6	35.1±2.3	56.7±1.3	58.7±1.0	59.4±2.5	64.0±1.6	43.3±1.5
Ger.-Spa.*	63.0±3.0	64.1±0.2	58.5±1.3	67.1±2.6	38.9±2.4	55.9±2.0	58.3±2.7	57.3±3.1	**64.6±1.5**	40.0±1.6
Ger.**	65.1±1.6	64.9±1.7	57.7±2.2	67.7±2.6	49.7±2.9	52.9±1.2	61.1±2.2	55.0±1.7	59.6±2.1	39.8±1.5
Spa.**	66.3±2.2	68.1±1.9	56.2±2.4	70.0±1.4	47.6±2.2	57.8±1.4	60.3±1.8	55.6±2.6	63.0±1.4	45.7±1.1
Ger.-Spa.**	67.6±1.1	66.0±1.2	58.3±2.4	70.4±1.9	37.3±5.6	55.6±1.3	60.2±1.8	55.2±4.2	62.4±2.4	44.2±1.9
Average	65.7±2.0	64.9±1.6	58.8±2.2	69.0±1.9	42.8±3.0	55.2±1.6	60.1±2.0	56.1±2.7	62.4±1.9	42.4±1.5

Table 4.8 shows more detailed information on the best results, including different measures of performance and concordance, such as accuracy balanced, F1 score, and Kappa coefficient (κ). The results show that the

monologue task (Acc.: 71.0%) has a better performance than the /pa-ta-ka/ task (Acc.: 64.6%). In addition, there are high levels of concordance (up to 0.57).

Table 4.8. Details of performance measures of the best results obtained in the classification of PD patients vs. HD patients vs. HC subjects. **Ger.**: German, **Spa.**: Spanish, **Fus.**: Fusion, **Phon.**: Phonation. * UBM trained with HC subjects. ** UBM trained with HC subjects and PD patients. Values are reported in terms of mean ± standard deviation.

UBM	Monologue				/pa-ta-ka/			
	Feat.	Accuracy	F1 score	Kappa	Feat.	Accuracy	F1 score	Kappa
Ger.*	Fus.	**71.0±1.4**	**73.6±1.4**	**0.57±0.02**	Phon.	61.9±2.2	66.6±2.0	0.46±0.03
Spa.*	Fus.	67.9±1.6	70.3±1.6	0.52±0.03	Fus.	64.0±1.6	67.8±1.4	0.48±0.02
Ger.-Spa.*	Fus.	67.1±2.6	70.2±2.2	0.51±0.04	Fus.	**64.6±1.5**	**67.9±1.1**	**0.48±0.02**
Ger.**	Fus.	67.7±2.6	71.2±2.4	0.53±0.04	Phon.	61.1±2.2	64.3±2.1	0.42±0.03
Spa.**	Fus.	70.0±1.4	73.3±1.3	0.57±0.02	Fus.	63.0±1.4	65.7±0.8	0.45±0.01
Ger.-Spa.**	Fus.	70.4±1.9	73.1±2.0	0.57±0.03	Fus.	62.4±2.4	68.1±1.6	0.44±0.03

Figure 4.5 shows the confusion matrices per task for the best results obtained in the multi-class classification. In both cases, it is possible to observe that the percentage of correctly classified controls is high (monologue 84.6% and /pa-ta-ka/ 82.2%). Also, for the case of HD patients relatively good percentages (monologue 70.5% and /pa-ta-ka/ 63.5%) are obtained, while PD patients are the worst classified speakers (monologue 57.8% and /pa-ta-ka/ 48.0%) because they are mainly confused with HC, this is due to the fact that most patients are in an early stage of the disease.

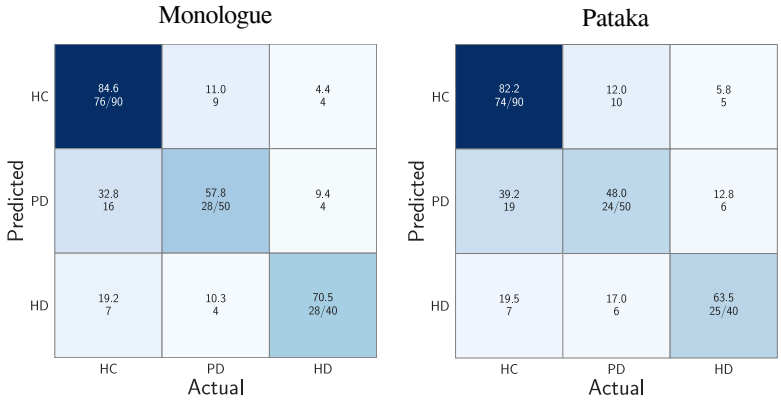

Figure 4.5. Confusion matrices of the best results obtained in the classification of PD patients vs. HD patients vs. HC subjects.

Finally, Figure 4.6 shows a representation per task of each group of the best results presented in Table 4.8. In this case, the concatenation of the 3

supervectors (articulation, phonation, and prosody) are used and performed Linear Discriminant Analysis (LDA) to reduce the final dimension of the supervectors to 2 dimensions. In both figures, it is possible to observe that each group has a region where the majority is located. In addition, the confusion between the HD patients and the other speakers is minimal, while there is overlap between the HC subjects and the PD patients (mostly in the /pa-ta-ka/ task), which supports what was observed in the confusion matrices.

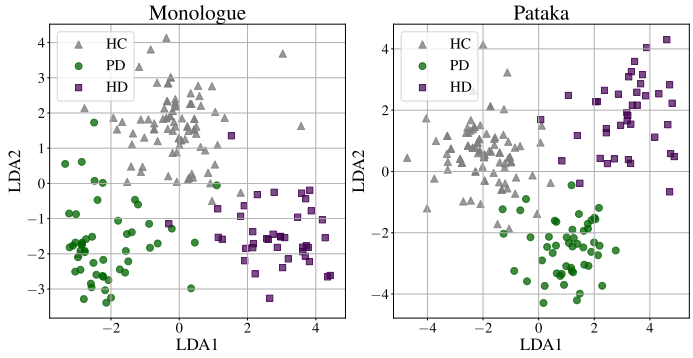

Figure 4.6. Visualization of the samples after applying Linear Discriminant Analysis (2 components) on the best results of the multi-class classification.

Discussion

In the case of classification between PD patients vs. HC subjects, the results are comparable to the state of the art [38]. According to the experiments, the best result obtained an accuracy of 69.8% using the supervector created from the phonation features. When each speech dimension in this experiment were analyzed, it is observed that articulation and phonation were the most discriminating features. This finding can be associated with the hypokinetic dysarthria characteristic of PD patients because they try to model the start of the movements due to the rigidity of the muscles. On the other hand, when each speech dimension in the classification of HD patients vs. HC subjects were analyzed, it was observed that phonation and prosody were the most discriminating dimensions. Therefore, it is possible to associate these dimensions with hyperkinetic dysarthria because in HD rapid movements (chorea) may be reflected as losses of vocal folds control that can be observed through phonation, but also as rapid movements that can be modeled in

prosody with high-level features based mainly on measures of F0 contour, energy, rate, and duration.

Previous studies such as [33], [40] have shown that the phonatory ability of patients with PD and HD is largely impaired as the neurodegenerative disease progresses. In this section, it is possible to verify that this dimension is fundamental for the discrimination of these pathologies, specifically for continuous speech (monologue). In addition, both pathologies present variations in the affectation of the phonatory ability of the patients. Therefore, it is possible to conclude that features that model vocal fold vibration such as variations in F0 and amplitude are useful for the discrimination of pathologies.

Another important point of view in this section is related to the UBM training. There was no trend about which language would best support the classification or if the combination of controls and patients would generate more discriminating supervectors, all the results were very balanced and none of them stood out. However, it can be concluded that it is possible to generate a UBM in different languages and with speakers related or not to the target phenomenon, and at the same time obtain accurate results when classifying. Also, it was possible to observe that there are big differences for the discrimination of patients with PD and HD obtaining accuracies between 75.2% and 86.2% despite the similarity of their symptoms, this could significantly support the evaluation and monitoring of both pathologies.

4.2 TL between different language corpora for disease classification

The classification of pathological speech in different languages has to be carefully conducted to avoid bias towards the linguistic content present in each language. For instance, Czech and German languages are richer than the Spanish language in terms of consonant production, which may cause that it is easier to produce consonant sounds by Czech HD patients than by Spanish PD patients [96]. Despite these language-dependent issues, the results in the classification of pathological speech in different languages could be improved using a TL strategy among languages, i.e., to train a base model with utterances from pathology and a language, and then, to perform a fine-tuning of the weights with utterances from the target language and pathology. Therefore, in this section, TL among languages and pathologies scenario was proposed. Base models were trained to classify PD (Spanish, German, Czech), HD (Czech), or larynx pathologies (LP, English) from HC subjects, and then apply a TL strategy using such pre-trained models to classify a different disease, not included in the base models. The main motivation behind the proposed experiments is to take advantage of existing corpora in order to improve the performance of deep learning models trained to classify problems where it is more difficult to collect high amounts of data. A perfect example is the use of base models trained to classify speech from PD patients vs. HC subjects, which is easier to perform because there is more available data due to the higher incidence of the disease and because of the existing databases. Then, those particular models can be used to improve the accuracy of systems designed to classify HD patients, which is a population very difficult to collect data from.

4.2.1 Methodology

The experiments are divided as follows: (1) CNNs are trained for each dataset with the aim to have base models for the TL among languages and diseases scenarios. (2) The weights of the base models are used to initialize CNNs to classify the PD corpora in different languages. (3) The base models for PD, HD, and LP are used to perform TL among diseases. Figure 4.7 summarizes the general methodology addressed in this section.

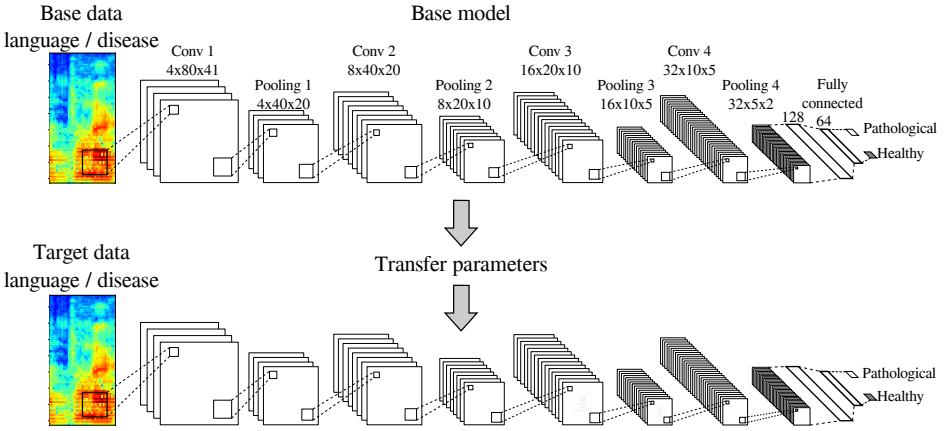

Figure 4.7. TL strategy proposed in this section to classify patients with pathological speech in different languages.

Pre-processing

For the pre-processing of speech signals, we consider the onset and offset segments to model the capability of the patients to start/stop the vocal fold vibration proposed in [66]. The change between a voiced and unvoiced segment is detected based on the presence of F0 values. Once the borders are detected, 80 ms of the signal to the left and to the right were taken, forming segments with 160 ms length. The segmented transitions are transformed into a Mel-scale spectrogram with 80 Mel filters and a time shift of 4 ms to get an 80x41 time-frequency representation to feed the CNN.

Baseline

The baseline model consists of the computation of 12 MFCCs and the energy content distributed in 22 frequency bands according to the Bark scale. These features are extracted for frames of 40 ms length with an overlap of 20 ms in the transition segments. The extracted features for each corpus were used to train an SVM classifier with a Gaussian kernel. The model is validated following a stratified k-fold cross-validation strategy with 10 folds. The hyper-parameters C and γ were optimized in grid-search with $C \in \{10^{-5}, 10^{-4}, \cdots, 10^{3}\}$ and $\gamma \in \{10^{-5}, 10^{-4}, \cdots, 10^{2}\}$ based on the accuracy in the development set. The optimal hyper-parameters are found based on the mode across the 10 folds.

CNN model

The CNN implemented consists of four convolutional layers of size 4, 8, 16, and 32 respectively, each one followed by a max-pooling layer of size 2x2. In addition, 3 fully connected layers of size 128, 64, and 2. ReLu activations are considered in the hidden layers, and a Softmax activation function is considered in the output to make the final decision. More details are shown in Table 4.9. For the training of the network, we used Pytorch with a cross-entropy loss function, using an Adam optimizer [97].

Table 4.9. Architecture of the CNN implemented in this experiment.

CNN Architecture	Input size	Output size
Conv (1x4x3,1)+dropout	1x80x41	4x80x41
Max Pool(2,2)	4x80x41	4x40x20
Conv (4x8x3,1)+dropout	4x40x20	8x40x20
Max Pool(2,2)	8x40x20	8x20x10
Conv (8x16x3,1)+dropout	8x20x10	16x20x10
Max Pool(2,2)	16x20x10	16x10x5
Conv (16x32x3,1)+dropout	16x10x5	32x10x5
Max Pool(2,2)	32x10x5	32x5x2
Lineal(320,128)+dropout	32x5x2	1x128
Lineal(128,64)+dropout	1x128	1x64
Lineal(64,2)	1x64	1x2

4.2.2 Experiments and results

All experiments followed a 10-fold speaker-independent stratified cross-validation strategy to guarantee that data from the same speaker is not included in the train and test sets and that the classes are balanced in each fold. In addition, the results obtained with the proposed methodology are compared to those achieved with the baseline model that consists.

Baseline

Results obtained in the baseline system are observed in Table 4.10 for the considered corpora. Experiments combining two of the three available languages for PD classification were included. The results for PD range from 67.2 to 74.0% depending on the language. The accuracy observed for HD is 85.2%, which is higher than the results observed for PD. Particularly, note that the highest accuracies for the neurodegenerative diseases are observed

for PD-Spanish and HD-Czech, which are the ones where the patients are in more severe stages of the disease (see Figure 3.1). Finally, the accuracy for LP is 94.3%, which is similar to the reported one in related studies where the same data were considered [98].

Table 4.10. Results obtained with the baseline model using articulation features and an SVM classifier. **Acc.**: Accuracy. **Sen.**: Sensitivity. **Spe.**: Specificity. **AUC**: Area under the ROC curve. Values are reported in terms of mean ± standard deviation.

Corpus	Acc. (%)	Spe. (%)	Sen. (%)	AUC
PD-Spanish	74.0±16.2	80.0±15.5	68.0±27.1	0.86
PD-German	69.9± 4.8	67.8±12.6	71.8±14.0	0.68
PD-Czech	70.2±18.0	76.0±21.5	64.3±23.6	0.76
PD-Spanish & PD-German	71.0± 9.3	69.1± 8.0	76.0±17.4	0.77
PD-Czech & PD-German	68.9± 7.4	63.2±13.9	74.9±13.6	0.77
PD-Spanish & PD-Czech	67.2± 6.6	62.0±12.5	72.2± 8.9	0.74
HD-Czech	85.2± 8.0	82.5±19.5	86.1± 9.6	0.94
LP-English	94.3± 4.8	90.7± 7.4	95.4± 4.3	0.99

Training of base models

Table 4.11 shows the results obtained when the CNNs are trained and tested on each corpus separately with random initialization, i.e., without any TL strategy (from scratch). The results obtained upon PD corpora show the highest accuracy for the PD-Spanish corpus (71.0%), while the accuracies in PD-German and PD-Czech are 63.1% and 68.5%, respectively. Similar accuracies are observed for the ones obtained with these base CNN models and the ones with the baseline system in Table 4.10, being the last ones slightly higher especially for PD-Spanish and PD-German. On the other hand, the accuracy reported for HD-Czech is 81.3% and for LP-English database is 98.2%.

Regarding the combination of 2 of the 3 available languages, different insights are observed in these results: the combination of PD-Spanish and PD-German data was the most accurate for the classification of PD (72.8%), improving the accuracy with respect to those obtained individually. However, the AUC for the combined model (0.79) is lower than the one observed only with PD-Spanish (0.82). The base models combining PD-Czech and PD-German highly improve the results compared to those obtained when considering only PD-German or PD-Czech. This could be explained because

possible similarities in the two languages and also in the neurological state of the patients of these two databases (see Figure 3.1). Conversely, the fusion of PD-Spanish with PD-Czech has the opposite effect due to the differences between the Spanish and Czech languages, and to the fact that Czech PD patients are in a lower stage of the disease compared to the Spanish PD patients, which is reflected in the total MDS-UPDRS-III score.

In addition, it is important to mention the variability that exists in the different performance metrics calculated, it is not favorable to have these large variations because they do not allow to quantify exactly the performance of the system, but for the proposed methodology, we can continue with these results, it is also possible to verify if the TL strategy helps to reduce this variability.

Table 4.11. Classification results obtained with the base models for each corpus. **Acc.**: Accuracy. **Sen.**: Sensitivity. **Spe.**: Specificity. **AUC**: Area under the ROC curve. Values are reported in terms of mean ± standard deviation.

Corpus	Acc. (%)	Spe. (%)	Sen. (%)	AUC
PD-Spanish	71.0±15.9	68.0±28.6	74.0±25.0	0.82
PD-German	63.1±11.7	83.1±17.7	43.1±38.0	0.68
PD-Czech	68.5±14.1	94.0±13.5	42.0±33.2	0.76
PD-Spanish & PD-German	72.8± 5.0	66.6±13.9	79.1±12.2	0.79
PD-Czech & PD-German	71.9± 8.5	67.2±15.2	76.4±26.5	0.76
PD-Spanish & PD-Czech	64.8±11.5	61.8±28.0	68.0±25.7	0.67
HD-Czech	81.3±20.6	67.5±19.1	95.0±10.5	0.88
LP-English	98.2± 2.8	76.7±36.3	100± 0.0	0.95

Transfer learning among languages

The results with the TL strategy among languages for the same pathology (PD) are shown in Table 4.12. A CNN trained with utterances from the base language is fine-tuned with utterances from the target language. The accuracy highly improves when the target languages are German and Czech, with respect to the results observed with the base CNN models shown in Table 4.11. The accuracy improved up to 77.3% for PD-German when the model is fine-tuned from PD-Spanish, and up to 72.6% for PD-Czech, also when the model is fine-tuned from PD-Spanish. These results outperform the ones obtained with the baseline system in Table 4.10. The accuracy for PD-Spanish also improves from 71.0% in the base model to 75.0% when the

network is initialized with the combination of PD-Czech and PD-German, improving also the accuracy with respect to the one observed with the baseline system. The results obtained with the TL strategy among languages are also more balanced in terms of the specificity and sensitivity than those observed with the base CNNs and the baseline experiments. In addition, the standard deviations of the transferred CNNs are also smaller, which leads to an improvement in the stability of the models.

Table 4.12. Classification results obtained in TL among languages using CNNs. **Acc.**: Accuracy. **Sen.**: Sensitivity. **Spe.**: Specificity. **AUC**: Area under the ROC curve. Values are reported in terms of mean ± standard deviation.

Base corpus	Target corpus	Acc. (%)	Spe. (%)	Sen. (%)	AUC
PD-German	PD-Spanish	70.0±12.5	62.0±19.9	78.0±23.9	0.78
PD-Czech		72.0±13.1	67.0±11.6	78.0±23.9	0.84
PD-Czech & PD-German		75.0±14.3	62.0±19.9	88.0±16.8	0.78
PD-Spanish	PD-German	77.3±11.3	86.2±13.8	68.3±14.3	0.82
PD-Czech		76.7± 7.9	87.5±11.0	66.0±15.6	0.79
PD-Spanish & PD-Czech		73.7±13.1	82.9±20.8	64.4±21.3	0.81
PD-Spanish	PD-Czech	72.6±13.9	82.0±14.8	62.0±28.9	0.83
PD-German		70.7±14.5	80.0±16.3	62.5±26.3	0.76
PD-Spanish & PD-German		61.5±16.3	90.0±10.5	55.1±21.6	0.76

Transfer learning among diseases

The results with the TL strategy among diseases are shown in Table 4.13. CNNs trained with all PD data to fine-tune models were considered to classify the HD-Czech and the LP-English corpora; and base CNNs from HD-Czech and LP-English to fine-tune models to classify the three PD corpora. The highest accuracy classifying HD-Czech (86.2%) and PD-German (83.7%) were observed when the model was pre-trained with the LP-English data. This is explained because the high accuracy observed for LP-English in the base CNN (98.2%), thus the classification benefit from the best initial model. Particularly, the accuracy observed for PD-German outperformed the best results observed either for the baseline system, for the base PD-German model, and for the TL among languages. The observed behavior does not apply when the target corpora are PD-Spanish or PD-Czech, where the best results were obtained from the TL among languages scenario. Finally, the results classifying the LP-English corpus also outperformed those observed

in the baseline system and in the base CNN model by 5.5% and by 1.6%, respectively.

The use of PD-Spanish as a target corpus to be fine-tuned from the HD-Czech and LP-English corpora is constrained due to the fact that the base corpora are at the same time from a different disease and a different language. An additional experiment was performed using the HD-Czech corpus as the base model and the PD-Czech as the target corpus, with the aim to evaluate only a TL among diseases, keeping constant the language of the data. In this case, we observed an accuracy of 71.0%, which is a good result for that corpus, compared to the other results obtained in these experiments and in the literature, where the same data was considered [38]. This results can also be explained because the patients from the HD-Czech and PD-Czech corpora performed the same speech tasks. TL approach from the LP-English corpus to the PD-Czech corpus was also considered for comparison purposes, and obtained an accuracy of 69.7%. In addition, fine-tuning was also performed using PD-Czech as the base model and HD-Czech as the target corpus, in this case the proposed TL strategy (Acc.: 85.0%) outperformed both the baseline and the from scratch model; however, the best result for the HD-Czech corpus classification was obtained using LP-English as the base with an accuracy of 86.2%.

Discussion

The difference in results obtained in the base models (see Table 4.11) for the PD data in the three languages can be explained considering the information provided in the Chapter 3. For patients in the PD-Spanish corpus, the total MDS-UPDRS-III score and its speech item is higher compared with the PD-German and PD-Czech patients, i.e., there are patients with higher disease severity in the PD-Spanish data compared to PD-German and PD-Czech. At the same time, the difference in the results obtained for HD patients with respect to those observed in PD patients can be likely explained because the speech of HD patients is more affected than in the PD corpora, causing that the results classifying HD are more accurate than those classifying PD. This behavior also causes that the CNNs pre-trained with the HD data are more accurate as base models than their counterparts pre-trained with PD utterances.

The combination of models in different languages has to be carefully con-

Table 4.13. Classification results obtained in TL among languages using CNNs. **Acc.**: Accuracy. **Sen.**: Sensitivity. **Spe.**: Specificity. **AUC**: Area under the ROC curve. Values are reported in terms of mean ± standard deviation.

Base corpus	Target corpus	Acc. (%)	Spe. (%)	Sen. (%)	AUC
HD-Czech	PD-Spanish	72.0±17.5	84.0±18.3	60.0±24.9	0.78
LP-English		70.0± 9.4	90.0±21.6	50.0±25.3	0.76
HD-Czech	PD-German	76.1± 6.3	66.9± 9.9	85.3±14.8	0.78
LP-English		83.7±13.5	77.3±19.5	90.0± 9.7	0.89
HD-Czech	PD-Czech	71.0±22.3	54.0±31.3	88.0±16.9	0.79
LP-English		69.7±11.6	61.5±25.3	78.0±14.8	0.77
PD-Spanish	HD-Czech	82.5±12.1	70.0±19.7	95.0±10.5	0.91
PD-German		82.5± 8.7	72.5±14.2	92.5±12.1	0.89
PD-Czech		85.0±17.5	77.5±32.1	92.5±12.1	0.91
PD-Czech & PD-German		83.8±11.9	80.0±15.8	87.5±17.6	0.88
PD-Czech & PD-Spanish		82.5± 8.7	75.0±16.6	90.0±12.9	0.94
PD-German & PD-Spanish		81.2±15.9	67.5±23.7	95.0±10.5	0.93
LP-English		86.2±10.9	87.5±17.7	85.0±12.9	0.89
PD-Spanish	LP-English	98.4± 2.0	79.2±23.1	100±0.0	0.99
PD-German		99.6± 0.8	95.0±10.5	100±0.0	0.99
PD-Czech		98.8± 0.6	96.7±10.5	100±0.0	0.99
PD-Czech & PD-German		98.8± 1.4	82.5±22.0	100±0.0	0.99
PD-Czech & PD-Spanish		99.0± 1.4	86.6±18.5	100±0.0	0.99
PD-German & PD-Spanish		98.7± 1.6	80.0±24.0	100±0.0	0.99
HD-Czech		99.6± 0.8	95.0±10.5	100±0.0	0.99

ducted to guarantee an improvement in the results. The data to be combined should be similar in terms of aspects like language, disease severity, and acoustic conditions. This is why the combination of PD-German and PD-Czech data is successful to improve the accuracy with respect to the results in separate languages, while the accuracy combining PD-Spanish and PD-Czech was lower than the one observed for instance only with PD-Spanish.

The highest accuracy for PD-German and PD-Czech was obtained when the base language is Spanish. This fact is explained considering that Spanish speakers have the best initial separability because those patients are in more severe stages of the disease. Hence, the other two languages benefit from the best initial model. An additional advantage observed after the TL among languages is that the stability of the trained models after pre-training is better than the one observed in the base CNN models without pre-training.

The TL among diseases improved the accuracy to classify HD when considering the base models trained to classify LP-English. The improvement in the results is explained because the high initial accuracy observed for LP-

English, and the fact that both diseases share some symptoms like phonatory impairments and abnormal vocal fold vibration [98]. On the other hand, when we performed the TL among diseases scenario but keeping constant the language and the speech tasks of the corpus, an improvement in the accuracy of the model to classify the target language was achieved, as it was shown when we performed the TL from HD-Czech to PD-Czech and PD-Czech to HD-Czech.

Finally, the results obtained in this section confirm that it is possible to transfer knowledge from a deep learning model trained for a specific domain to improve the accuracy of a different corpus, which includes speech data from a different disease or language.

4.3 Transfer learning in PD based on emotions

In these experiments, we want to validate whether a trained model for emotion recognition improves the classification of depression in PD using TL techniques. Studies have shown that different emotions in PD can be identified and classified using speech signals [99], [100]. Therefore, the IEMOCAP database was used (described in Section 3.7) created for emotion classification as a base model to perform fine-tuning in a PD corpus (described in Section 3.6) using two different approaches: (1) classification of PD patients with depression (D-PD) vs. PD patients without depression (ND-PD), (2) classification of depression level in PD patients.

4.3.1 Methodology

Figure 4.8 summarizes the methodology addressed for this experiment. Initially, two architectures based on CNNs were trained using the IEMOCAP database for the classification of four emotions: anger, neutral, sadness, and happiness. The same procedure performed in the previous section was implemented to obtain the spectrograms, and the LeNet architecture (see Table 4.9). In addition, a new architecture based on a ResNet topology was proposed. Then, from the models trained with the IEMOCAP database, we performed a fine-tuning of the base model for two different approaches: (1) classification of D-PD patients vs. ND-PD patients; D-PD patients are participants with a depression score greater than or equal to one on the MDS-UPDRS-I scale. ND-PD are participants with a depression score equal to zero. (2) Multi-class classification of the depression score on the MDS-UPDRS-I scale (range 0-3).

ResNet architecture

In addition to the previously mentioned LeNet architecture, we implemented a ResNet model with 6 residual blocks and 3 main blocks with 16, 32, and 64 feature maps, respectively. Table 4.14 shows details of the implemented architecture. The convolutional layers are represented as $(I \times O \times K,\ S)$, where I and O are the numbers of input and output channels, respectively, K is the kernel size (square), and S is the stride (square). ReLu activations are considered in the hidden layers, and a Softmax activation function is applied in the output to make the final decision.

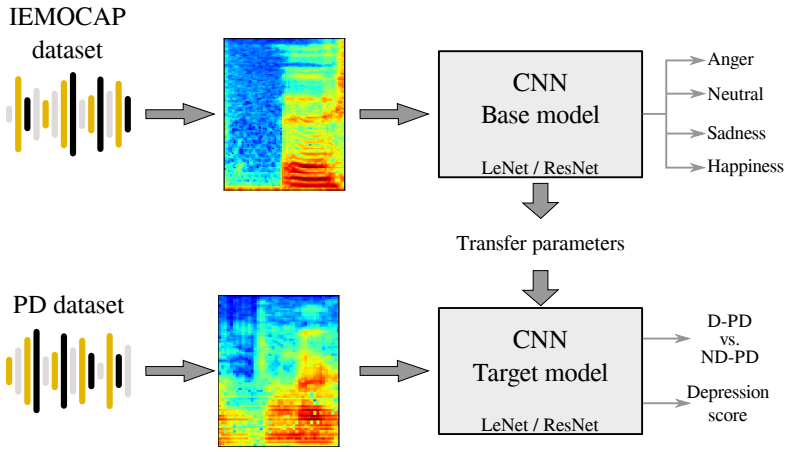

Figure 4.8. General methodology

For this architecture, all raw signals are segmented into 500 ms chunks with 250 ms time-shift. From each chunk, a time-frequency representation is obtained using the STFT to feed the CNNs. Each spectrogram is obtained using Hanning windows of 32 ms and a step-size of 8 ms, forming 63 time frames per chunk. Each representation is transformed into a Mel-scale spectrogram with 128 Mel filters; therefore, the final representation to feed the CNN is 128×63.

Table 4.14. Architecture of the ResNet-based model. Conv.: Convolution, Avg. Pool.: Average pooling, FC: Fully connected.

Stage	Type of layer	Output size
Conv. 1	Conv. (1×16×3,1)	128×63×16
Block 1	Conv. (16×16×3,1) Conv. (16×16×3,1) } ×2	128×63×16
Block 2	Conv. (16×32×3,2) Conv. (32×32×3,1) } ×2	64×32×32
Block 3	Conv. (32×64×3,2) Conv. (64×64×3,1) } ×2	32×16×64
Pooling	Avg. Pool. (32,16)	1×1×64
Output	FC (64,2)	1×1×2

4.3.2 Experiments and results

To obtain the base models, the architectures were trained using sessions 1-4 of the IEMOCAP database, the validation of each CNN was performed from

session 5 of the same database. Experiments for the target database (PD-Depression) followed a 5-fold speaker-independent stratified cross-validation strategy to guarantee that data from the same speaker is not included in the train and test sets, in addition to the classes are balanced in each fold. The results obtained by performing a fine-tuning from the base model are compared with the same architecture trained from scratch in order to validate whether the TL strategy supports the classification of depression in PD.

Base model

The results obtained for the LeNet and ResNet architectures trained with the IEMOCAP database are shown in Table 4.15. It is possible to observe that the ResNet architecture has a better performance over the LeNet architecture with a balanced accuracy of 48.6%; however, it is important to mention that the LeNet architecture is simpler and it is only one percentage point below the ResNet.

Table 4.15. Emotion classification using the IEMOCAP database. **Acc**: Accuracy. **B. Acc.**: Balanced Accuracy.

Architecture	**Acc. (%)**	**B. Acc. (%)**	**F1-Score (%)**	κ
LeNet	42.5	47.4	41.0	0.23
ResNet	47.1	48.6	43.7	0.28

Figure 4.9 shows the confusion matrix of the results reported in the previous table. In both cases, we can observe that the sad emotion is the best classified with a percentage of success of 85.3% for LeNet and 80.0% for ResNet. On the other hand, the neutral trait is the second easiest emotion to predict for the ResNet architecture with an accuracy of 65.4%, while for the LeNet architecture, the anger emotion is the second best performing emotion with an accuracy of 44.7%. Finally, for both architectures, the emotion happiness is the most difficult to predict, being mainly confused with the neutral and sad emotions.

Bi-class classification

For the bi-class classification, we divided the PD-Depression (see Table 3.5) database into 2 groups, patients with a depression score greater than or equal to one (D-PD), and patients with a depression score equal to zero

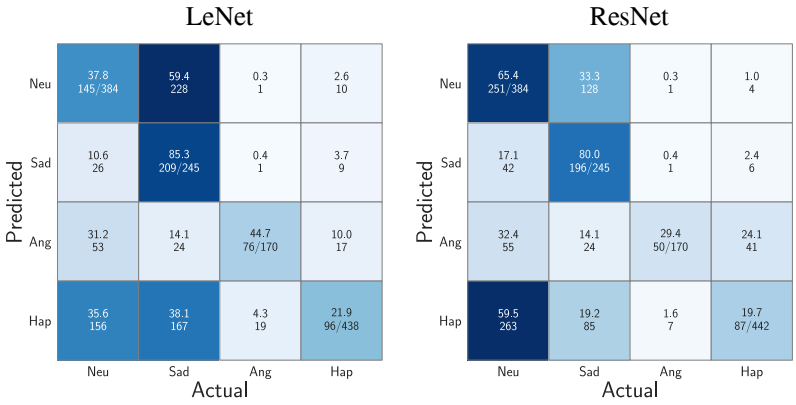

Figure 4.9. Confusion matrices obtained in emotion classification using the IEMOCAP database. Neu: Neutral. Sad: Sadness. Ang: Angry. Hap: Happy.

(ND-PD). The results obtained for this experiment are shown in Table 4.16. The results show that in both architectures the TL strategy improved the system performance by 3.3% compared to the model trained from scratch. In addition, the LeNet architecture is better than the ResNet architecture with an accuracy of 86.7% and 73.3%, respectively.

Table 4.16. Classification of D-PD patients vs. ND-PD patients. **Arch.**: Architecture. **Acc**: Accuracy. **Spe.**: Specificity. **Sen.**: Sensitivity. **F1**: F1-Score. Values are reported in terms of mean ± standard deviation.

	From scratch				Transfer learning			
Arch.	**Acc. (%)**	**Spe. (%)**	**Sen. (%)**	**F1 (%)**	**Acc. (%)**	**Spe. (%)**	**Sen. (%)**	**F1 (%)**
LeNet	83.3± 8.3	85.7±13.1	80.0±24.5	82.8± 9.2	**86.7±7.45**	**91.4±12.8**	**80.0± 2.2**	**86.6±7.3**
ResNet	70.0±13.9	77.1±21.7	60.0±31.4	67.8±15.2	73.3±7.0	77.1±16.3	68.0±17.9	72.9±6.5

Figure 4.10 shows the distribution of the scores in both architectures separately when the fine-tuning strategy was implemented. In the case of the LeNet architecture (left), it is possible to observe that the number of misclassified ND-PD patients (black bins greater than 0.5) are few, while there is a greater error when classifying D-PD participants as ND-PD (purple bins less than 0.5). The above mentioned can be supported from Table 4.16 in the TL approach, where we can observe that there is a higher specificity (91.4%) than sensitivity (80.0%), which implies that the system allows classifying more easily ND-PD patients than D-PD patients. On the other hand, for the ResNet architecture (right) a similar behavior is observed in the figure but with more samples misclassified in both classes, for this case D-PD pa-

tients are mostly confused with ND-PD (Sen.: 68.0%), than ND-PD patients classified as D-PD (Spe.: 77.1%).

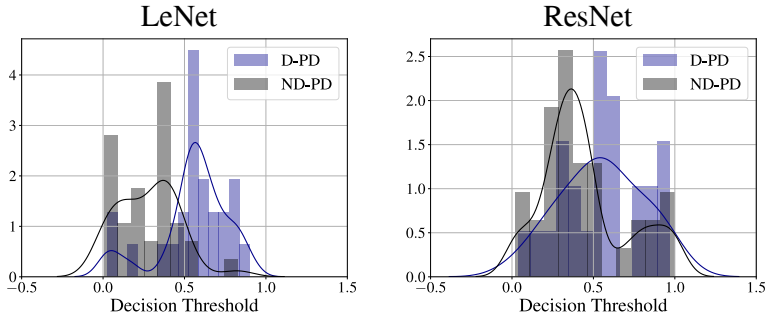

Figure 4.10. Histograms and the corresponding probability density distributions of the scores obtained in the classification of D-PD patients vs. ND-PD patients.

Multi-class classification

The level of depression of each participant was predicted based on a multi-class classification, the value of the label for each patient is that assigned by a specialist in the depression item on the MDS-UPDRS-I scale, which varies in a range of 0-3. The results obtained are shown in Table 4.17. In this case, the LeNet architecture outperformed ResNet again and although the fine-tuning strategy helped to improve the performance metrics, the results are low, with a balanced accuracy of 36.7% for LeNet and 32.4% for ResNet.

Table 4.17. Classification of the depression level in PD patients. **Arch.**: Architecture. **Acc**: Accuracy. **B. Acc.**: Balanced Accuracy. **F1**: F1-Score. Values are reported in terms of mean ± standard deviation.

	From scratch				Transfer learning			
Arch.	**Acc. (%)**	**B. Acc.(%)**	**F1 (%)**	κ	**Acc. (%)**	**B. Acc.(%)**	**F1 (%)**	κ
LeNet	50.0±16.7	25.6±9.7	41.7±1.2	-0.03±0.18	**60.0±9.1**	**36.7±11.7**	**53.1±13.1**	**0.17±0.25**
ResNet	51.7± 7.0	29.0±5.7	42.3±5.1	-0.02±0.08	53.3±4.6	32.4± 5.7	44.0± 3.2	0.05±0.13

Despite the low performance obtained in the multi-class classification, we plotted the confusion matrix in both architectures when TL was performed (see Figure 4.11). In the confusion matrices, it can be observed that the system is over-fitting to class zero and one, this is due to the fact that these are the classes that contain the largest number of participants, 35 and 18, respectively. This is the reason why in both matrices the success rates for these classes are higher. Regarding classes 2 and 3, the system did not predict

any sample, but this is because these classes do not have a sufficient amount of participants to train a deep approach.

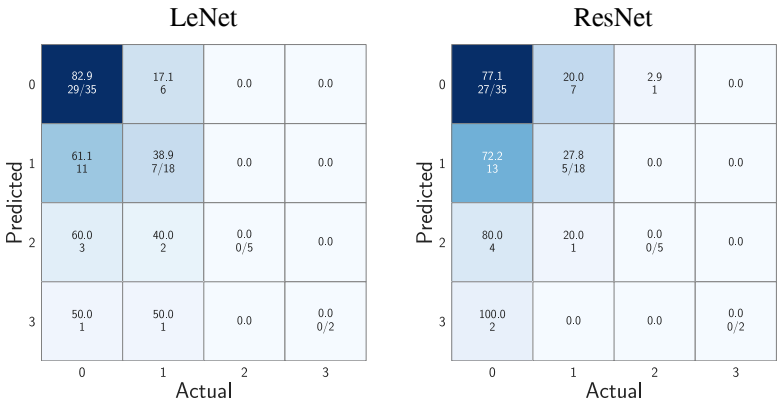

Figure 4.11. Confusion matrices obtained in classification of the depression level in PD patients.

4.4 Freezing layers to classify PD from speech

This section search to explore whether by performing TL between the same pathology but in different languages it is possible to improve the results obtained in the previous sections but performing a freezing layer strategy [79], [101], [102]. Therefore, the base models trained in 3 different languages (German, Czech, and Spanish) for PD, i.e., PD-German, PD-Czech, and PD-Spanish were used. Then, we performed TL between each database in order to evaluate how much the classification of PD patients vs. HC subjects can support base models trained with the same pathology but in a different language. Unlike the previous sections, in these experiments, a TL with layer freezing is performed. This section was motivated by the work performed in [103], where the efficiency of partial layer freezing was demonstrated for image recognition using CNNs when TL is applied to a small target database. Therefore, a TL with partial freezing of convolutional layers is performed, that is, we show the effect of copying all layers and fine-tune a sequential number of layers, freezing some of them from the base model.

4.4.1 Methodology

The methodology of this section is divided as follows: (1) CNNs are trained to classify PD patients vs. HC speakers in each language individually (the same models obtained in the previous sections). Thus, they can be used as a base model in TL. For this case, we use a random parameter initialization. The results are the same as those obtained in Table 4.11 and are used as the baseline of this experiment. (2) CNNs from each language are re-trained with data from the remaining two languages using a TL strategy with layer freezing, i.e., we sequentially freeze the convolutional layers in order to keep a constant part of the learned weights from the base models, Figure 4.12 summarizes this procedure. For this case, the number in the lock means the number of sequentially frozen convolutional layers.

The CNN architecture implemented is the same as described in Section 4.2 (see Table 4.9). In addition, the pre-processing and extraction of the spectrograms were the same as developed in this section, i.e, 80x41 time-frequency representation to feed the CNN. All experiments were performed through a 10-fold speaker-independent stratified cross-validation strategy. In addition, McNemar's test was performed between the baseline and the best performing model after TL with layer freezing. This test checks if the disagreements be-

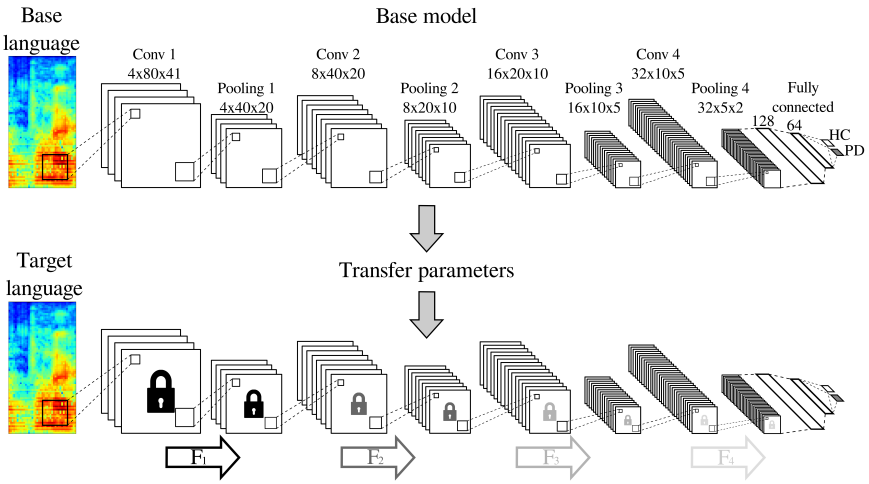

Figure 4.12. TL strategy proposed in this section to classify PD patients vs. HC subjects from speech with utterances from different languages. F_x, $x \in \{1, 2, 3, 4\}$ indicates the number of sequentially frozen convolutional layers.

tween the two cases match. In terms of comparing two binary classification algorithms, the test evaluates whether the two models make errors in the same proportion [104].

4.4.2 Experiments and results

CNNs were fine-tuned for each language using a TL strategy with layer freezing denoted as F_x, where x indicates the number of sequentially frozen convolutional layers. For example, F_3 means freezing of the first 3 convolutional layers, i.e., only the fourth convolutional layer is adjusted. Table 4.18 shows results when base models in Spanish and German are fine-tuned to classify Czech speakers. In this case, the best results are obtained in F_0 for both base models, i.e., TL without layer freezing. The results improve by 4% with respect to the baseline when the Spanish base model is re-trained (from 68.5% to 72.5% for the Spanish base model, and 70.6% for the German base model). McNemar's test was performed to compare the baseline and the best performing model (Spanish base model), a significance level of 0.05 was used. In this case, the p-value obtained exceeded the significance level ($p > 0.05$), which implies that both models make errors in the same proportion, i.e., TL had no significant effect on the results. This is because the layer freezing increases the sensitivity of the models, but reduces the specificity, similar to

the baseline (94% sensitivity). This behavior generates over-fitting in the system and also low accuracy.

Table 4.18. Classification of PD vs HC subjects using TL with Czech as target language. **T.Lang.:** Target language. **F.L.:** Frozen layers. **Acc.**: Accuracy. **Sen.**: Sensitivity. **Spe.**: Specificity. Values are reported in terms of mean ± standard deviation.

		Spanish base model			German base model		
T.Lang.	**F.L.**	**Acc. (%)**	**Sen. (%)**	**Spe. (%)**	**Acc. (%)**	**Sen. (%)**	**Spe. (%)**
Czech	F_0	**72.5±13.9**	**82.0±14.7**	**62.0±28.9**	**70.6±14.6**	**80.0±16.3**	**62.5±26.4**
	F_1	65.7±12.6	86.0±18.9	45.0±22.7	65.9±11.5	84.0±12.6	48.0±31.5
	F_2	60.4±17.1	76.0±30.9	44.0±30.9	65.4±16.5	86.0±25.0	45.5±30.7
	F_3	65.9±19.4	78.0±17.5	54.0±35.3	66.5±17.3	92.0±13.9	40.0±32.7
	F_4	60.6±15.1	94.0± 9.6	27.0±34.6	58.5±14.5	92.0±13.9	25.0±34.4
Czech baseline		68.5±14.1	94.0±13.5	42.0±33.2	68.5±14.1	94.0±13.5	42.0±33.2

The results when the German database was used as the target model are shown in Table 4.19. On the one hand, when we used the Spanish base model, the best result is obtained with F_2 with an accuracy of 78.3%, mainly because the sensitivity improves with respect to F_0. This is because the Spanish base model has higher sensitivity (74.0%) than specificity (68.0%) (see Table 4.11). On the other hand, when we used the Czech base model, CNN with F_4 improved the accuracy by 5.3% compared to F_0 (without layer freezing). In addition, this result improves the accuracy by 18.9% compared to the German base model (without using TL). For this case, the statistical test produced a $p \ll 0.05$, which implies that TL has a significant effect on the performance of the model.

Table 4.19. Classification of PD vs HC subjects using TL with German as target language. **T.Lang.:** Target language. **F.L.:** Frozen layers. **Acc.**: Accuracy. **Sen.**: Sensitivity. **Spe.**: Specificity. Values are reported in terms of mean ± standard deviation.

		Spanish base model			Czech base model		
T.Lang.	**F.L.**	**Acc. (%)**	**Sen. (%)**	**Spe. (%)**	**Acc. (%)**	**Sen. (%)**	**Spe. (%)**
German	F_0	77.3±11.3	86.3±13.8	68.3±14.2	76.7± 7.9	87.5±11.0	66.0±15.6
	F_1	75.5± 6.7	86.0±13.6	65.0±14.1	77.9±11.9	82.1±18.1	73.8±19.8
	F_2	**78.3±11.3**	**90.0±11.0**	**68.6±13.5**	75.0±12.6	78.4±21.8	71.5±13.2
	F_3	75.7± 8.5	78.8±21.8	72.6±17.7	74.7±10.6	80.7±14.4	68.8±17.0
	F_4	75.0± 9.8	71.4±18.7	78.6±16.7	**82.0±11.0**	**93.3±7.7**	**70.7±18.4**
German baseline		63.1±11.7	43.1±38.0	83.1±17.7	63.1±11.7	43.1±38.0	83.1±17.7

Table 4.20 shows the results of classifying the speakers of the Spanish corpus using base models in German and Czech. For this case, we obtained the

best result when the first layer was frozen (F_1), improving the accuracy by up to 7% compared to the CNN without layer freezing (F_0). This improvement is supported by McNemar's test with a $p \ll 0.05$ which implies a significant change in model performance because of layer freezing. These results confirm that the TL strategy with layer freezing can improve the classification accuracy of systems when fine-tuning some layers from the base model.

Table 4.20. Classification of PD vs HC subjects using TL with Spanish as target language. **T.Lang.:** Target language. **F.L.:** Frozen layers. **Acc.**: Accuracy. **Sen.**: Sensitivity. **Spe.**: Specificity. Values are reported in terms of mean ± standard deviation.

		Czech base model			**German base model**		
T.Lang.	**F.L.**	**Acc. (%)**	**Sen. (%)**	**Spe. (%)**	**Acc. (%)**	**Sen. (%)**	**Spe. (%)**
Spanish	F_0	72.0±13.1	67.0±11.6	78.0±23.9	70.0±12.5	62.0±19.9	78.0±29.0
	F_1	**78.0±16.8**	**76.0±20.6**	**80.0±18.8**	**77.0±11.6**	**64.0±18.4**	**90.0±14.1**
	F_2	74.0±10.7	72.0±14.0	76.0±24.6	71.0±20.8	54.0±26.7	88.0±19.3
	F_3	74.0±18.4	66.0±25.0	82.0±17.5	76.0±13.5	60.0±23.1	92.0±14.0
	F_4	73.0± 6.7	60.0±13.3	86.0± 9.7	68.0±14.7	48.0±30.1	88.0±10.3
Spanish baseline		71.0±15.9	74.0±25.0	68.0±28.6	71.0±15.9	74.0±25.0	68.0±28.6

From the results obtained in this section, it is possible to conclude that the proposed sequential layer freezing strategy appropriately adjusts the network parameters from a baseline model trained in another language. Specifically, in all languages, it was possible to outperform the baseline, i.e., the model trained from scratch. While in 2 of the 3 languages, the sequential layer freezing outperformed the fine-tuning of all convolutional layers (F_0). Therefore, we can again observe the potential of these TL strategies, based on fine-tuning from a base model or freezing of layers from a system training for the same pathology but for a different language.

4.5 Interpretability of CNN in PD classification

Motivated by the results obtained in the previous section on layer freezing in TL, we started to explore studies about the interpretation and understanding of other speaker or environmental traits possibly learned during the training of the neural network. This could be relevant because the detection of additional demographic information possibly learned by the network could help in the process to *open the black-box* and detect possible algorithmic biases or improvement opportunities. An approximation to this goal was presented in the field of automatic speech recognition by J. A. Thompson in [105]. The authors evaluated the transferability of a neural network to be fine-tuned with utterances in different languages, in order to track the language-specificity of each layer. The results indicated that the first 2 layers are fully transferable among languages, layers 2-8 are highly transferable but they evidence some specific-language features. Finally, the last layers are language-dependent, but they can be fine-tuned to the target specific languages.

The main objective of this section is to perform experiments to help in discovering which are the layers of a CNN that are involved in the process of learning particular information potentially useful to characterize different speaker traits such as age, gender, native language, speech disorders, and others. To address the aforementioned objective, we trained a ResNet-based model with data from three corpora in different languages (PD-Spanish, PD-German, and PD-Czech) to discriminate between PD and HC speakers. Then, four intermediate representations or embedding vectors from the network were used in order to classify additional demographic data from the speakers.

4.5.1 Methodology

The methodology addressed in this section consists of the following main stages: the corpora are pre-processed, then a CNN is trained using spectrograms, then embeddings are extracted at different layers of the network, and finally, the classification of each trait is performed using an SVM. This methodology is summarized in Figure 4.13.

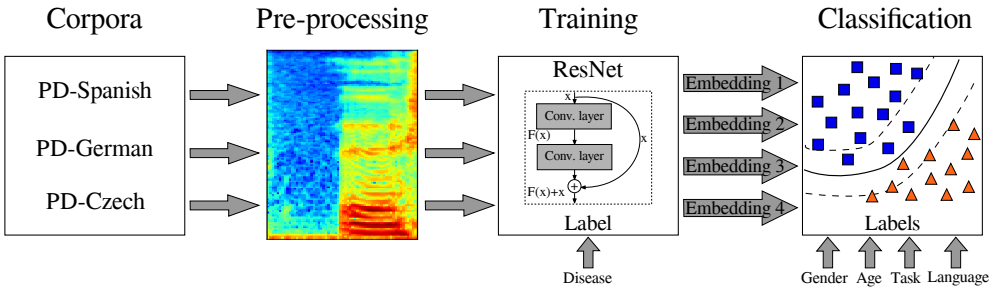

Figure 4.13. General methodology

Embedding extraction

For the classification of the different speaker traits, we extract intermediate network representations or embedding vectors from the previously trained models. A total of 4 representations of the network are obtained. The first embedding corresponds to the output of the first convolutional layer. The second, third, and fourth embedding vectors are obtained at the output of Blocks 1, 2, and 3 from the ResNet model, respectively. A flatten operation upon each representation is performed to convert the matrix into a vector. Then, we computed the mean and standard deviation of each representation obtained from all spectrograms of the same participant in order to create a fixed-length vector to represent each subject. Finally, the dimension of the embedding vectors is reduced using PCA with 95% of the total variance. Table 4.21 also indicates the original dimension of the extracted embedding vectors, and the number of components after PCA, which are used for the later classification of the speaker traits.

Table 4.21. Architecture of the ResNet-based model implemented in this section. Conv.: Convolution, Avg. Pool.: Average pooling, FC: Fully connected. Emb.: embedding vector.

Stage	Type of layer	Output size	Emb. size	Emb. size after PCA
Conv. 1	Conv. (1×16×3,1)	128×63×16	258048	72
Block 1	Conv. (16×16×3,1) Conv. (16×16×3,1) } ×2	128×63×16	258048	101
Block 2	Conv. (16×32×3,2) Conv. (32×32×3,1) } ×2	64×32×32	131072	121
Block 3	Conv. (32×64×3,2) Conv. (64×64×3,1) } ×2	32×16×64	65536	117
Pooling	Avg. Pool. (32,16)	1×1×64	–	–
Output	FC (64,2)	1×1×2	–	–

4.5.2 Experiments and results

The embedding representations extracted from the neural network in the different blocks are used to classify different aspects from the speakers such as age, gender, native language, and speech tasks. SVM classifier with a Gaussian kernel was used. Each experiment is trained and evaluated following a speaker independent 10-fold cross-validation strategy. The procedure is repeated 10 times for a better generalization of the results. The parameters of the classifier are optimized through a grid-search where $C \in \{0.001, 0.005, \cdots, 50, 100\}$ and $\gamma \in \{0.0001, 0.001, \cdots, 100\}$. Finally, the performance of each classifier is evaluated in terms of the accuracy, the AUC, the F1-score, and Cohen's kappa coefficient (κ).

Base models

Different models are trained to classify PD patients vs. HC subjects by implementing the architecture shown in Table 4.14 and using PD-Spanish, PD-German, and PD-Czech corpora in order to train three language-dependent models. In addition, we consider the combination of the 3 corpora in order to train a language independent model. The results are obtained following a speaker independent 10-fold cross-validation strategy. Table 4.22 shows the results obtained for each model. Note that the best model is obtained with the PD-Spanish, followed by PD-German and finally PD-Czech. The results are consistent with the ones observed in Figure 3.1, where it is shown that PD-Spanish patients are in the most advanced stage of the disease, thus their classification should be easier. PD-German patients are in an intermediate stage and PD-Czech patients are in an early stage. Finally, the resulting model when combining the 3 databases yields an accuracy of 70% and an AUC value of 0.75.

Table 4.22. Classification of PD vs. HC subjects for the three language dependent and for the language independent models. **AUC**: Area under the ROC curve. Values are reported in terms of mean ± standard deviation.

Corpus	Accuracy (%)	F1-Score (%)	AUC
PD-Spanish	90.0 ± 3.5	89.9 ± 3.7	0.91 ± 0.06
PD-German	75.0 ± 6.3	74.8 ± 6.4	0.74 ± 0.10
PD-Czech	68.8 ± 2.2	68.8 ± 2.3	0.68 ± 0.04
3-Databases	69.9 ± 9.8	68.8 ± 9.4	0.75 ± 0.09

Age classification

Four embedding vectors extracted from intermediate activations of the previously trained neural network are used to classify age, gender, and the speech tasks produced by each subject from the corpora. For the age classification, we divided the ages of the participants from each database into 3 equally distributed classes to perform a multi-class classification. The results obtained for each embedding on each corpus and for the age classification problem are shown in Table 4.23. The results show that in general, the network is not able to accurately identify the age of the participants in none of the intermediate embedding.

Table 4.23. Age classification using intermediate embeddings (Emb.). Values are reported in terms of mean ± standard deviation.

	PD-Spanish Accuracy (%)	**PD-German Accuracy (%)**	**PD-Czech Accuracy (%)**
Emb. 1	35.1 ± 1.6	32.8 ± 0.0	36.2 ± 1.0
Emb. 2	36.5 ± 1.0	32.8 ± 0.0	35.6 ± 0.6
Emb. 3	37.0 ± 1.4	32.8 ± 0.0	33.7 ± 2.0
Emb. 4	36.0 ± 2.1	32.6 ± 0.4	30.9± 2.8

Figure 4.14 shows the histograms and box-plot of age for the 3 databases, where it can be seen that all speakers are elderly and are in a similar range of age (mean and standard deviation). Therefore, in this case, it is very difficult to classify age per database due to the variability of the data.

Gender classification

For the gender classification, the results are showed in Table 4.24 where gender is accurately recognized in intermediate layers of the neural network. These results indicate that our model implicitly learned information about the gender of the speakers. Specifically, this trait is identified for PD-Spanish from the first layers of the network with an accuracy of up to 84.6%, while for PD-German and PD-Czech, information about gender is only evidenced in the fourth embedding with an accuracy of 85.6% and 78.9%, respectively.

Figure 4.15 shows a visual representation of the activations using PCA in the last embedding for PD-German and PD-Czech, and the second embedding for PD-Spanish. The aim is to observe how the gender information is present in intermediate parts of the network for each corpus. The figure

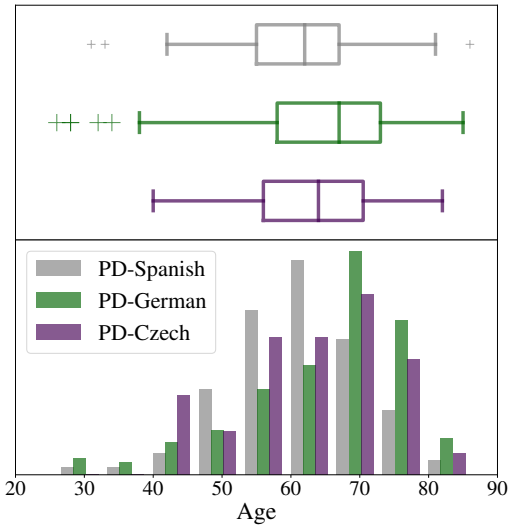

Figure 4.14. Distributions of the age for each database.

Table 4.24. Gender classification using intermediate embeddings (Emb.). Values are reported in terms of mean ± standard deviation.

	PD-Spanish Accuracy (%)	**PD-German Accuracy (%)**	**PD-Czech Accuracy (%)**
Emb. 1	81.2 ± 1.4	49.4 ± 0.0	60.6 ± 0.0
Emb. 2	**84.6 ± 0.8**	49.4 ± 0.0	60.6 ± 0.0
Emb. 3	84.5 ± 1.6	49.4 ± 0.0	63.6 ± 0.8
Emb. 4	81.9 ± 1.9	**85.6 ± 0.8**	**78.9 ± 0.8**

shows how it is possible to identify the gender of the speakers in the three corpora, which confirms that the network implicitly learns the gender of the speakers. Specifically, for the PD-Spanish plot (left), there are 2 clusters where it is possible to identify male and female speakers. It is also possible to differentiate patients and controls, where patients (circles) are located in the bottom right part of the figure, while controls (triangles) are distributed in the left part of the figure. For the case of PD-German (center), the discrimination between male and female speakers is clearer. There is a trend for patients to be found in the bottom part of the figure, and controls in the upper part of the figure. Finally, for PD-Czech (right), it is observed that males are on the left side, while females are on the right side of the figure. For this case, it is not possible to observe clusters to differentiate controls and

patients. This is why this model has the lowest accuracy in that classification problem (see Table 4.22).

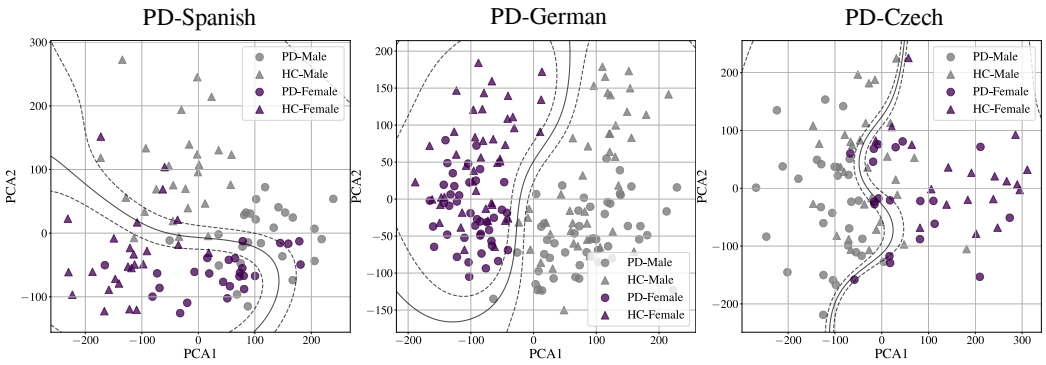

Figure 4.15. Visualization of best embeddings for gender classification after applying PCA with 2 principal components.

Speech task classification

For the specific case of speech task classification we only considered 3 tasks that are common in the 3 databases: (1) DDKs, (2) monologue, and (3) read text. The results obtained in the multiclass classification are showed in Table 4.25. In this case, it is possible to observe that the information about speech tasks in PD-Spanish and PD-Czech appear in the last layers of the network with an accuracy of 64.2% and 74.2%, respectively. While the model trained for PD-German does not find information about the speech task performed by the speakers.

Table 4.25. Speech task classification using intermediate embeddings (Emb.). Values are reported in terms of mean ± standard deviation.

	PD-Spanish Accuracy (%)	**PD-German Accuracy (%)**	**PD-Czech Accuracy (%)**
Emb. 1	59.1 ± 1.4	**48.9 ± 0.6**	63.4 ± 0.6
Emb. 2	58.8 ± 1.2	43.6 ± 1.1	66.5 ± 0.6
Emb. 3	55.2 ± 0.7	42.8 ± 1.2	72.0 ± 0.9
Emb. 4	**64.2 ± 1.5**	48.0 ± 0.6	**74.2 ± 0.7**

The scatter plots in Figure 4.16 indicates that for the three corpora it is possible to differentiate the DDK tasks from continuous speech utterances like the monologue and the read text, which are highly overlapped. These dif-

ferences are observed even for PD-German, in which the accuracy for speech-task classification is very small (48.9%). This result is expected because the nature of the DDK tasks, which makes it totally different than spontaneous speech utterances like the monologue or the read text. In addition, note in the scatter plot on the left (PD-Spanish) that it is possible to differentiate PD patients and HC subjects. HCs (triangles) are observed in the upper part of the figure, while the patients are mainly found in the bottom part of the figure. This behavior is only observed for the PD-Spanish model because it is the most accurate one (90.0%) when discriminating PD patients from HC subjects.

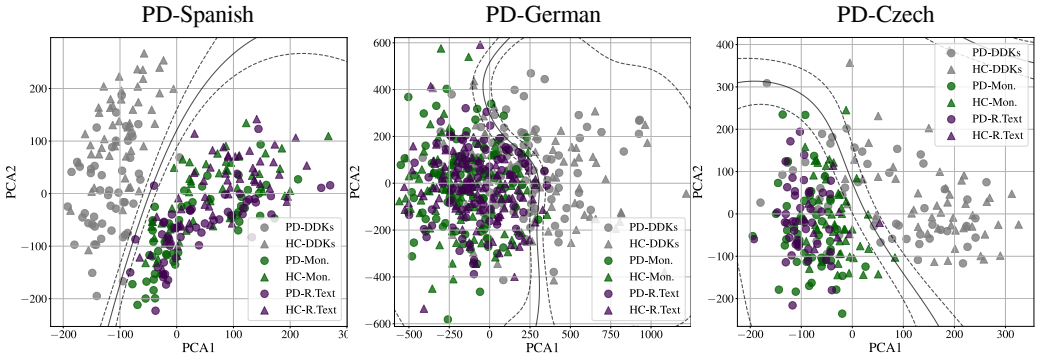

Figure 4.16. Visualization of best embedding for speech task classification after applying PCA with 2 principal components.

Native language classification

The trained model using the three corpora is considered for this experiment. The same four embedding activations are extracted to test whether there is information about the language of the participants in intermediate representations of the CNN. The results are shown in Table 4.26. Note that the native language of the participants can be identified only in the fourth embedding with an accuracy of 82.1%. This fact indicates that despite never using language information during the training process, later stages of the CNN are able to differentiate the language of each participant, which can be considered as a deeper and more complex information about the speaker, similar to the presence of the disease. Moreover, this result agrees with the one obtained in [105], where the authors conclude that the first layers of a CNN trained with several languages are fully transferable between languages,

while the last layers are highly language dependent, as it was confirmed by our experiments.

Table 4.26. Language classification using embeddings (Emb.) extracted from a model trained with the 3 databases. Values are reported in terms of mean ± standard deviation

	Accuracy (%)	**F1-Score (%)**	$\boldsymbol{\kappa}$
Emb. 1	35.6 ± 0.16	33.6 ± 0.2	0.040 ± 0.002
Emb. 2	33.8 ± 0.15	32.4 ± 1.3	0.007 ± 0.003
Emb. 3	41.5 ± 1.08	42.6 ± 0.1	0.141 ± 0.021
Emb. 4	**82.1 ± 0.23**	**84.3 ± 0.1**	**0.765 ± 0.002**

A visual representation using PCA of the activations obtained from the last embedding is shown in Figure 4.17. Clusters of each language are observed. Particularly, speakers from the Czech corpus are very well separated, while some of the Spanish and German speakers are overlapped. Additionally, it is possible to differentiate PD patients and HC subjects for the PD-Spanish corpus, where the patients are clustered in the lower part of the figure, while the controls are in the central part the plot. For the Czech and German corpora, the distinction between patients and HCs is not that clear as in PD-Spanish. However, in some cases patients tend to group in the center of each language-cluster, while the controls tend to surround each of these subgroups.

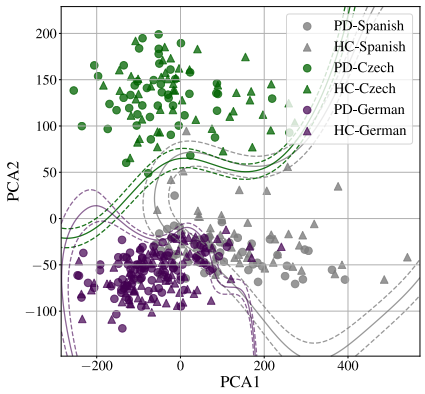

Figure 4.17. Visualization of the 4th embedding for language classification after applying PCA with 2 principal components.

Chapter 5

Conclusions and future work

This work evaluated the suitability of using DL for pathological speech processing applications. In particular, the use of TL methods was proposed, where the main idea is to use the experience/knowledge obtained in certain models to improve the learning of new target phenomena. Different approaches were implemented, including transfer knowledge in classical methods by crossing pathologies and languages for the classification of PD and HD in the Czech language. In addition, different CNNs models were trained to perform a fine-tuning strategy from cross-language and cross-pathology for the discrimination of patients with PD and HD with respect to HC subjects and the estimation of their depression level. Subsequently, a freezing layer strategy was performed for the classification of PD patients vs. HC subjects, in order to evaluate which percentage of information extracted from an original base model is required to perform the automatic classification of the disease. Finally, a novel DL strategy based on embeddings is proposed to know which additional demographic information a CNN learns when it is trained for pathological speech classification

Regarding the transfer knowledge in classical methods, the use of 3 speech dimensions (articulation, phonation, and prosody) was explored for the classification of patients with PD and HD in Czech language using supervectors extracted from the GMM-UBM. For the UBM training, recordings of patients and controls in German and Spanish languages were considered, in addition, feasibility of cross-language and combination of controls and patients in the UBM was evaluated. The results indicate that obtaining a supervector from a UBM trained in other languages can be a good representation to discriminate between patients and/or HC speakers. Besides, it

is possible to observe a trend to obtain the best result by combining the 3 speech dimensions; therefore, we can conclude that these are complementary and each one evaluates different paralinguistic traits that allow to effectively discriminate the different proposed scenarios. Finally, accuracies of up to 86.2% were observed to classify PD vs. HD, and up to 71.0% in a multi-class classification (PD vs. HD vs. HC). This indicates that although both diseases share some symptoms, speech is an important bio-marker to differentiate them and support the correct diagnosis of the pathologies.In general for this approach, it is possible to generate a UBM in different languages and with speakers related or not to the target phenomenon, and at the same time obtain accurate results when classifying. This allows to conclude that using classical techniques, in this case with GMM-UBM, it is possible to perform successfully the knowledge transfer and improve the performance obtained for each experiment with respect to its baseline. This can help to generalize much better the automatic learning systems and to robust the models in charge of characterizing pathological speech.

Besides, the use of a TL strategy based on fine-tuning was considered to classify speech data from patients with neurodegenerative diseases such as PD, HD or LP which were collected from patients in different languages: Spanish, German, Czech, and English. Experiments were conducted in two scenarios: TL among languages and TL among diseases. The results indicate that the TL among languages improved the accuracy to classify PD in German and Czech languages when a base model trained with Spanish utterances is used. The TL among languages also produced more balanced results in terms of specificity-sensitivity and have a lower variance. In general, the TL schemes improved the accuracy in the target corpus only when the base model was robust enough. This was observed when the model trained with PD-Spanish utterances was used to initialize models for PD-German and PD-Czech languages, when the LP-English model was used as base to classify HD-Czech utterances, or when HD-Czech data were used for the base model to fine-tune the PD-Czech model. Also, data from two different languages to train a base model to classify PD were considered. For that case, it is possible to conclude that the data to be combined should be similar in terms of aspects like language, disease severity, or acoustic conditions, in order to have highly accurate models.

The same approach mentioned above was implemented for the classification of depression in PD patients. Base model trained with the IEMOCAP

database for the classification of 4 emotions was considered. Then, based on these trained models, a fine-tuning was performed in the PD-Depression database for 2 different approaches: (1) classification of PD patients with depression vs. PD patients without depression, (2) classification of depression level in PD patients. The results showed that it is possible to improve the performance for the classification of depression in PD using a trained model for emotion recognition as a base. In these experiments, it is again confirmed that an appropriate TL strategy can support the classification of models containing small amounts of data, regardless of whether the base and target models are not closely related. However, it is also possible to conclude that a minimum amount of samples is necessary for the prediction of each label, even though CNN started from previously acquired knowledge, it is necessary for it to adjust its parameters to its new content in order to obtain a better performance. This is the reason why the multiclass classification was low, for this case it is necessary to increase considerably the samples corresponding to levels 2-3 of the depression item.

Finally, in order to understand and learn more about DNNs, 2 experiments were proposed focused on the classification of PD patients. In the first case, methodology based on TL with layer freezing was proposed to classify PD patients and HC subjects from speech in three different languages: Spanish, German, and Czech. The objective is to improve the performance of the network with respect to a random initialization or a full transfer of the parameters. The methodology was implemented using a language as a target model and the remaining 2 languages as base models. The results show that the proposed strategy of freezing layers improves the accuracy of CNNs by up to 7% compared to a full fine-tune of the CNN. In addition, it was observed that the accuracy of the models improves in a range from 4% to 18% compared to the base models without any TL. For the second case, we proposed a methodology to analyze which aspects and speaker traits a CNN can learn when it is trained to classify pathological speech signals. This approach considers 3 language-dependent models trained with different PD databases in Spanish, German and Czech, as well as a language-independent model combining the three corpora. Then, four intermediate activations of the trained CNNs were extracted to evaluate different aspects and traits such as age, gender, speech task, and the native language of the participants. The results indicate that the network not only extracts information about the presence of the disease, but also acquires knowledge at different stages to

recognize the gender of each participant, the speech tasks uttered by the subjects, and the native language of the speakers. From the results, it is possible to observe that the gender can be detected in the first layers of the network for the case of PD-Spanish, and at intermediate and later stages for PD-German and PD-Czech. For the recognition of speech tasks, high accuracies were obtained in the last stage of the network for the Spanish and Czech corpora. Finally, the results showed that only the final stage of the CNN contains information about the native language of the participants.

In general, in each approach, it could be observed that the different TL strategies can considerably improve the performance of models trained with a reduced amount of data, including when the base and target data are not closely related. In addition, it was observed that it is necessary to have a robust base model to perform the parameter transfer, otherwise, the TL strategy may not help or even worsen the performance of the target model. This research work contributes to the state of the art on how to correctly perform a TL strategy for pathological speech classification, including cross-language and/or pathology. In addition to this, this work not only shows that it is possible to increase the performance of the systems from a TL strategy, but it also allows for exploring and reducing the gap of how to create a base model for TL in the field of pathological speech, it also generates discussion of what TL strategy to use for models that include cross-language, cross-pathology and even cross-trait models. Finally, it can also be considered from the last implemented methodology, that this work opens the gap to performing a selection of transferable layers or even the implementation of transferability methods such as a weight assignment at different stages of the network. As future work, we will continue with the fine-tuning trend but in more robust architectures such as CNN-LSTM where frequency and temporal information of the input signal is considered. In addition, to include experiments where a transfer of specific layers is performed based on the results obtained and observe how much it can improve the classification of different traits and/or aspects such as gender, age, native language, among others. Finally, the strategies proposed and implemented in this work could be applied to other bio-markers such as handwriting, gait, and face for the monitoring and classification of patients with neurodegenerative diseases, even a multimodal analysis could be considered.

Appendix A

Publications emerging from the development of this master's thesis

A.1 Journals

- ✓ J. C. Vásquez-Correa, **C. D. Rios-Urrego**, T. Arias-Vergara, M. Schuster, J. Rusz, E. Nöth, and J. R. Orozco-Arroyave. "Transfer learning helps to improve the accuracy to classify patients with different speech disorders in different languages." *Pattern Recognition Letters*, 2021.
- ✓ J. R. Orozco-Arroyave, J. C. Vásquez-Correa, P. Klumpp, P. A. Pérez-Toro, D. Escobar-Grisales, N. Roth, **C. D. Rios-Urrego**, et al., "Apkinson: the smartphone application for telemonitoring Parkinson's patients through speech, gait and hands movement." *Neurodegenerative Disease Management*, 10(3), 2020, pp. 137-157.

A.2 Book Chapters

- ✓ **C. D. Rios-Urrego**, J. C. Vásquez-Correa, J. R. Orozco-Arroyave, and E. Nöth. "Is There Any Additional Information in a Neural Network Trained for Pathological Speech Classification?." *International Conference on Text, Speech, and Dialogue*. Springer, Cham, 2021, pp. 435-447.

✓ **C. D. Rios-Urrego**, J. C. Vásquez-Correa, J. R. Orozco-Arroyave, and E. Nöth. "Transfer Learning to Detect Parkinson's Disease from Speech In Different Languages Using Convolutional Neural Networks with Layer Freezing." *International Conference on Text, Speech, and Dialogue*. Springer, Cham, 2020, pp. 331-339.

A.3 Conferences

✓ D. Escobar-Grisales, **C. D. Rios-Urrego**, D. A López-Santander, J. D. Gallo-Aristizabal, J. C. Vásquez-Correa, E. Nöth, and J. R. Orozco-Arroyave. "Colombian Dialect Recognition Based on Information Extracted From Speech and Text Signals." in *Proccedings of ASRU*, 2021. Accepted.

✓ J. C. Vásquez-Correa, T. Arias-Vergara, **C. D. Rios-Urrego**, M. Schuster, J. Rusz, J. R. Orozco-Arroyave, and E. Nöth. "Convolutional Neural Networks and a Transfer Learning Strategy to Classify Parkinson's Disease from Speech in Three Different Languages." in *Proccedings of CIARP*. Springer, Cham, 2019, pp. 697-706. **Best Paper Award.**

✓ J. C. Vásquez-Correa, T. Arias-Vergara, P. Klumpp, M. Strauss, A. Küderle, N. Roth, S. Bayerl, N. Garcıa-Ospina, P. A. Pérez-Toro, L. F. Parra-Gallego, **C. D. Rios-Urrego**, D. Escobar-Grisales, J. R. Orozco-Arroyave, B. Eskofier, and E. Nöth, "Apkinson: a Mobile Solution for Multimodal Assessment of Patients with Parkinson's Disease." in *Proccedings of INTERSPEECH*, 2019, pp. 964-965.

List of Figures

List of Tables

Bibliography

[1] L. M. Babrak, J. Menetski, M. Rebhan, G. Nisato, M. Zinggeler, *et al.*, "Traditional and digital biomarkers: Two worlds apart?" *Digital Biomarkers*, vol. 3, no. 2, pp. 92–102, 2019.

[2] A. Coravos, S. Khozin, and K. D. Mandl, "Developing and adopting safe and effective digital biomarkers to improve patient outcomes," *NPJ Digital Medicine*, vol. 2, no. 1, pp. 1–5, 2019.

[3] J. Robin, J. E. Harrison, L. D. Kaufman, F. Rudzicz, W. Simpson, and M. Yancheva, "Evaluation of speech-based digital biomarkers: Review and recommendations," *Digital Biomarkers*, vol. 4, no. 3, pp. 99–108, 2020.

[4] V. Boschi, E. Catricala, M. Consonni, C. Chesi, A. Moro, and S. F. Cappa, "Connected speech in neurodegenerative language disorders: A review," *Frontiers in Psychology*, vol. 8, p. 269, 2017.

[5] D. Rijk, L. Launer, K. Berger, M. Breteler, J. Dartigues, *et al.*, "Prevalence of Parkinson's disease in europe: A collaborative study of population-based cohorts. neurologic diseases in the elderly research group," *Neurology*, vol. 54, S21–3, 2000.

[6] J. Jankovic, "Parkinson's disease: Clinical features and diagnosis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[7] S. Pinto, C. Ozsancak, E. Tripoliti, S. Thobois, P. Limousin-Dowsey, and P. Auzou, "Treatments for dysarthria in Parkinson's disease," *The Lancet Neurology*, vol. 3, no. 9, pp. 547–556, 2004.

[8] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients," *Journal of Speech and Hearing Disorders*, vol. 43, no. 1, pp. 47–57, 1978.

[9] M. D. Rawlins, N. S. Wexler, A. R. Wexler, *et al.*, "The prevalence of Huntington's disease," *Neuroepidemiology*, vol. 46, no. 2, pp. 144–153, 2016.

[10] F. O. Walker, "Huntington's disease," *The Lancet*, vol. 369, no. 9557, pp. 218–228, 2007.

[11] G. M. Peavy, M. W. Jacobson, J. L. Goldstein, *et al.*, "Cognitive and functional decline in Huntington's disease: Dementia criteria revisited," *Movement Disorders*, vol. 25, no. 9, pp. 1163–1169, 2010.

[12] C. Saldert, A. Fors, S. Ströberg, and L. Hartelius, "Comprehension of complex discourse in different stages of Huntington's disease," *International Journal of Language & Communication Disorders*, vol. 45, no. 6, pp. 656–669, 2010.

[13] L. Hartelius, A. Carlstedt, M. Ytterberg, M. Lillvik, and K. Laakso, "Speech disorders in mild and moderate Huntington disease: Results of dysarthria assessments of 19 individuals," *Journal of Medical Speech-Language Pathology*, vol. 11, no. 1, pp. 1–15, 2003.

[14] S. Skodda, U. Schlegel, R. Hoffmann, and C. Saft, "Impaired motor speech performance in Huntington's disease," *Journal of Neural Transmission*, vol. 121, no. 4, pp. 399–407, 2014.

[15] H. H. Nguyen and M. A. Cenci, *Behavioral Neurobiology of Huntington's Disease and Parkinson's Disease*. Springer, 2015, vol. 22.

[16] W. Poewe and R. Granata, "Pharmacological treatment of Parkinson's disease," *Movement Disorders: Neurologic Principles and Practice (Watts RL, Koller WC, eds)*, pp. 201–219, 1997.

[17] J. H. Kordower, "Introduction to Parkinson's and Huntington's disease," in *CNS Regeneration*, Elsevier, 1999, pp. 295–298.

[18] D. P. Salmon, P. F. Kwo-on-Yuen, W. C. Heindel, N. Butters, and L. J. Thal, "Differentiation of Alzheimer's disease and Huntington's disease with the dementia rating scale," *Archives of Neurology*, vol. 46, no. 11, pp. 1204–1208, 1989.

[19] T. Arias-Vergara, J. C. Vásquez-Correa, J. R. Orozco-Arroyave, and E. Nöth, "Speaker models for monitoring Parkinson's disease progression considering different communication channels and acoustic conditions," *Speech Communication*, vol. 101, pp. 11–25, 2018.

[20] M. Perez, W. Jin, D. Le, *et al.*, "Classification of Huntington disease using acoustic and lexical features," in *Proceedings of INTERSPEECH*, 2018, pp. 1898–1902.

[21] J. R. Orozco-Arroyave, J. C. Vásquez-Correa, J. F. Vargas-Bonilla, *et al.*, "Neurospeech: An open-source software for Parkinson's speech analysis," *Digital Signal Processing*, vol. 77, pp. 207–221, 2018.

[22] J. R. Orozco-Arroyave, J. C. Vásquez-Correa, P. Klumpp, P. A. Pérez-Toro, C. D. Rios-Urrego, *et al.*, "Apkinson: The smartphone application for telemonitoring Parkinson's patients through speech, gait and hands movement," *Neurodegenerative Disease Management*, vol. 10, no. 3, pp. 137–157, 2020.

[23] D. Van Compernolle, W. Ma, F. Xie, and M. Van Diest, "Speech recognition in noisy environments with the aid of microphone arrays," *Speech Communication*, vol. 9, no. 5-6, pp. 433–442, 1990.

[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[25] H. Wang and B. Raj, "On the origin of deep learning," *ArXiv preprint arXiv:1702.07800*, 2017.

[26] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[27] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[28] D. Wang and T. F. Zheng, "Transfer learning for speech and language processing," in *Proceedings of APSIPA*, IEEE, 2015, pp. 1225–1237.

[29] J. C. Vásquez-Correa, T. Arias-Vergara, C. D. Rios-Urrego, *et al.*, "Convolutional neural networks and a transfer learning strategy to classify Parkinson's disease from speech in three different languages," in *Proceedings of CIARP*, Springer, 2019, pp. 697–706.

[30] M. Wodzinski, A. Skalski, D. Hemmerling, J. R. Orozco-Arroyave, and E. Nöth, "Deep learning approach to Parkinson's disease detection using voice recordings and convolutional neural network dedicated to image classification," in *Proceedings of EMBC*, IEEE, 2019, pp. 717–720.

[31] J. Mallela, A. Illa, B. Suhas, *et al.*, "Voice based classification of patients with amyotrophic lateral sclerosis, Parkinson's disease and healthy controls with CNN-LSTM using transfer learning," in *Proceedings of ICASSP*, IEEE, 2020, pp. 6784–6788.

[32] J. Rusz, C. Saft, U. Schlegel, R. Hoffman, and S. Skodda, "Phonatory dysfunction as a preclinical symptom of Huntington disease," *PloS one*, vol. 9, no. 11, e113412, 2014.

[33] J. Rusz, J. Klempiř, E. Baborová, *et al.*, "Objective acoustic quantification of phonatory dysfunction in Huntington's disease," *PLoS One*, vol. 8, no. 6, e65881, 2013.

[34] L. Moro-Velazquez, J. A. Gomez-Garcia, J. I. Godino-Llorente, *et al.*, "A forced Gaussians based methodology for the differential evaluation of Parkinson's disease by means of speech processing," *Biomedical Signal Processing and Control*, vol. 48, pp. 205–220, 2019.

[35] L. Moro-Velazquez, J. A. Gomez-Garcia, J. I. Godino-Llorente, *et al.*, "Phonetic relevance and phonemic grouping of speech in the automatic detection of Parkinson's disease," *Scientific Reports*, vol. 9, no. 1, pp. 1–16, 2019.

[36] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, "New spanish speech corpus database for the analysis of people suffering from Parkinson's disease.," in *Proceedings of LREC*, 2014, pp. 342–347.

[37] J. Rusz, R. Cmejla, T. Tykalova, *et al.*, "Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task," *The Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 2171–2181, 2013.

[38] M. Novotn, P. Dušek, I. Daly, E. Ržička, and J. Rusz, "Glottal source analysis of voice deficits in newly diagnosed drug-naïve patients with Parkinson's disease: Correlation between acoustic speech characteris-

tics and non-speech motor performance," *Biomedical Signal Processing and Control*, vol. 57, p. 101 818, 2020.

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

[40] T. Arias-Vergara, J. C. Vásquez-Correa, and J. R. Orozco-Arroyave, "Parkinson's disease and aging: Analysis of their effect in phonation and articulation of speech," *Cognitive Computation*, vol. 9, no. 6, pp. 731–748, 2017.

[41] G. Solana-Lavalle, J. Galán-Hernández, and R. Rosas-Romero, "Automatic Parkinson disease detection at early stages as a pre-diagnosis tool by using classifiers and a small set of vocal features," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, pp. 505–516, 2020.

[42] L. L. Murray and L. P. Lenz, "Productive syntax abilities in Huntington's and Parkinson's diseases," *Brain and Cognition*, vol. 46, no. 1-2, pp. 213–219, 2001.

[43] M. Novotn, J. Rusz, R. Čmejla, H. Ržičková, J. Klempıř, and E. Ržička, "Hypernasality associated with basal ganglia dysfunction: Evidence from Parkinson's disease and Huntington's disease," *PeerJ*, vol. 4, e2530, 2016.

[44] J. Hlavnička, T. Tykalová, O. Ulmanová, *et al.*, "Characterizing vocal tremor in progressive neurological diseases via automated acoustic analyses," *Clinical Neurophysiology*, 2020.

[45] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, and E. Nöth, "Convolutional neural network to model articulation impairments in patients with Parkinson's disease," in *Proceedings of INTERSPEECH*, 2017, pp. 314–318.

[46] P. Khojasteh, R. Viswanathan, B. Aliahmad, S. Ragnav, P. Zham, and D. Kumar, "Parkinson's disease diagnosis based on multivariate deep features of speech signal," in *Proceedings of LSC*, IEEE, 2018, pp. 187–190.

[47] D. Korzekwa, R. Barra-Chicote, B. Kostek, T. Drugman, and M. Lajszczak, "Interpretable deep learning model for the detection and reconstruction of dysarthric speech," in *Proceedings of INTERSPEECH*, 2019, pp. 1–5.

[48] D. R. Rizvi, I. Nissar, S. Masood, M. Ahmed, and F. Ahmad, "An LSTM based deep learning model for voice-based detection of Parkinson's disease," *International Journal of Advanced Science and Technology*, vol. 29, no. 5, p. 8, 2020.

[49] B. E. Sakar, M. E. Isenkul, C. O. Sakar, *et al.*, "Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 4, pp. 828–834, 2013.

[50] J. C. Vasquez-Correa, T. Arias-Vergara, M. Schuster, J. R. Orozco-Arroyave, and E. Nöth, "Parallel representation learning for the classification of pathological speech: Studies on Parkinson's disease and cleft lip and palate," *Speech Communication*, vol. 122, pp. 56–67, 2020.

[51] C. Quan, K. Ren, and Z. Luo, "A deep learning based method for Parkinson's disease detection using dynamic features of speech," *IEEE Access*, vol. 9, pp. 10 239–10 252, 2021.

[52] A. H. Shahid and M. P. Singh, "A deep learning approach for prediction of Parkinson's disease progression," *Biomedical Engineering Letters*, vol. 10, no. 2, pp. 227–239, 2020.

[53] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, *et al.*, "Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): Scale presentation and clinimetric testing results," *Movement Disorders*, vol. 23, no. 15, pp. 2129–2170, 2008.

[54] B. Sonawane and P. Sharma, "Speech-based solution to Parkinson's disease management," *Multimedia Tools and Applications*, vol. 80, no. 19, pp. 29 437–29 451, 2021.

[55] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of CVPR*, 2016, pp. 770–778.

[57] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

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

[59] J. D. Arias-Londoño and J. A. Gómez-Garcıa, "Predicting updrs scores in Parkinson's disease using voice signals: A deep learning/transfer-learning-based approach," in *Proceedings of AAPS*, Springer, 2019, pp. 100–123.

[60] O. Karaman *et al.*, "Robust automated parkinson disease detection based on voice signals with transfer learning," *Expert Systems with Applications*, vol. 178, p. 115 013, 2021.

[61] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of CVPR*, 2017, pp. 4700–4708.

[62] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[63] X. Zhang, J. Ma, Y. Li, P. Wang, and Y. Liu, "Few-shot learning of parkinson's disease speech data with optimal convolution sparse kernel transfer learning," *Biomedical Signal Processing and Control*, vol. 69, p. 102 850, 2021.

[64] Q. Yu, Y. Ma, and Y. Li, "Enhancing speech recognition for parkinson's disease patient using transfer learning technique," *Journal of Shanghai Jiaotong University*, vol. 27, no. 1, pp. 90–98, 2022.

[65] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[66] J. R. Orozco-Arroyave, *Analysis of Speech of people with Parkinson's disease*. Logos Verlag Berlin GmbH, 2016, vol. 41.

[67] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2095–2103, 2007.

[68] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 3, pp. 235–238, 1977.

[69] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Ismir*, Citeseer, vol. 270, 2000, pp. 1–11.

[70] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B*, vol. 39, no. 1, pp. 1–22, 1977.

[71] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proceedings of EUROSPEECH*, 1997.

[72] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[73] R. Kuhn, P. Nguyen, J.-C. Junqua, *et al.*, "Eigenvoices for speaker adaptation," in *Proceedings of ICSLP*, 1998.

[74] R. Fletcher, *Practical methods of optimization*. John Wiley & Sons, 2013.

[75] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[76] C. M. Bishop, *Neural networks for pattern recognition*. Oxford University press, 1995.

[77] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *ArXiv Preprint ArXiv:1502.03167*, 2015.

[78] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[79] H. Wu, J. Soraghan, A. Lowit, and G. Di Caterina, "A deep learning method for pathological voice detection using convolutional deep belief networks," in *Proceedings of INTERSPEECH*, vol. 2018, 2018.

[80] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, 1994.

[81] Z. Liu, Z. Wu, T. Li, J. Li, and C. Shen, "GMM and CNN hybrid method for short utterance speaker recognition," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 7, pp. 3244–3252, 2018.

[82] G. Zaccone, M. R. Karim, and A. Menshawy, *Deep Learning with TensorFlow*. Packt Publishing Ltd, 2017.

[83] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proceedings of ECCV*, Springer, 2016, pp. 630–645.

[84] R. Reed and R. J. MarksII, *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks*. Mit Press, 1999.

[85] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[86] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[87] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[88] C. Busso, M. Bulut, C.-C. Lee, *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[89] J. Rusz, "Detecting speech disorders in early Parkinson's disease by acoustic analysis," 2018.

[90] T. Bocklet, S. Steidl, E. Nöth, and S. Skodda, "Automatic evaluation of Parkinson's speech-acoustic, prosodic and voice related cues," in *Proceedings of INTERSPEECH*, 2013, pp. 1149–1153.

[91] J. Rusz, J. Klempıř, T. Tykalová, *et al.*, "Characteristics and occurrence of speech impairment in Huntington's disease: Possible influence of antipsychotic medication," *Journal of Neural Transmission*, vol. 121, no. 12, pp. 1529–1539, 2014.

[92] K. Kieburtz *et al.*, "Unified Huntington's disease rating scale: Reliability and consistency," *Neurology*, vol. 11, no. 2, pp. 136–142, 2001.

[93] V. Parsa and D. Jamieson, "Acoustic discrimination of pathological voice sustained vowels versus continuous speech," *J. Speech Lang. Hear. Res.*, vol. 44, no. 2, pp. 327–339, 2001.

[94] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[95] I. Hertrich and H. Ackermann, "Acoustic analysis of speech prosody in Huntington's and Parkinson's disease: A preliminary report," *Clinical Linguistics & Phonetics*, vol. 7, no. 4, pp. 285–297, 1993.

[96] S. Pinto, A. Chan, I. Guimarães, R. Rothe-Neves, and J. Sadat, "A cross-linguistic perspective to the study of dysarthria in Parkinson's disease," *Journal of Phonetics*, vol. 64, pp. 156–167, 2017.

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

[98] J. R. Orozco-Arroyave, E. A. Belalcazar-Bolanos, J. D. Arias-Londoño, *et al.*, "Characterization methods for the detection of multiple voice disorders: Neurological, functional, and laryngeal diseases," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 6, pp. 1820–1828, 2015.

[99] S. Zhao, F. Rudzicz, L. G. Carvalho, C. Márquez-Chin, and S. Livingstone, "Automatic detection of expressed emotion in Parkinson's disease," in *Proceedings of ICASSP*, IEEE, 2014, pp. 4813–4817.

[100] J. Skibińska and R. Burget, "Parkinson's disease detection based on changes of emotions during speech," in *Proceedings of ICUMT*, IEEE, 2020, pp. 124–130.

[101] L. Vavrek, M. Hires, D. Kumar, and P. Drotár, "Deep convolutional neural network for detection of pathological speech," in *Proceedings of SAMI*, IEEE, 2021, pp. 000 245–000 250.

[102] Y.-T. Hsu, Z. Zhu, C.-T. Wang, S.-H. Fang, F. Rudzicz, and Y. Tsao, "Robustness against the channel effect in pathological voice detection," *arXiv Preprint arXiv:1811.10376*, 2018.

[103] M. C. Kruithof, H. Bouma, N. M. Fischer, and K. Schutte, "Object recognition using deep convolutional neural networks with complete transfer and partial frozen layers," in *Proceedings of SPIE*, International Society for Optics and Photonics, vol. 9995, 2016, 99950K.

[104] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[105] J. A. Thompson, M. Schönwiesner, Y. Bengio, and D. Willett, "How transferable are features in convolutional neural network acoustic models across languages?" In *Proceedings of ICASSP*, IEEE, 2019, pp. 2827–2831.